Machine Learning Methods for Radio Host Cross-Identification with Crowdsourced Labels

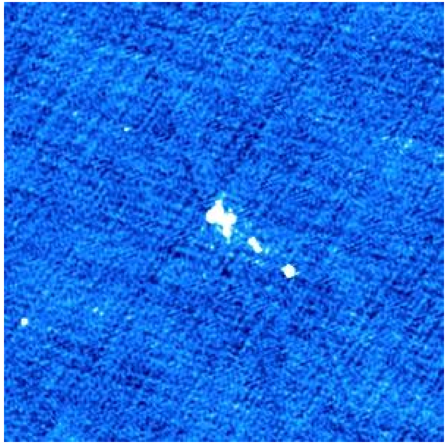
Matthew Alger (ANU),
Julie Banfield (ANU/WSU),
Cheng Soon Ong (Data61/ANU),
Ivy Wong (ICRAR/UWA)

Slides: http://www.mso.anu.edu.au/~alger/sparcs-vii

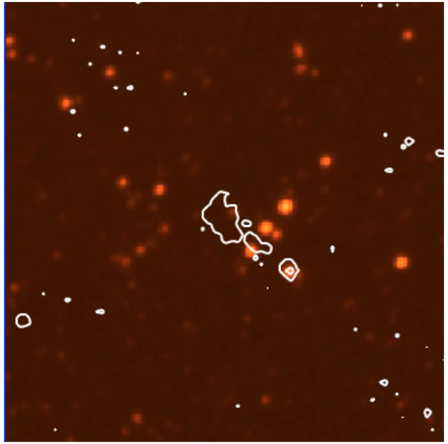
Australian National University

Host Galaxy Cross-Identification

- Problem: match radio emission to its host galaxy at other wavelengths
- Hard:
  - Radio emission can be extended at scales of tens of arcminutes
  - Often no clear relationship between radio emission and host galaxy



FIRSTJ023838.0+023450
at 1.4 GHz.
Image: FIRST

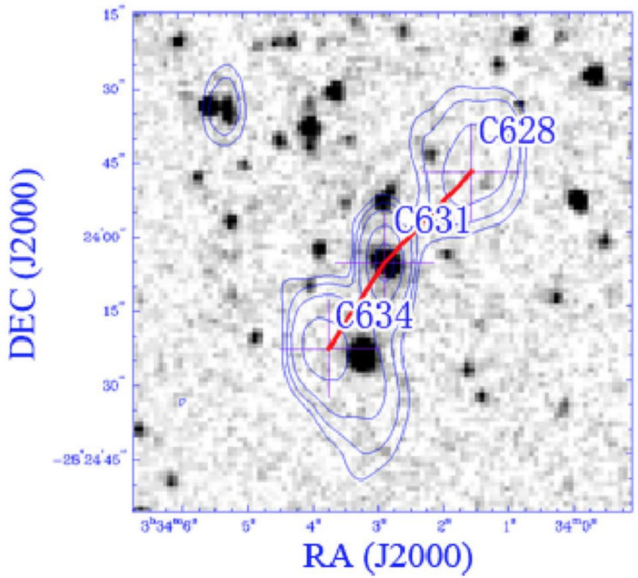


FIRSTJ023838.0+023450
ininfrared.
Image: WISE

Host Galaxy Cross-Identification

Current approaches:

- Manual
- Crowdsourcing
- Nearest neighbours
- Bayesian methods
- Likelihood ratio



Bayesian model fit to a radio triple.
Image: ATLAS (radio), SWIRE (infrared), Fan et al. 2015

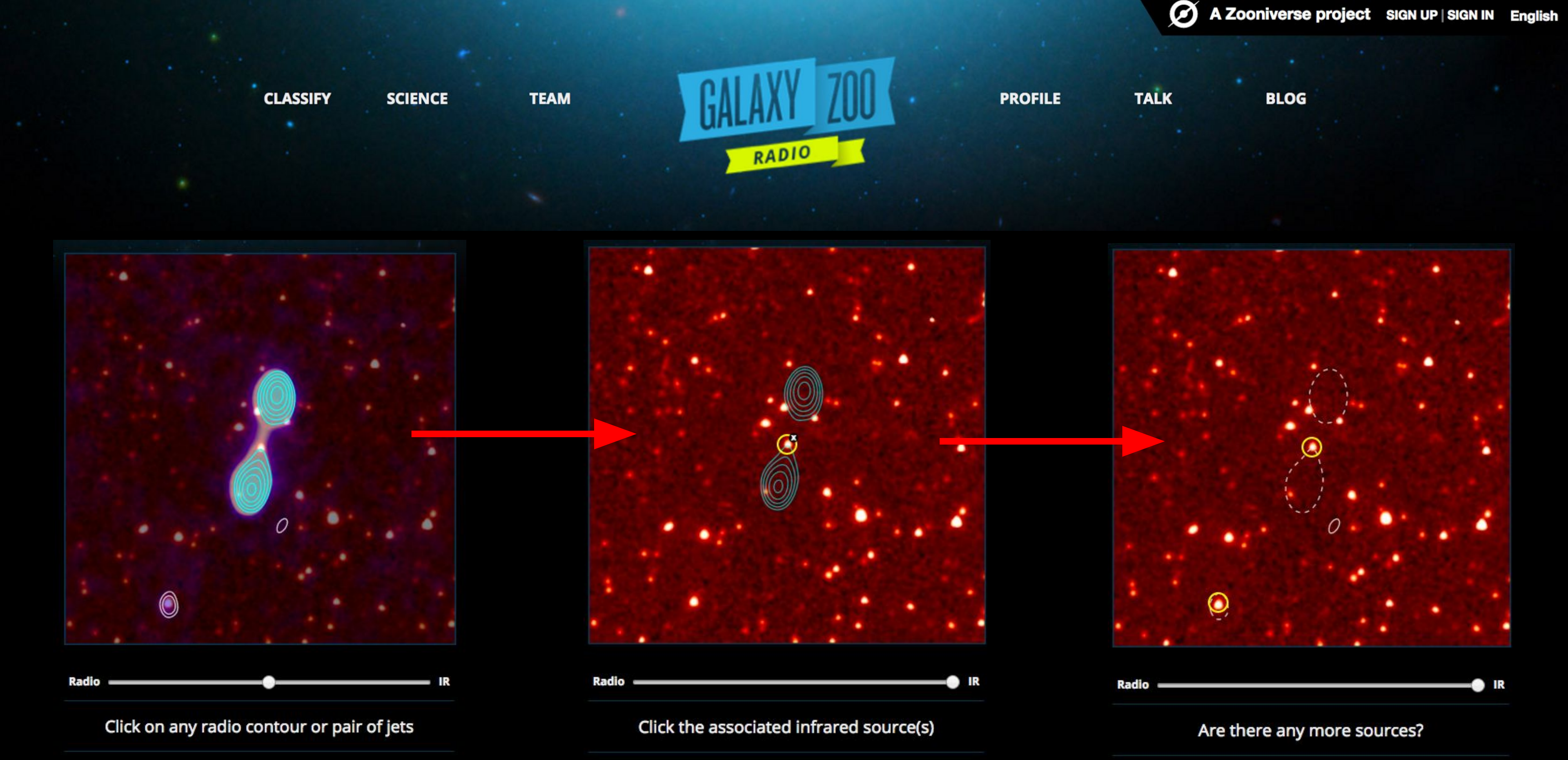
Host Galaxy Cross-Identification

Our approach:

- Casts cross-identification as *object localisation* so we can use algorithms from computer vision
- Allows training cross-identification methods using existing cross-identification datasets

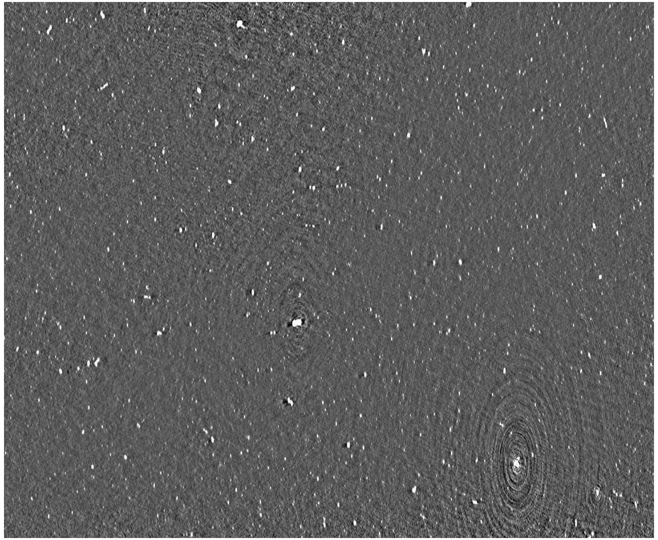
Radio Galaxy Zoo

- Crowdsourced, citizen science project
- Volunteers cross-identify radio emission from FIRST and ATLAS-CDFS with infrared host galaxies from WISE and SWIRE-CDFS
- See Ivy’s talk later this session


A Zooniverse project
SIGN UP | SIGN IN
English
CLASSIFY
SCIENCE
TEAM
GALAXY ZOO
RADIO
PROFILE
TALK
BLOG
Radio
IR
Click on any radio contour or pair of jets
Radio
IR
Click the associated infrared source(s)
Radio
IR
Are there any more sources?

ATLAS-CDFS

- ~2000 sources in ATLAS DR3
- Radio Galaxy Zoo source identifications and SWIRE host cross-identifications
- ~500 sources cross-identified with SWIRE in ATLAS DR1



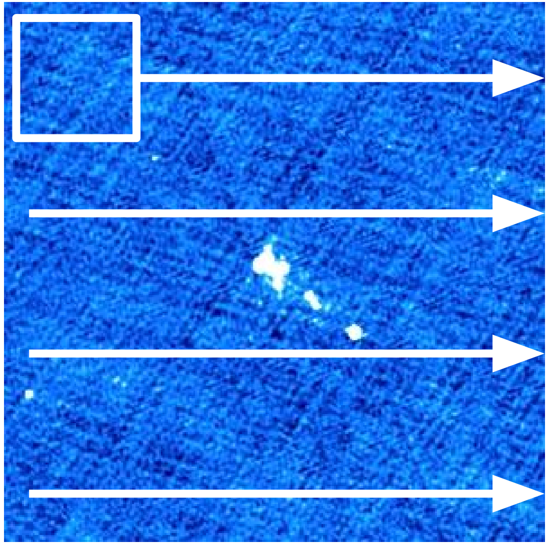
ATLAS observations of CDFS.
Image: ATLAS, Franzen et al. 2015

Supervised Machine Learning

- Encompasses classification, regression, and other function approximation tasks
- Promising methods for handling very large datasets
- Training requires a large set of labelled data
- Application requires converting problem into a function approximation problem
- Binary classification best understood

Machine Learning for Cross-Identification

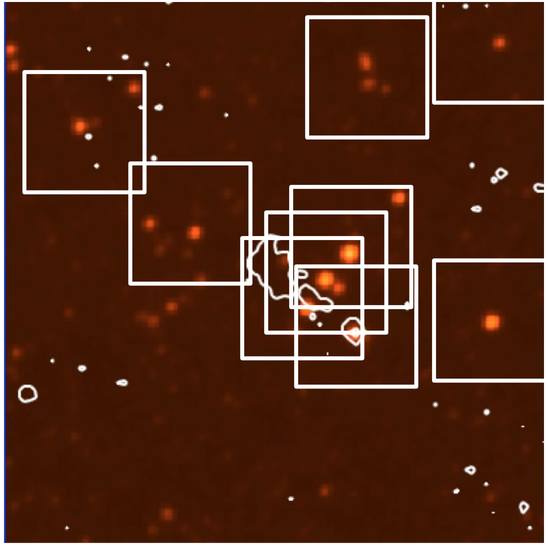
- Allows use of Radio Galaxy Zoo data for training
- Need to convert cross-identification into a machine learning task
- First pass from computer vision:
  - *Sliding window* approach
  - Given an image of radio emission, classify each square patch based on whether the AGN is located there
  - Not terribly efficient
  - Binary classification!



Scanning to find the host galaxy.
Image: FIRST

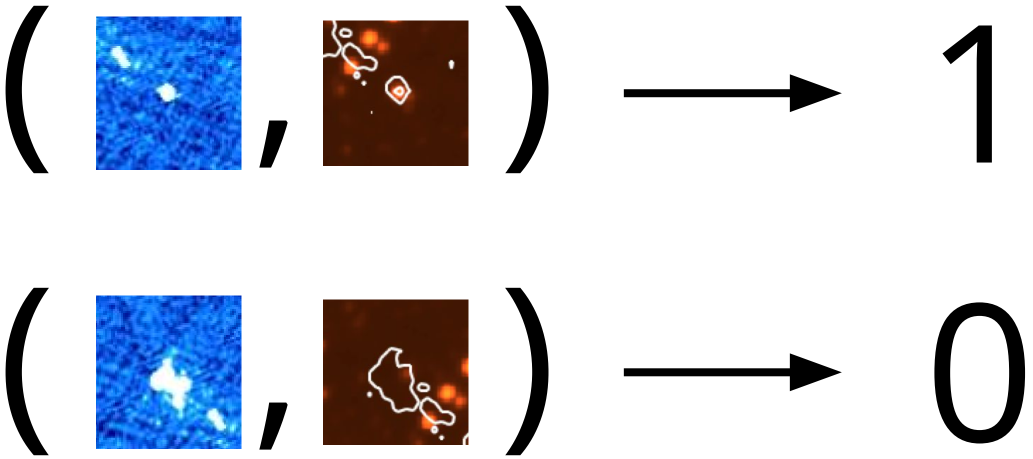
Machine Learning for Cross-Identification

- Second attempt:
  - Assume host galaxies visible in infrared
  - Given an image of radio emission, classify each candidate host galaxy in that image based on whether it is the host galaxy
  - Much more efficient!



Candidate host galaxies.
Image: FIRST/WISE

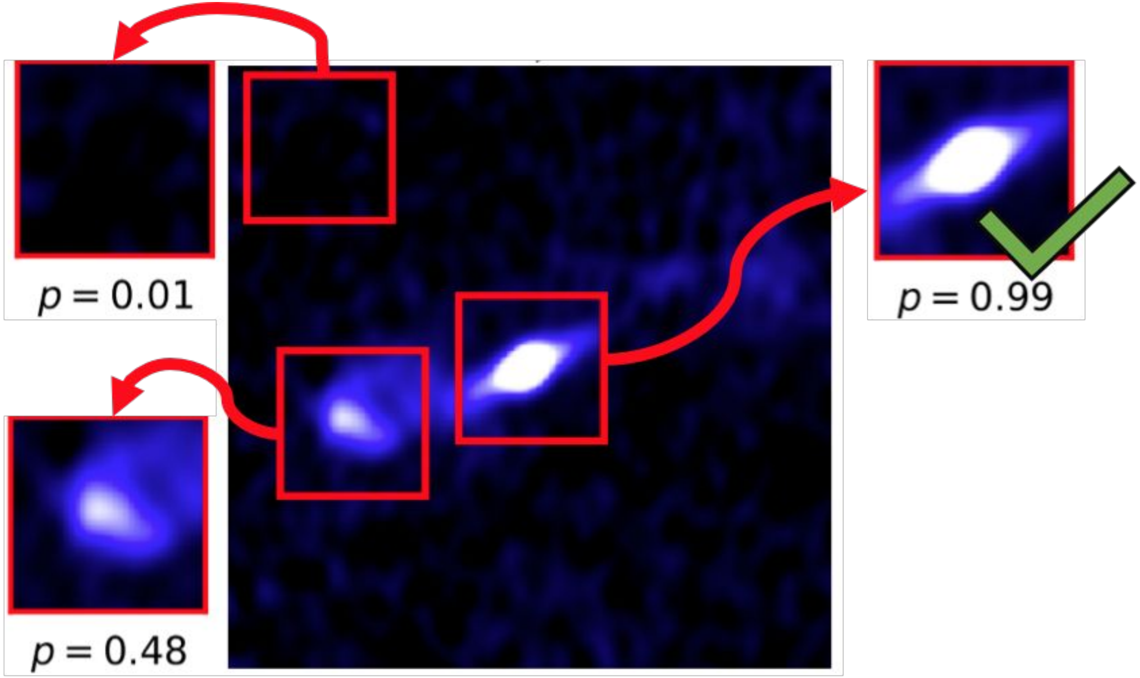
Cross-Identification with Binary Classification



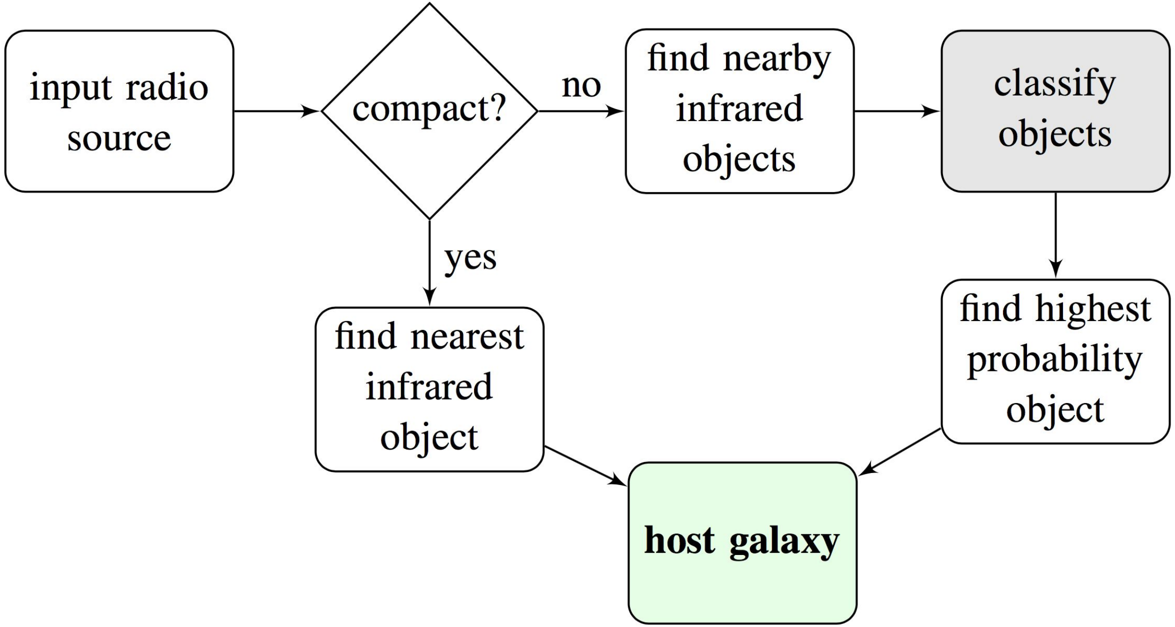
Representation of galaxy

Whether galaxy has an AGN

Cross-Identification with Binary Classification



Cross-Identification with Binary Classification

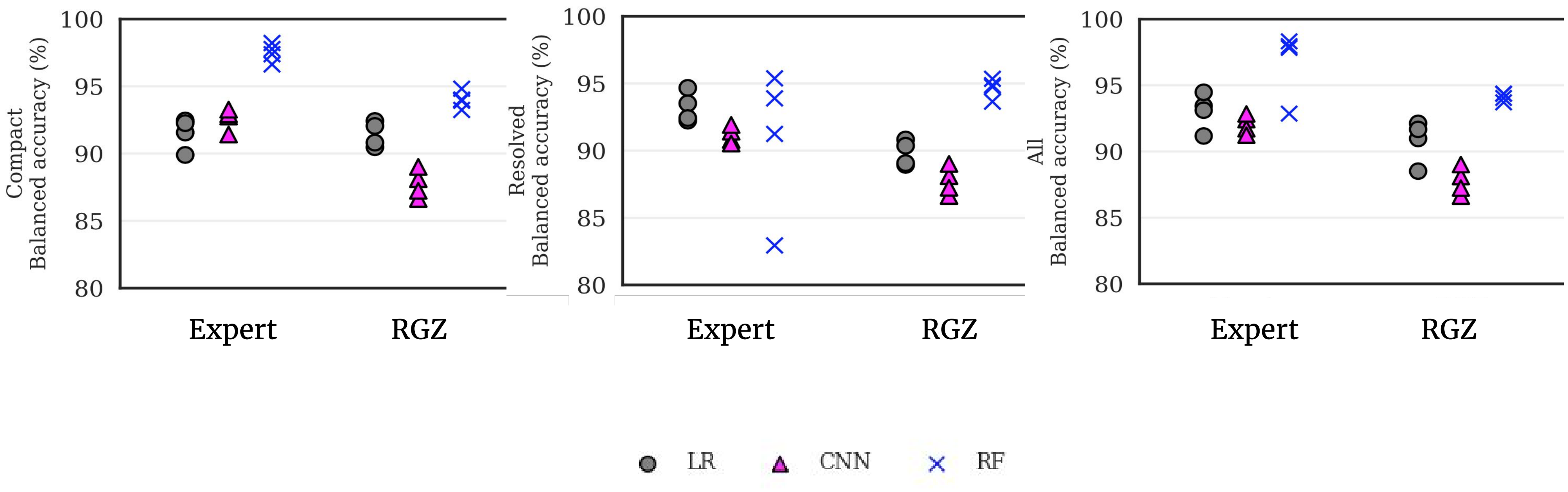


Experimental Method

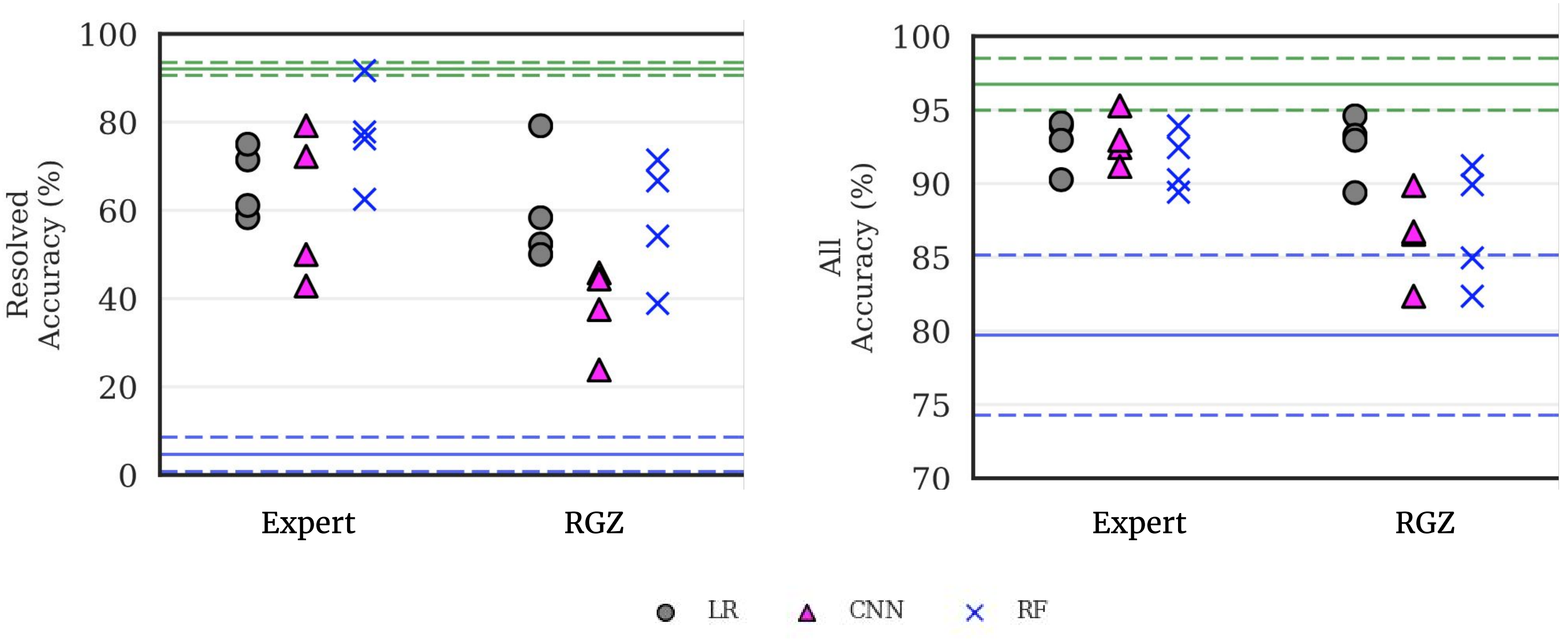
- Three classifiers:
    - Logistic regression
    - Random forests
    - Convolutional neural networks
- Labelled training data:
    - Inputs are square image cutouts centred on candidate host galaxies
    - Expert labels from ATLAS DR1
    - Crowdsourced labels from Radio Galaxy Zoo

- Split CDFS into resolved/compact sources
- Train on 75% of CDFS
- Test by comparing outputs to ATLAS DR1 on remaining 25%

Classification Accuracy on SWIRE-CDFS



Cross-Identification Accuracy on SWIRE-CDFS

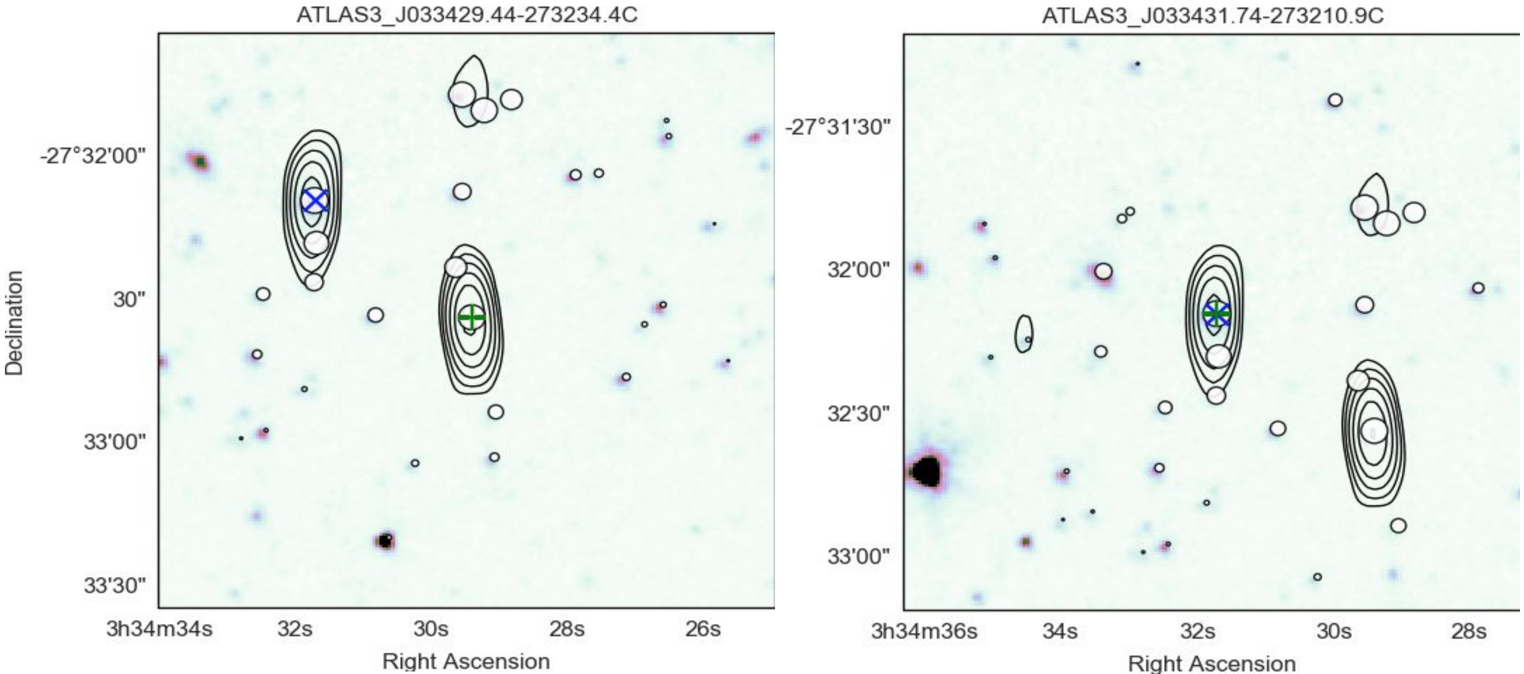


Key Assumptions

- Assumptions on search radius:
  - One host galaxy in radius
  - All radio emission from a source is contained in radius
- Assumptions on candidate host galaxies:
  - Host galaxies visible in infrared
- Assumptions on sliding window radius:
  - Information in sliding window sufficient to determine host galaxy
- We defer these problems for now

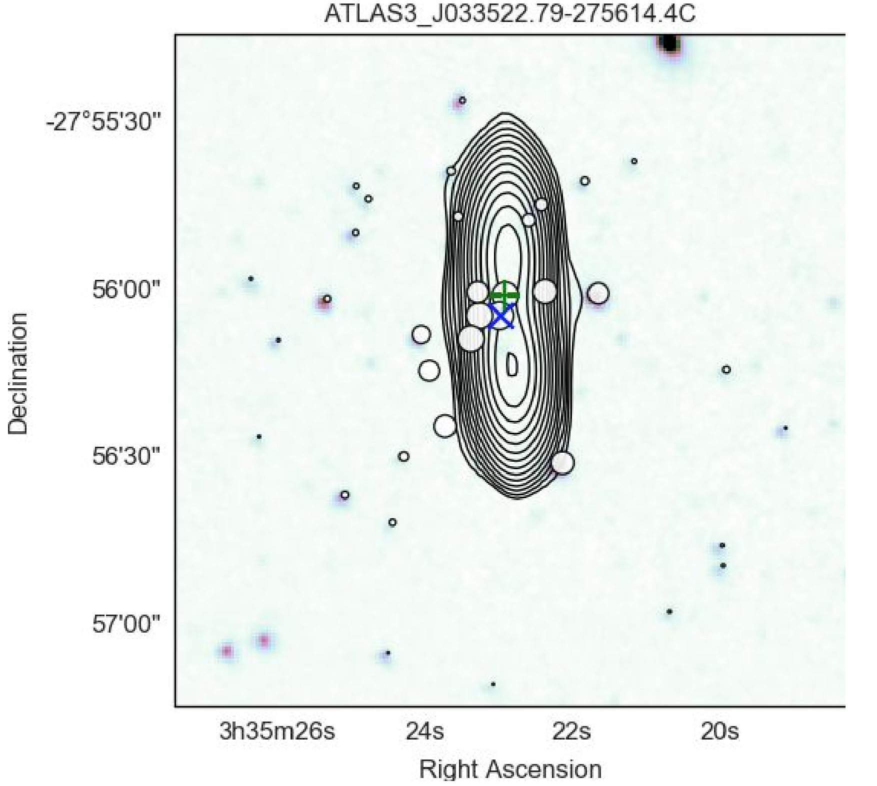
Failure Case — Multiple Hosts

- Assumption: One host galaxy in search radius
  - Search radius = 1’ (as in Radio Galaxy Zoo)
  - Assumption often broken



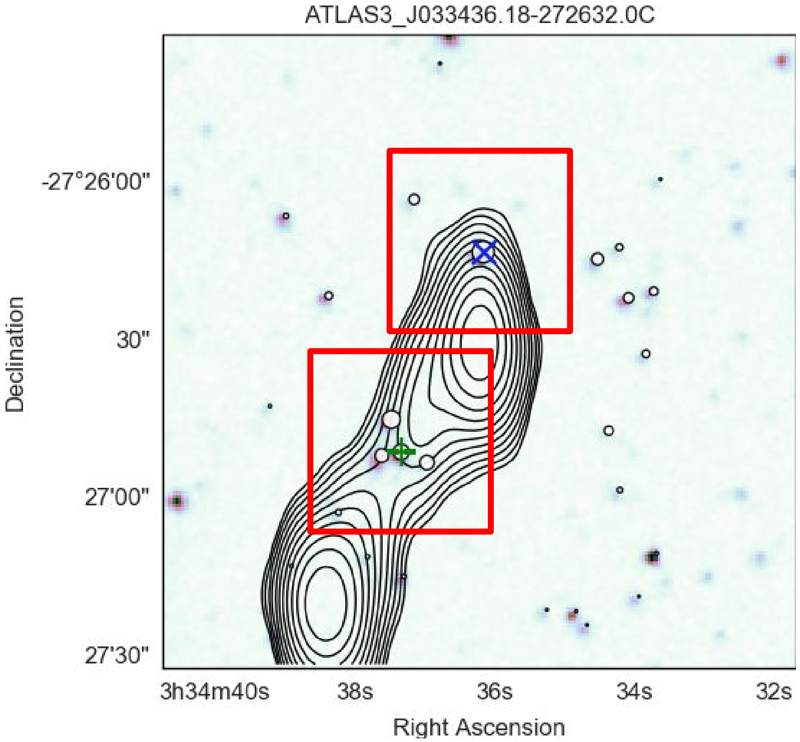
Failure Case — Nearby Candidate Hosts

- Hard to distinguish between nearby candidate hosts
- A prior could help resolve this issue



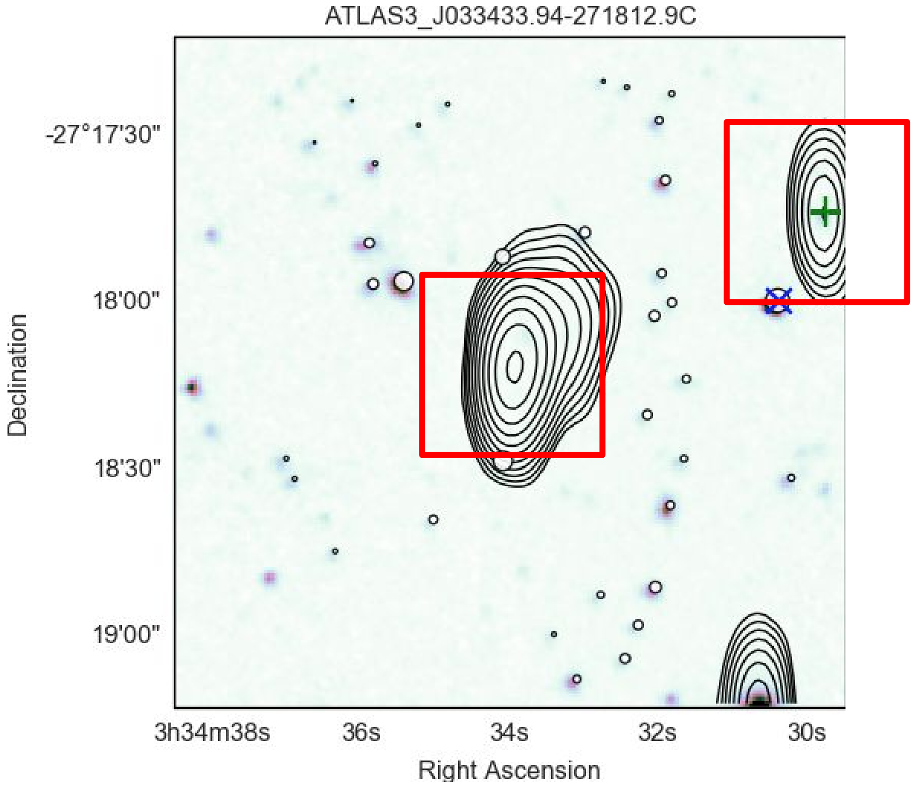
Failure Case — Misidentified Lobe

- Abundance of compact objects in training data bias the classifier toward bright radio lobes
- Larger datasets with more varied radio doubles would likely resolve this issue
- Larger window sizes can help (but too large provides the classifier with too many inputs)



Failure Case — Search Radius

- Search radius of 1’ too small to find all host galaxies
- ...But making the search radius too large worsens the problem of multiple hosts



Future Work

- More data for convolutional neural network training
  - Radio Galaxy Zoo-FIRST?
  - Simulations?
- Dynamically choose window sizes and search radii
- Combine computer vision methods with radio source identification methods

Summary

- We developed a machine learning approach for host galaxy cross-identification
- We trained the method on both expert cross-identifications from ATLAS DR1 and volunteer cross-identifications from Radio Galaxy Zoo
- Crowdsourcing provides a promising source of supervised machine learning training data
- Better model selection and incorporating source identification would improve accuracy