Executive Summary: A new understanding of star formation in the local Universe.

A key step in the evolution of galaxies—the building blocks of the Universe—is the formation of the billions of stars they contain. The star formation process is very sensitive to the local environment, almost ceasing in regions of high galaxy density. This effect has been recognised for decades, but we still do not know if it is due to the formation of different galaxy types in these regions or a direct suppression of star formation. We propose a new approach to separate these two hypotheses by combining radio and optical data to measure the star formation *efficiency* in galaxies, rather than just the overall star formation rate as in previous work. The specific aims of the project are:

1. We will develop and apply new machine learning techniques to match galaxies from very large radio and optical catalogues. These techniques have not previously been applied to such problems in astrophysics.
2. We will use these matched data to measure how the star formation efficiency in galaxies varies with the local galaxy density. This will allow us to test the two rival star formation models discussed above.
3. This project will help UQ develop the expertise necessary to join a national consortium preparing to seek major government funding to support Australian participation in the International Virtual Observatory project.

Aims and Significance of the Project

To understand galaxies, we must understand how and when the stars in them formed. A key constraint on galaxy formation theories is the density-morphology relation [1]: a dramatic decrease in the fraction of galaxies forming new stars in regions of high galaxy density. This relation has been recognised for half a century, but it is still unclear which of several possible physical processes are responsible. There are two main hypotheses: either fewer star-forming galaxies form in high-density regions, or a physical process in these regions suppresses star formation. Previous studies [2, 3] have not been conclusive because the galaxy samples were selected from optical data which are strongly biased by current star formation activity, the very effect we are trying to measure.

We propose a new approach to this problem by investigating galaxies selected from radio data which measure their neutral hydrogen gas content. Neutral hydrogen is the raw material from which stars form. The project is only possible due to the recent completion of the HI Parkes All-Sky Survey (HIPASS) [4] which has made a complete map of neutral hydrogen gas (HI) in the southern sky. We will measure the current rate of star formation in these galaxies with optical data from the SuperCOSMOS [5] sky survey. The ratio of star formation rate to hydrogen mass gives us a new parameter: the efficiency of star formation. The dependence of this quantity on environment will show which of the two processes above contribute most to the density-morphology relation. Previous studies [e.g. 6] have attempted to measure the effect of environment on the hydrogen in galaxies, but they have been inconclusive because the galaxies were all selected from optical samples. Our use of the new HIPASS survey data will overcome this bias.

The major challenge of this project is the correlation of the radio and optical data as there are several optical galaxies within the position uncertainty of each radio galaxy. This problem is common in astrophysics and the traditional approach is to estimate a likelihood ratio for each possible identification [7]. This method is still used currently [8], but is seriously limited by assuming a prior knowledge of the galaxy properties. It is somewhat surprising that learning systems have not been previously applied to this problem, given that there is additional information available that is not used in the likelihood ratio approach. *Our aim is to apply the latest machine learning algorithms to this pattern classification problem, drawing upon a variety of techniques that have been developed within the machine learning community. This will be the first time these methods have been applied to such a problem in astrophysics.*

Neural networks have been used extensively for galaxy classification (as discussed by [9]), but only for morphological classification in the single data set case, and they have problems in relation to architecture selection and to the existence of local minima in the error surface leading to multiple solutions. A more promising method for our problem is the support vector machine [10] which formulates pattern classification as a quadratic programming problem and therefore has a unique solution. We will also consider other methods, including the method of boosting [11], which allows employment of multiple classifiers to improve classification performance.

This project is of strategic importance because of recent national and international initiatives towards an International Virtual Observatory (IVO). The IVO will allow the scientific community to make the best use of the enormous data flow from the latest telescopes worldwide: setting up a system of linked data archives, but more importantly developing high-level software to do new kinds of science. Very large IVO projects have been funded in the USA and Europe. A consortium of Australian observatories and universities (Melbourne and Sydney but not UQ) recently requested 12 months ARC Linkage funding for "phase 1" of Australian participation in the IVO (termed "Australian e-Astronomy"). They proposed three pilot studies focusing on the packaging of existing data archives for inclusion in the IVO. Our independent project is complementary to that proposal. We are taking a cross-discipline approach to develop new techniques for the vital next stage of the IVO: the correlation of very disparate data sets. Our initiative will put UQ in a strong position to join the consortium in a much larger funding application for "phase 2" of Australian e-Astronomy.

Research Plan, Methods and Techniques

Before detailing the individual steps of our research plan we give a short explanation of the classification problem itself.

The classification problem

For each of the 5057 radio galaxies in the HIPASS catalogue we wish to find the most likely optical counterpart in the SuperCOSMOS sky survey catalogue. This problem is illustrated in Figure 1: within the position uncertainty of each radio source there may be several possible optical counterparts. To identify each counterpart unambiguously we would require detailed high-resolution radio imaging of each HIPASS galaxy to give an accurate position and/or spectroscopy of all the possible optical counterparts to see which has a velocity matching the radio source. These approaches are not practical at present due to the large number of observations required. Instead we will develop a statistical approach using machine learning techniques to identify the most likely counterpart to each radio source based on the other optical properties of the possible matching galaxies. For each possible counterpart of a radio source we have the following parameters measured: the distance on the sky from the nominal radio source position, the optical flux in 3 bands (Blue, Red and Infrared) and the morphology of the image (area, semi-major axis length, ellipticity, classification as star or galaxy). We also have radio flux of the radio source and its velocity (which indicates its distance).

The optical counterpart is rarely the closest galaxy to the radio position, as seen in Figure 1. In that example, the flux and distance of the radio source make the larger galaxy "X" the most likely counterpart. Other factors like the "colour" of each optical galaxy (given by the ratio of fluxes in different bands) may be a very important as bluer galaxies tend to be forming more stars and thus contain more gas and have more radio emission. It is not known *a priori* which combination of these parameters will provide the best identification of the optical counterparts: this is what we will determine with the machine learning approach.

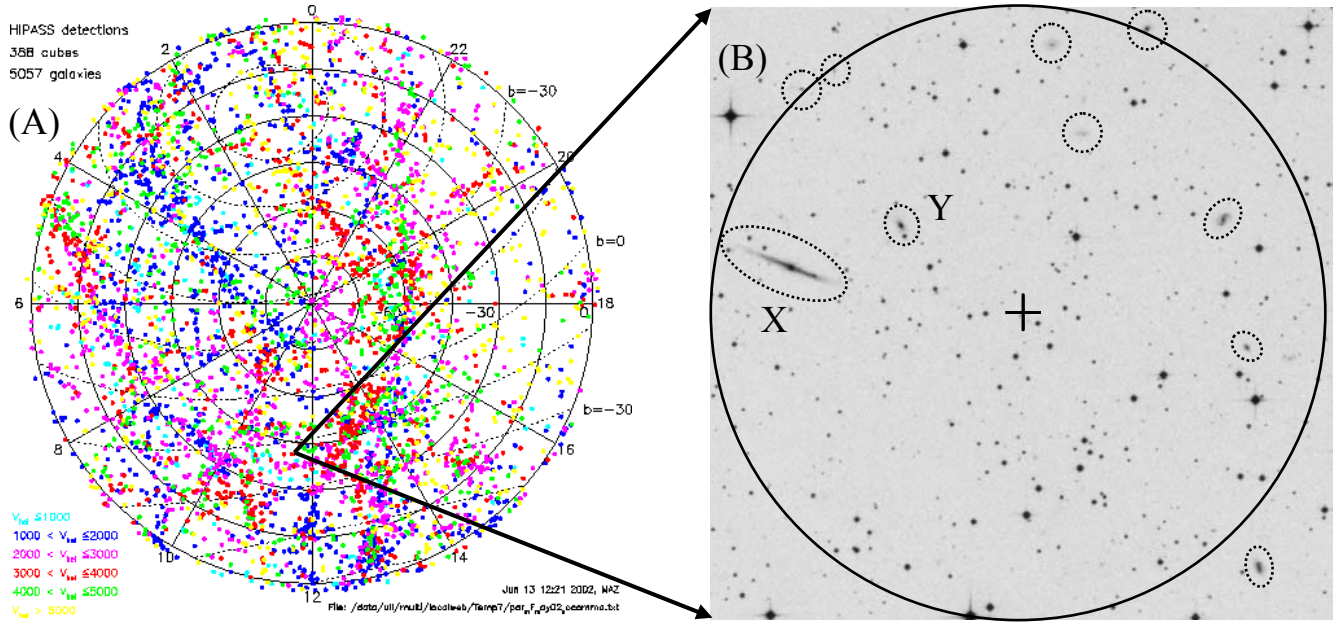

Figure 1: The identification of optical counterparts of the radio galaxies. **(A)** The distribution on the sky of all 5057 radio galaxies in the HIPASS catalogue. The plot shows the whole southern sky with the South Celestial Pole in the centre and the equator at the edge. **(B)** The optical sky survey image of a small region of the sky (in negative representation) centred on the position of a single HIPASS radio galaxy. The large solid circle shows the (5 arc minute radius) position uncertainty of the radio position; within this circle are several galaxies in the optical SuperCOSMOS catalogue, indicated by dotted ellipses. In this particular case the velocity of the large galaxy "X" is published and matches that of the radio source, so this is the correct identification, not the closest galaxy "Y".

Detailed Research Plan

1. Preparation of the input sets of galaxy data.

Drinkwater, as a member of the HIPASS team already has access to the catalogue of 5057 radio galaxies (Zwaan et al., in preparation). For each of these we need to extract all possible optical counterparts from the SuperCOSMOS catalogue at the University of Edinburgh. We expect a total of about 50 000 possible counterparts to be found. The SuperCOSMOS catalogue can be interrogated online, but for such a large project it would be most efficient if we could send the Research Assistant (RA) to work directly with our colleagues in Edinburgh. This is particularly true if we wished to consider options of removing incorrectly classified objects from the catalogue at this stage: this process is very labour-intensive requiring visual inspection of all possible counterparts and could only be done by sending the RA to Edinburgh. Drinkwater has an existing collaboration with the SuperCOSMOS group who will be able to provide some assistance if we are unable to send the RA to visit them.

A vital step in machine learning is the training of the chosen method using a "training set" of data for which the correct answers are known independently. This is available for our problem in the form of optical galaxies with previously measured velocities in the literature that match that of one of our radio sources as shown in Figure 1. It is relatively simple for us to generate a training set of several hundred such examples using the NASA Extragalactic Database (NED). However some software development will be needed to process and correlate the results.

2. Basic problem and candidate learning techniques

The basic classification problem is to identify the object in the SuperCOSMOS catalogue that is responsible for a given radio source. As stated above, the NED database contains several hundred examples where this identification has been made and these examples will provide

data for training our learning systems. We will commence with the most straightforward approach to training in which the full set of available features from both the radio source and the corresponding optical region will be employed as inputs to the learning system. That is, the learning system will be fed radio flux, source velocity, distance (in the optical image) from the radio source, the optical flux in 3 bands, and the parameters defining the morphology of the image (detailed above). This will provide an input vector to the learning system with 10 entries.

The examples in the NED database where identifications have been made will provide positive instances only. That is, instances where the input data correspond to the true radio source. We will also require negative instances, and these are abundantly available, as is clear from Fig. 1(B). It is likely, however, that a high proportion of these negative instances will contribute little or nothing to the discriminating power of a trained classifier due to being weak candidates for recognition as a true radio source. It will be necessary to identify these weak candidates so that processing time is not wasted on thousands of data vectors that are essentially redundant. The procedure we will follow is detailed in the next section. It is also important to recognise that the positive examples in the NED database are there because, by and large, they are the ones most easily matched with radio sources. They are mainly galaxies for which both the radio flux and the optical flux are high. This means that they are somewhat atypical and that, in this sense, the training data are biased. This bias will have to be dealt with in some way and this problem is also discussed in the next section.

The first learning system we will consider is the support vector machine (SVM) [10]. The SVM computes a nonlinear mapping that transforms its input data into a high-dimensional feature space where patterns of different classes can be separated by a hyper plane. This hyper plane then provides a nonlinear decision surface in input space. In cases where complete separation of the two classes is not warranted, the best generalization performance is obtained by reducing the nonlinearity in the mapping and a suitable reduction can be determined by cross-validation. The SVM has been developed in the last few years and gives state-of-the-art classification performance. But no classifier gives best performance on all data sets so we will apply other leading classifier systems to the data sets under consideration.

The other high-performance learning classifiers that we will deploy are boosting and Gaussian processes. Boosting [11] is a procedure that trains a sequence of relatively weak classifiers, with those later in the sequence being trained in a way that places emphasis on those training examples that earlier classifiers tended to misclassify. The method works best on so-called unstable classifiers such as decision trees and neural networks. We will use neural networks with a regularization technique to avoid overfitting. (Note that the tendency of neural networks to produce multiple solutions is not an issue when they are subjected to boosting). Gaussian processes [12] were originally developed for regression problems and are related to the technique of “kriging” employed in the spatial statistics community. Only very recently has a version applicable to classification appeared in the literature, but it does have the advantage of assigning probabilities to its class estimates.

3. Learning System Methodology

As stated above, there are two issues that need to be resolved regarding the data before we can expect our learning systems to provide accurate classifications.

(a) The first issue concerns the fact that each positive example in the training data has associated with it a very large number of negative examples and most of these negative examples will be redundant. These redundant examples need to be identified and eliminated as early as possible. Fortunately, we have recently developed a technique that allows this to be achieved when using a support vector machine. The method is an outgrowth of [13] and will be published in [14]. Basically the method involves identifying and discarding those training vectors that are linearly dependent in the feature space of the SVM. Recall that the feature space is arrived at via a nonlinear mapping from input space so this method will not be directly applicable to other learning systems like boosting and Gaussian processes.

However, it is expected that the method can be made to work for these systems by making appropriate changes to the procedure. This will be one of our first investigations.

(b) The other issue concerns the presence of bias in the training data due to positive examples being largely drawn from regions of high radio and optical flux. Various techniques for dealing with this kind of bias have been described in the statistics literature (see, for instance, [15], [16]) and we will identify the method that is most suitable for this particular problem. It is likely that the method will require some adaptation for application in a machine learning situation. To assist with solution of this problem we will also seek to make use of the fact that some input features (eg galaxy morphology) vary in a statistically predictable fashion as one moves from the region of high radio and optical flux (where most of the training data lie) to regions of low flux.

The classification work will be able to commence immediately with the application of support vector machines to data that has had redundant vectors removed. Once we have established the best way to remove redundant vectors for the other learning systems, these, too, can be applied to the available data. And as we develop methods for dealing with the bias in the data, we will be able to demonstrate improved performance. Ultimately we will be able to identify the most suitable (and complete) technique to employ and we will then be able to tackle the full set of data.

4. Star formation analysis.

Once the optical counterparts of each radio galaxy are identified, we will proceed to measure the physical parameters of each system: hydrogen mass (radio flux), stellar mass of galaxy (infra-red optical flux), current star formation rate (blue optical flux) and the efficiency parameter. We will make a preliminary analysis of these properties to test the data set for errors, notably by examining any significant outliers from the broad correlations expected between these parameters.

We will then make our primary test of the two star formation hypotheses. We will measure the relationship between the star formation efficiency parameter and the local galaxy density. If the efficiency decreases significantly in regions of higher density we will have shown that a physical process is suppressing star formation. Conversely, if no change is detected we will have shown that a different mix of galaxies has formed in these regions. In either case we will use our data to further constrain the physical mechanisms involved.

Our use of novel learning systems in the classification stage of our project may also open up the possibility of developing new insights into the relationship between the optical and radio properties of galaxies. Unlike artificial neural networks, techniques such as the support vector machines actually specify the rules that were derived to make the classifications: we will also analyse these rules to see if they reveal any new insights into the properties of these galaxies.

Justification of Budget

Research Assistant (RA): This project involves the collection of a large data set and the subsequent development, testing and applications of several machine learning algorithms to match the radio and optical data. Once these are matched there is a further stage of calculating physical parameters to analyse the relationship between star formation and environment. The CIs together have the expertise to lead and direct this project, but do not have sufficient time available to undertake the work involved. The appointment of a research assistant familiar with either the astrophysics or machine learning fields is therefore essential to the project at priority A.

Computer: The project centres on the correlation of two large data sets: the 5000 radio galaxies and the 50 000 potential optical counterparts. We therefore need to provide a

dedicated, fast PC computer for the use of the Research Assistant for the duration of the project at priority A.

Travel: The strategic or "development" aspect of this project focuses on the new techniques we can bring to a future large collaborative proposal to fund "Australian eAstronomy". To this end the CIs need to participate in meetings with the other Australian researchers involved in the national proposal, so a meeting in Melbourne is required at priority B. The project also depends significantly on the provision of the optical data by the SuperCOSMOS group at Edinburgh: they can provide the required data by posting tapes to us, but we can potentially obtain higher quality data if the RA visits Edinburgh to work directly with them. This is requested at priority C as the project is still possible without it.

Roles and Responsibilities of the Investigators

Drinkwater: will provide overall leadership of project, and in particular will arrange access to the data and formulate the physical questions to be addressed. Drinkwater is a member of the HIPASS team with access to the new galaxy catalogue; he also has extensive experience in the areas of star formation and radio-optical studies [17, 18]. He will train the RA as required in analysis of the galaxy catalogues and studies of the new efficiency parameters. He will write up the astrophysical results.

Downs: will provide key intellectual input regarding treatment of the special peculiarities of this problem in a machine learning setting. These peculiarities include the presence of bias in the data and the fact that the data contain vastly more negative examples than positives (addressed in his recent work [13, 14]). He will also provide the machine learning software, train the RA in its use, and produce the paper on this new application of machine learning techniques.

Timetable

2003 Jan-Feb:	Establish access to the HIPASS and SuperCOSMOS data and make preliminary selection of galaxies from both catalogues. If possible, the RA will travel to Edinburgh. Commence investigations on the problem of redundant data.
Mar:	Obtain some preliminary results using the support vector machine.
Apr-July	Apply other classification algorithms to the training data, investigating the problems of bias and continuing the study of redundant data if necessary.
Aug-Sep	Apply the best method(s) to the complete data set to determine the optical counterparts; measure their properties and create a final catalogue of the matched data.
Sep (?)	CIs meet with colleagues in Melbourne to discuss national funding plans.
Oct-Nov	Analyse the data for the dependence of star formation efficiency on environment and determine which hypothesis is best supported.
Nov-Dec	Prepare results for publication: papers on new machine learning techniques; the dependence of star formation on environment.

References

[1] Dressler A, “Galaxy morphology in rich clusters - Implications for the formation and evolution of galaxies”, *Astrophysical Journal*, 236, 351, 1980
[2] Postman M, Geller M, “The morphology-density relation - The group connection”, *Astrophysical Journal*, 281, 95, 1984
[3] Hashimoto Y, et al. “The Influence of Environment on the Star Formation Rates of Galaxies”, *Astrophysical Journal*, 499, 589, 1998
[4] Barnes D G, et al., “The HI Parkes All Sky Survey: southern observations, calibration and robust imaging”, *Monthly Notices of the Royal Astronomical Society*, 322, 486, 2001
[5] Hambly N C, et al., “The SuperCOSMOS Sky Survey - I. Introduction and description”, *Monthly Notices of the Royal Astronomical Society*, 326, 1279, 2001
[6] Schröder A, Drinkwater M J, Richter O-G, “The neutral hydrogen content of Fornax cluster galaxies”, *Astronomy & Astrophysics*, 376, 98, 2001
[7] Sutherland W, Saunders W, “On the likelihood ratio for source identification”, *Monthly Notices of the Royal Astronomical Society*, 259, 413, 1992
[8] Mann R G, et al., “Observations of the Hubble Deep Field South with the Infrared Space Observatory - II. Associations and star formation rates”, *Monthly Notices of the Royal Astronomical Society*, 332, 549, 2002
[9] Odewahn S C, et al., “Automated Galaxy Morphology: A Fourier Approach”, *Astrophysical Journal*, 568, 539, 2002
[10] Burges C J C, “A tutorial on support vector machines for pattern recognition” *IEEE Transactions on Data Mining and Knowledge Discovery*, **2**, 121-167, 1998.
[11] Schapire R E and Singer Y, “Improved boosting algorithms using confidence-rated predictions”, *Machine Learning*, **37**, 297-336, 1999.
[12] Williams C K I and Barber D, “Bayesian classification with Gaussian processes”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1342-1351, 1998.
[13] Downs T, Gates K E and Masters A, “Exact simplification of support vector solutions”, *Journal of Machine Learning Research*, **2**, 293-297, 2001.
[14] Gates K E and Downs T, “Eliminating redundant vectors prior to the training of support vector machines”, (in preparation).
[15] G J McLachlan, “Discriminant analysis and statistical pattern recognition”, Wiley, 1992.
[16] Little R J A and Rubin D B, “Statistical analysis with missing data”, Wiley, 1987.
[17] Drinkwater M.J., Gregg M.D., Holman B.A., Brown M.J.I., “The evolution and star formation of dwarf galaxies in the Fornax Cluster”, *MNRAS*, 326, 1076, 2001
[18] Drinkwater M.J. et al., “The Parkes Half-Jansky Flat-Spectrum Sample”, *MNRAS*, 284, 85, 1997