Learning from Crowd Labels to find Black Holes

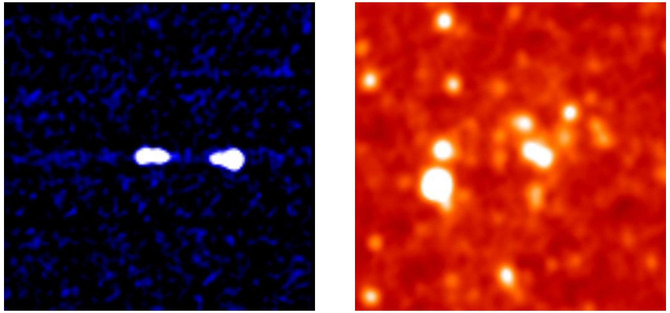
Matthew Alger

Supervisors: Dr Cheng Soon Ong & Dr Julie Banfield

Slides: http://goo.gl/UkgFVc

Project sketch

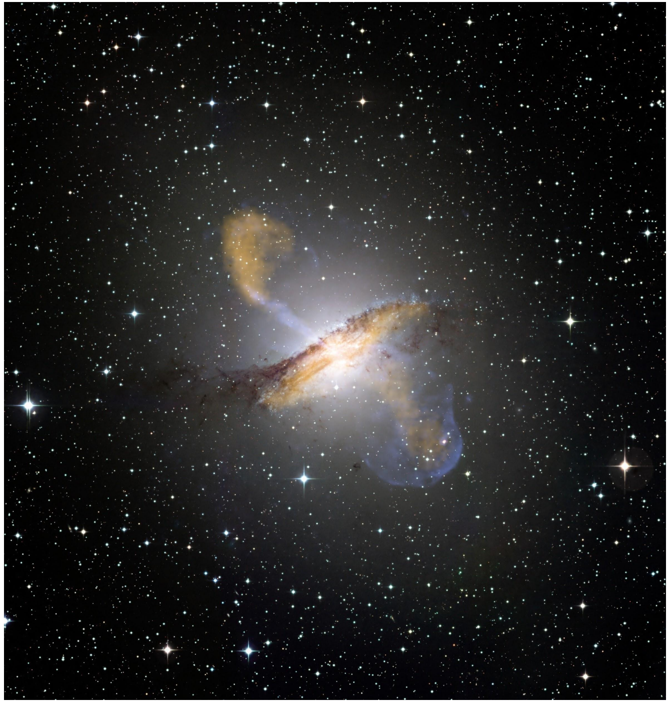
- We want to automate radio cross-identification, a problem in radio astronomy
- We need to automate this because new radio surveys and telescopes will generate more data than we can currently deal with
- I developed a naïve method for automated cross-identification, trained with crowdsourced labels from Radio Galaxy Zoo



The same patch of sky in both radio (left) and infrared (right).
Image: FIRST (Radio); WISE (Infrared)

Presentation outline

- Motivation
  - Radio cross-identification
  - SKA, ASKAP, and the Australia Telescope
  - Radio Galaxy Zoo
- The Galaxy Classification Task
  - Finding black holes is object localisation
  - Binary classification with logistic regression
  - Representing galaxies as vectors
- Learning from the Crowd
  - Majority vote
  - Raykar et al.
- Active Learning



Centaurus A, a nearby radio AGN.

Image: ESO/WFI (Optical); MPIfR/ESO/APEX/A.Weiss et al. (Submillimetre); NASA/CXC/CfA/R.Kraft et al. (X-ray)

Motivation

Radio cross-identification in astronomical surveys

- Radio cross-identification
  - Multi-wavelength surveys
  - Source cross-identification
- Radio telescopes and surveys
  - Square Kilometre Array
  - ASKAP and EMU
  - Australia Telescope and ATLAS
- Radio Galaxy Zoo

Observations of the sky at different wavelengths

Optical
Image: ESO/WFI/M.Rejkuba et al.

Infrared
Image: NASA

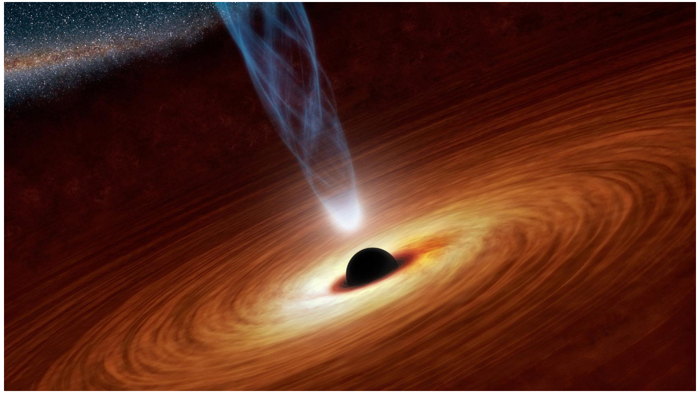
X-ray
Image: NASA/CXC/U.Birmingham/M.Burke et al.

Radio
Image: NSF/VLA/Univ.Hertfordshire/M.Hardcastle

Images of Centaurus A at different wavelengths.

Radio active galactic nuclei (AGNs)

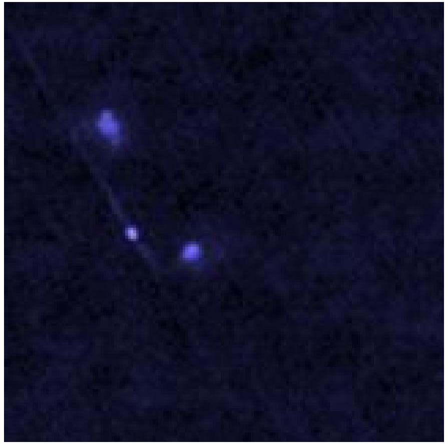
- Supermassive black holes at the centre of galaxies
- May be involved in galactic evolution and star formation
- Has jets that emit synchrotron radiation in radio wavelengths
    - We see these with radio telescopes



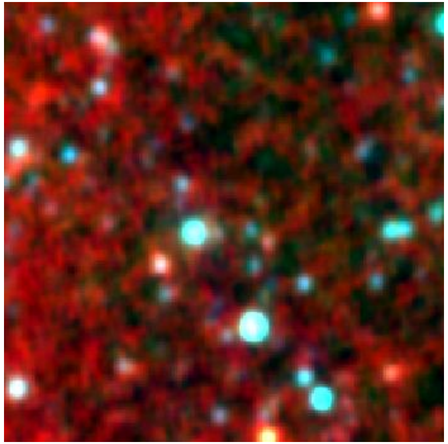
Artist’s impression of an AGN.
Image: NASA/JPL-Caltech

Radio cross-identification

- Match a radio object (i.e. a black hole) to its corresponding object in other wavelengths
- Hard!
  - Multiple components
  - Arbitrarily large
  - Complex components
  - Unclear relationship to other wavelengths



A complex radio object.
Image: FIRST



The same image in infrared.
Image: WISE

The Square Kilometre Array

- Very (very) big radio telescope
- Expected to be constructed by 2024
- Very powerful
  - 50 times more sensitive than other radio telescopes
  - 10000 times faster than other radio telescopes
- Will generate *a lot* of data
  - Phase I will produce 160 TB/s
  - Phase II could produce up to 10 PB/s

Artist's impression of the SKA.

Image: Swinburne Astronomy Productions/SKA Program Development Office

SKA pathfinders

- New telescopes built to test SKA technologies
- Also very big
- Starting to receive data now

MeerKAT.
Image: Mike Peel (www.mikepeel.net)

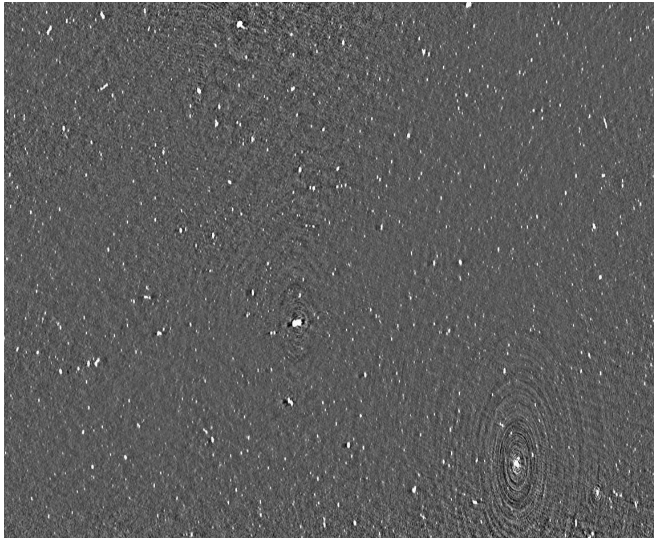
Australian SKA Pathfinder.
Image: CSIRO

The Evolutionary Map of the Universe (EMU)

- Upcoming radio survey of the southern sky
  - Using ASKAP
  - To be completed ~2018
- Very big
  - Will cover over 75% of the entire sky
  - Expected to find 70 million new radio galaxies compared to 2.5 million known now
  - ~10% won’t be cross-identifiable with existing algorithms

The Australia Telescope Large Area Survey (ATLAS)

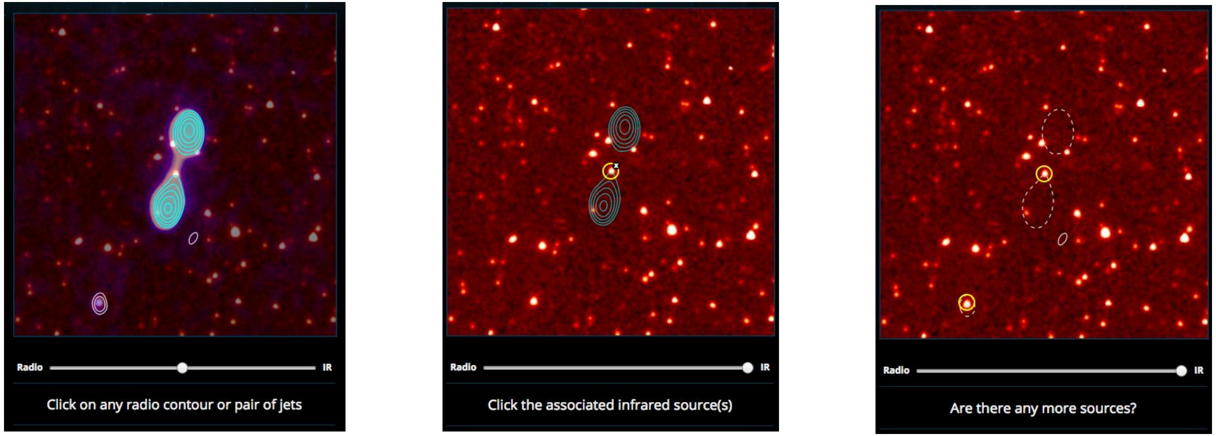
- Pilot survey for EMU
  - Similar resolution
  - Similar sensitivity
  - Similar wavelengths
- Not very big!
  - ~4000 radio objects
  - ~0.02% of the sky



Radio image of the CDFS field.
Image: ATLAS/Franzen et al. 2015

Radio Galaxy Zoo (RGZ)

- Citizen science project to cross-identify radio galaxies
- Crowdsources the cross-identification problem
- Cross-identified ~100 000 galaxies so far!



The Radio Galaxy Zoo web interface.
Image: http://radio.galaxyzoo.org/

Some numbers

- ~6000 cross-identifications by expert astronomers (ever)
- ~100 000 cross-identifications by Radio Galaxy Zoo volunteers (in 3 years)
- ~70 000 000 radio galaxies in EMU
  - It would take 2100 years for Radio Galaxy Zoo to cross-identify EMU!

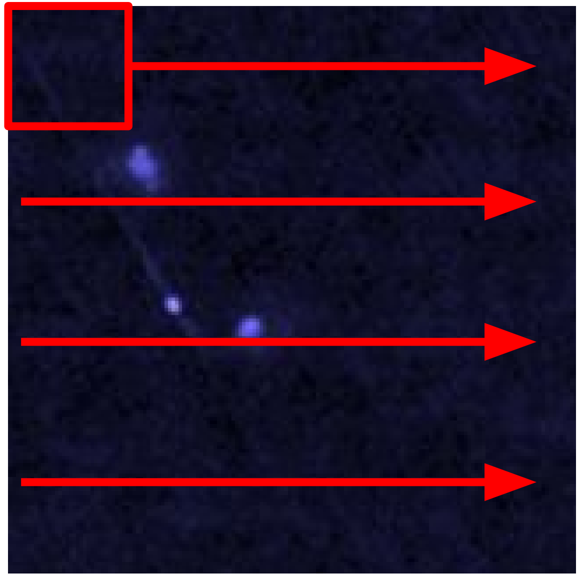
The Galaxy Classification Task

Radio cross-identification in a machine learning context

- Object localisation
  - Finding black holes on the sky
  - Candidate objects from WISE
- Classification
  - Binary classification
  - Logistic regression
  - Expert labels as groundtruth
- Feature selection
  - What does infrared tell us?
  - What does radio tell us?
  - Convolutional neural networks

Finding black holes as object localisation

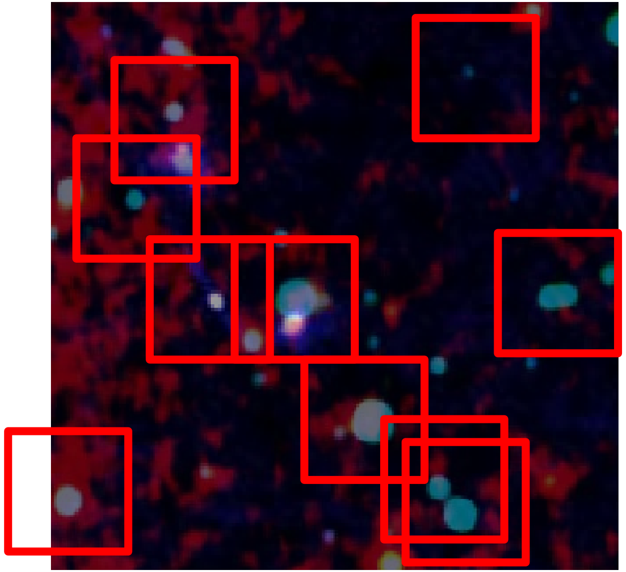
- First attempt:
    - Given an image of black hole jets, check each square patch to see if the black hole is located there
    - "Classic" technique from object localisation, a machine learning problem
    - Not terribly efficient



Scanning to find the black hole.
Image: FIRST

Making use of candidate galaxies

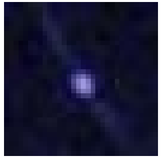
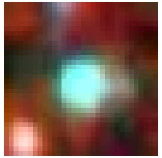
- Second attempt:
  - Given an image of black hole jets, check each *galaxy* in that image to see if it looks like it contains the black hole
  - Get galaxies from an infrared survey (e.g. WISE)
  - Much more efficient!
  - Problems...
    - Scale
    - Patch size
    - Galaxies don’t always show up in infrared

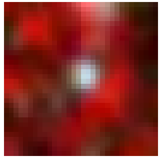


Candidate host galaxies.

Image: FIRST/WISE

A classification problem

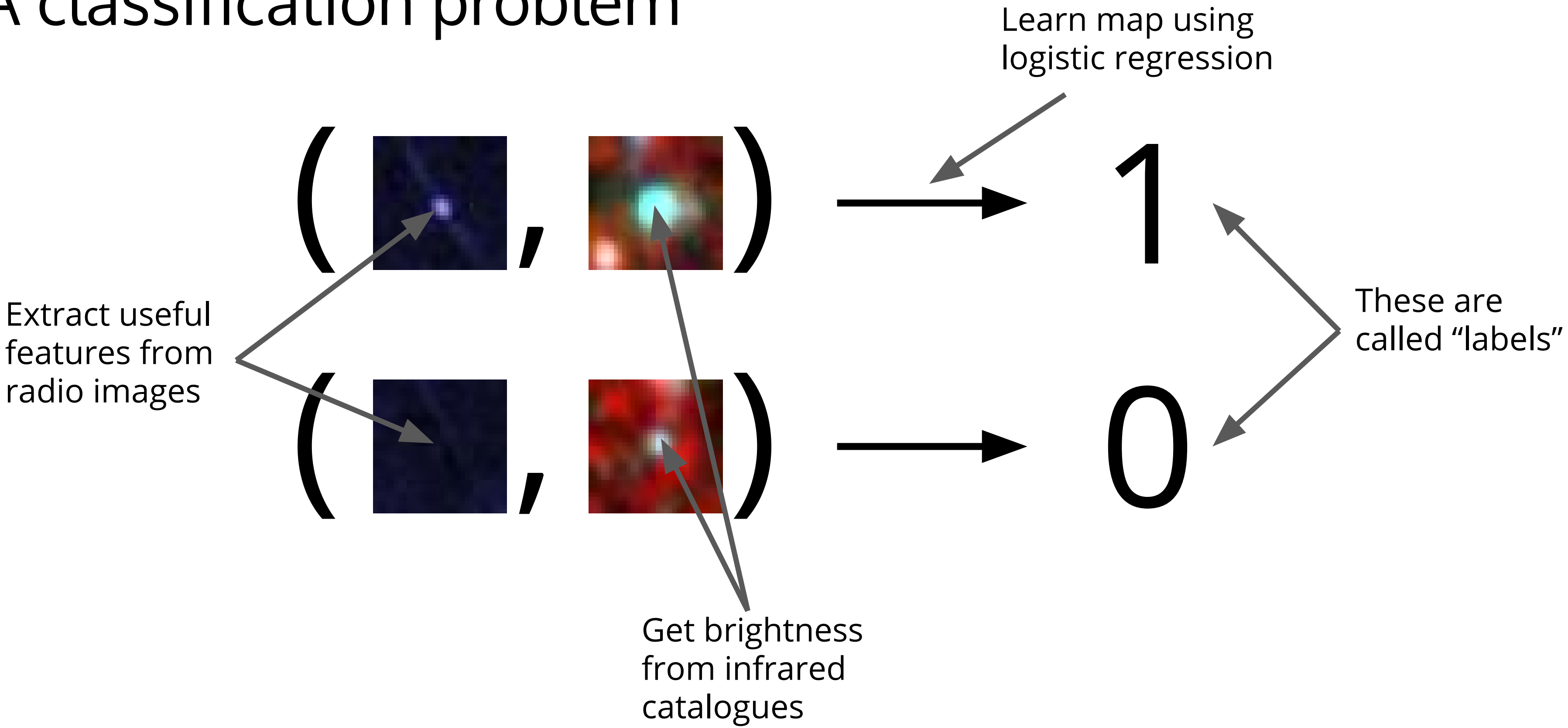
(, ) $\longrightarrow$ 1

(, ) $\longrightarrow$ 0

Representation of galaxy

Whether galaxy has an active galactic nucleus

A classification problem



Logistic regression

- We want to approximate the map:

galaxy → whether the galaxy contains a black hole

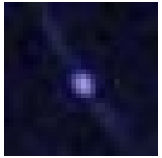
- Logistic regression model:

$$y(x; w) = (1 + exp(-w^T x))^{-1}$$

  - x is a vector representing the galaxy we are looking at
  - w is a vector of weights that we want to find
  - $y(x; w)$ is the probability that x contains a black hole under the given weights w
  - Find the weights that best approximate the Radio Galaxy Zoo cross-identifications
  - Test performance against expert cross-identifications (Norris et al. 2006)

Representing galaxies as vectors

- Infrared images tell us a few things
  - Star formation
  - Dust
- Radio images tell us most of the story
  - Complexity of the black hole
  - Shape of the jets
  - Location?



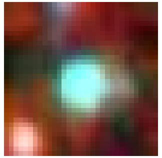
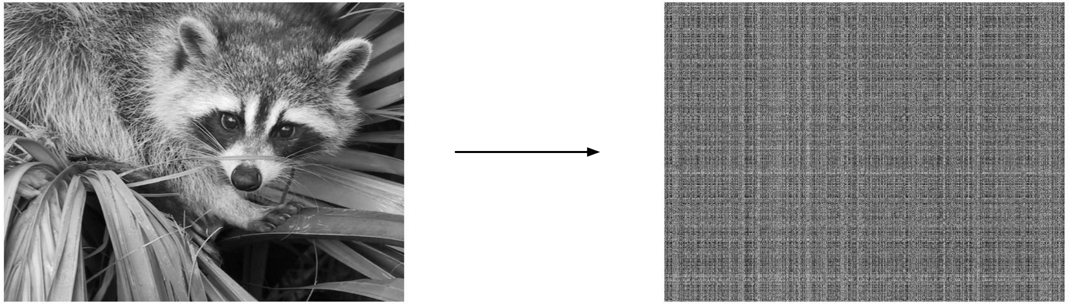


Image patches representing a galaxy.
Image: FIRST/WISE

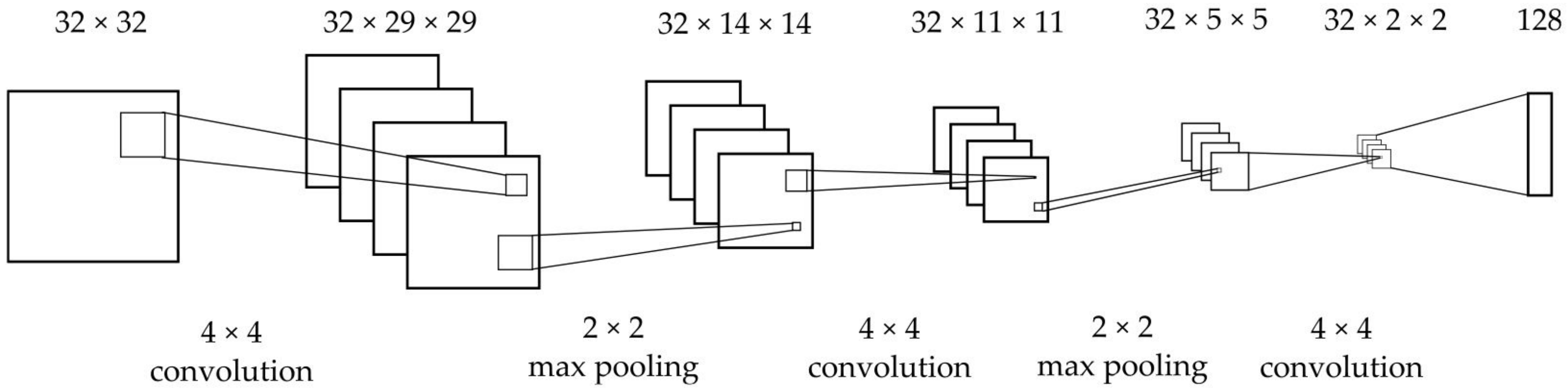
How can we extract information from radio images?

- We could just treat pixels as independent values
  - The location of pixels in images *matters*
  - Treating pixels independently would lose much of the information



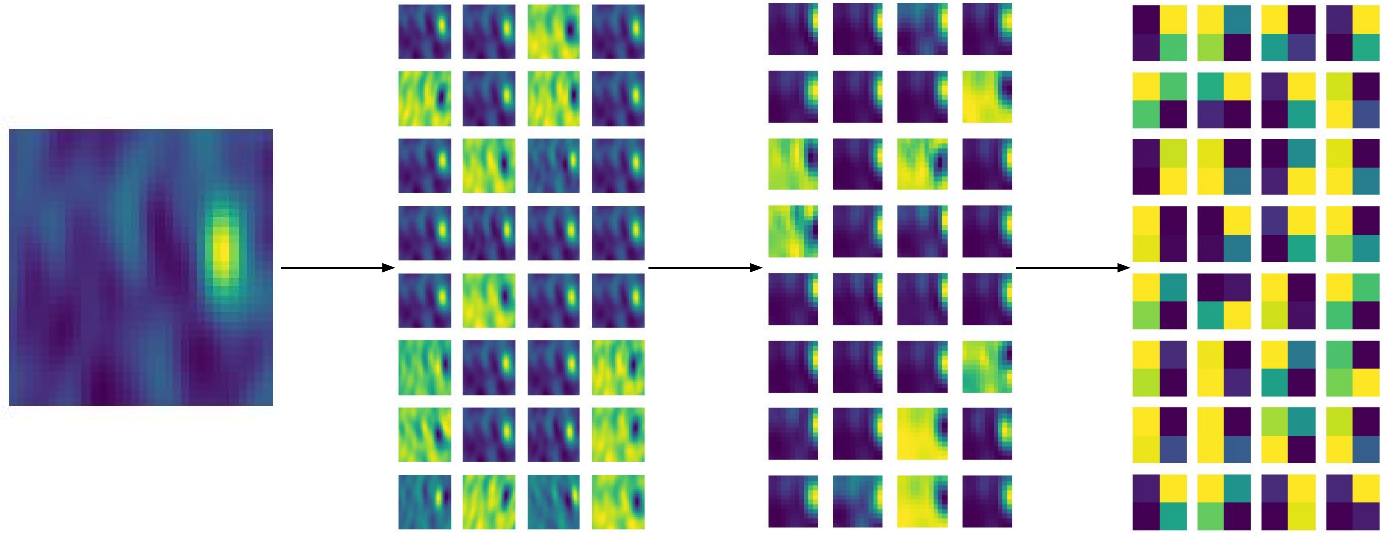
Convolutional neural networks

- Biologically-inspired image feature extraction method
  - Based on the retina and brain
  - Very good results on image classification in recent literature



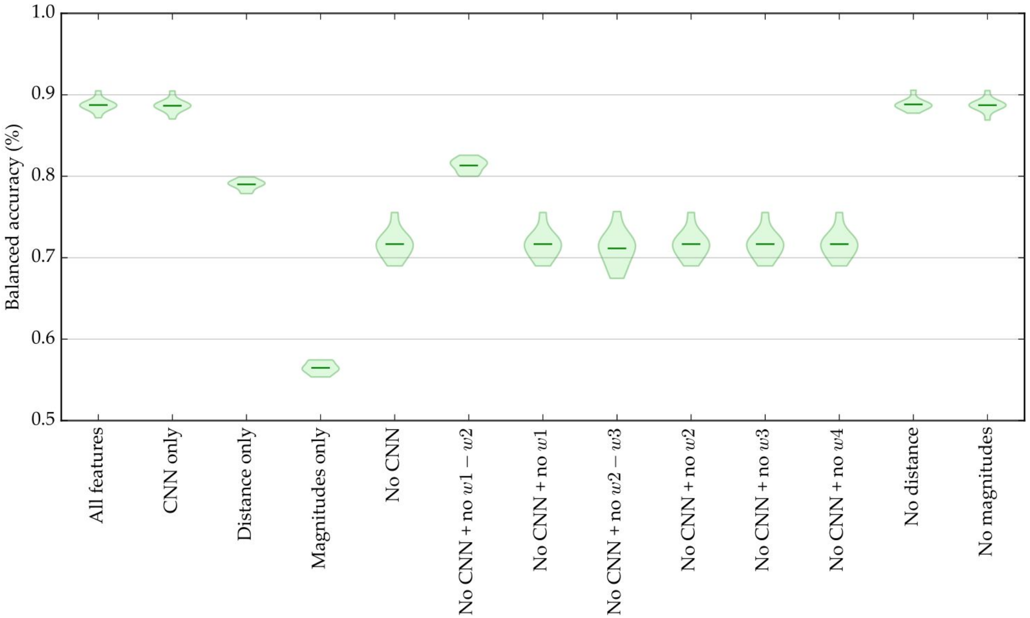
A convolutional neural network.

Radio features



Outputs of each layer of the convolutional neural network, given a radio image as input.
Image: ATLAS

Feature analysis



Accuracy on the galaxy classification task with different features. CNN features dominate.

Learning from the Crowd

Using Radio Galaxy Zoo labels to train the classifier

- Label aggregation
  - Majority vote
  - Raykar et al.
- Classifier performance

Radio Galaxy Zoo label aggregation

- How do we take volunteers' Radio Galaxy Zoo clicks and turn them into a training set?
- Training set is a set of (galaxy, 0 or 1) pairs

volunteers' clicks $\rightarrow$ {(galaxy, 0), (galaxy, 0), (galaxy, 1)}

Radio Galaxy Zoo label aggregation

- First step: Match the click to a corresponding infrared galaxy (easy)
- Next step: Combine these somehow into a single data set (hard)
  - Take the most common label for each galaxy?
  - Take the most common label, but weight different volunteers somehow?
  - Don’t combine them at all and somehow use the labels without combination?

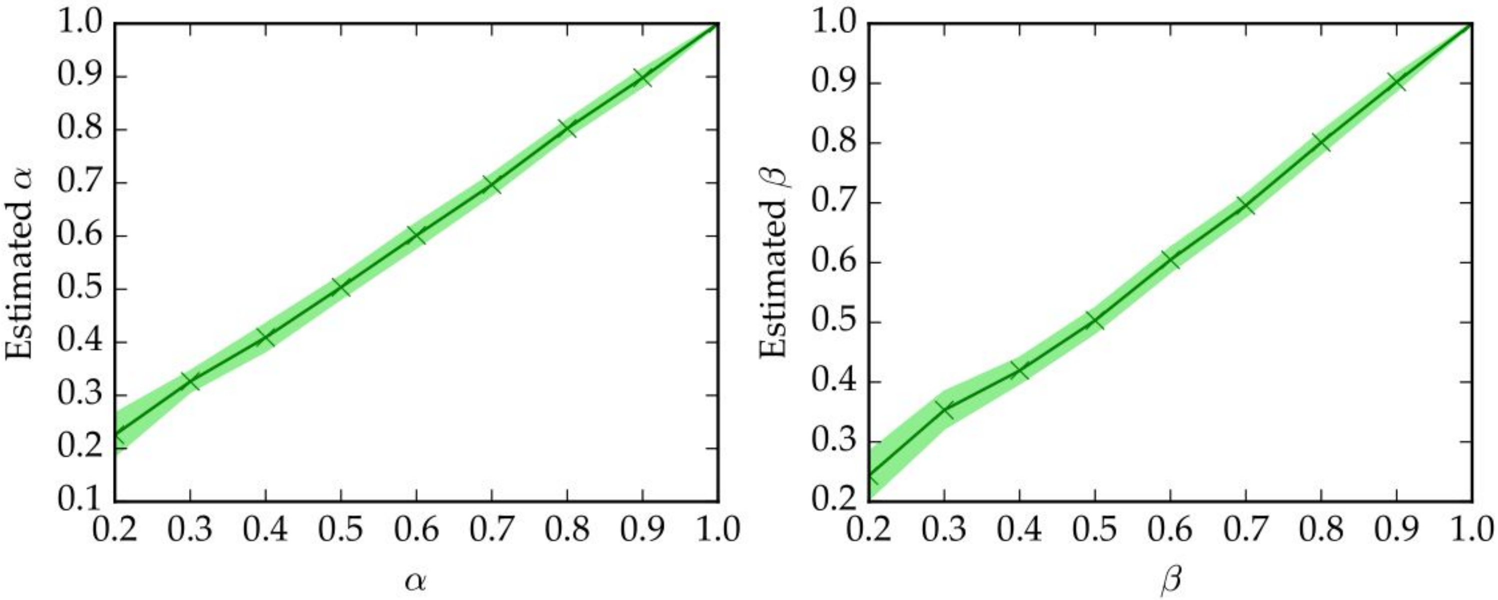
The two options I worked with

- Majority vote
  - For each galaxy, use the most common label
  - Very simple
- Raykar et al. (2010) expectation-maximisation algorithm
  - Attempts to estimate labellers' accuracies
  - *At the same time,* finds the weights for logistic regression
  - Not very simple
  - I produced an open source implementation of this algorithm, available on GitHub

Raykar et al. coin flip model

- Assumes labellers can be described by two biased coins
  - Flip one coin if the true label is 1
  - Flip the other coin if the true label is 0
  - The assigned label is the result of the coin flip
- Try to approximate the biases of the coins (for all labellers at once)
- Try to approximate the galaxy classification map (at the same time!)

Raykar et al. model can recover accuracies



Raykar predicted true positive (α) and true negative (β) rates on a toy data set.

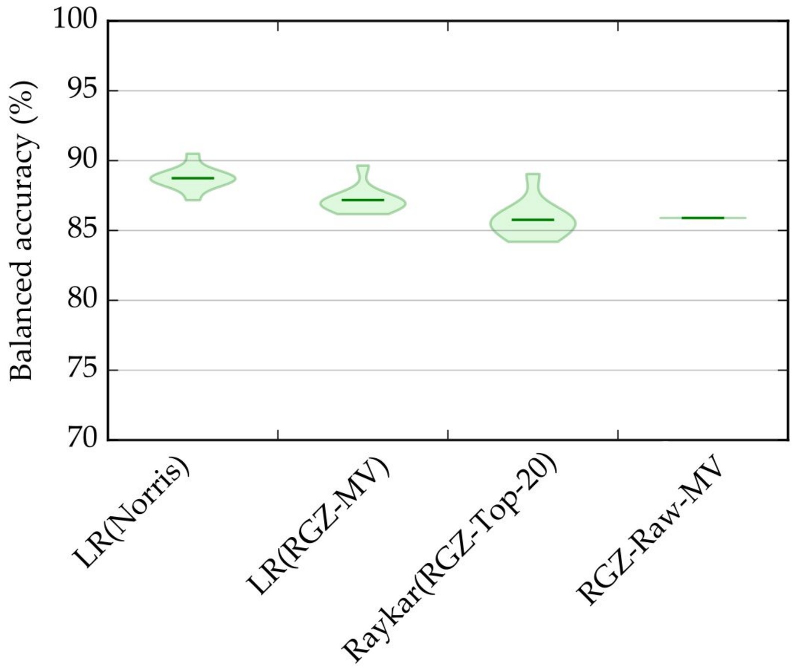
Logistic regression benchmark

- Need a “best-case scenario” for a machine learning algorithm using the features that I decided upon
- Train logistic regression *on the expert cross-identifications*
    - The experts are “perfect” at cross-identification, so this is the best-case for our labels
    - Good approximation to an upper bound of performance

Results

- LR(Norris)
  - Logistic regression + expert labels
- LR(RGZ-MV)
  - Logistic regression + majority vote
- Raykar(RGZ-Top-20)
  - Raykar + 20 best RGZ volunteers
- RGZ-Raw-MV
  - Majority vote of all volunteers (no machine learning)

Experts estimated to be correct ~91% of the time



Observations

- Simple works!
  - Majority vote does well
  - Logistic regression reasonably accurate
- Majority vote outperforms explicitly handling the crowd
  - Maybe Raykar converges to a worse local minimum?
  - Maybe the coin flip model doesn’t describe citizen scientists?
- Crowd labels are useful
  - With current features, training with citizen scientists’ labels is almost as good as training with experts’ labels

Active Learning

Making the best use of volunteers' time

- Existing crowd strategies
  - Brief introduction to methods
  - Where they fall down
- Concept experiment
  - Query-by-committee
  - Results
- Where to next?

Active learning scenario

- Lots of unlabelled data
- Hard to get labels
  - Expensive
  - Time-consuming
  - Intractable
- Examples:
  - Scientific experiments (hypotheses are cheap; experiments are expensive)
  - Text classification (lots of text on the internet; time-consuming to label)
  - Radio cross-identification (many black holes; not many labellers & experts are expensive)

Basic idea

- Train a classifier with our labelled data
- Label all the unlabelled data with the classifier
- Find the *most informative* unlabelled data point
- Request a label from an expert and add it to the labelled data
- Repeat!

- Good theoretical motivation (e.g. Angluin 1988, Dasgupta 2011)
- Good experimental results (e.g. Lewis & Gale 1994, Cohn et al. 1996)

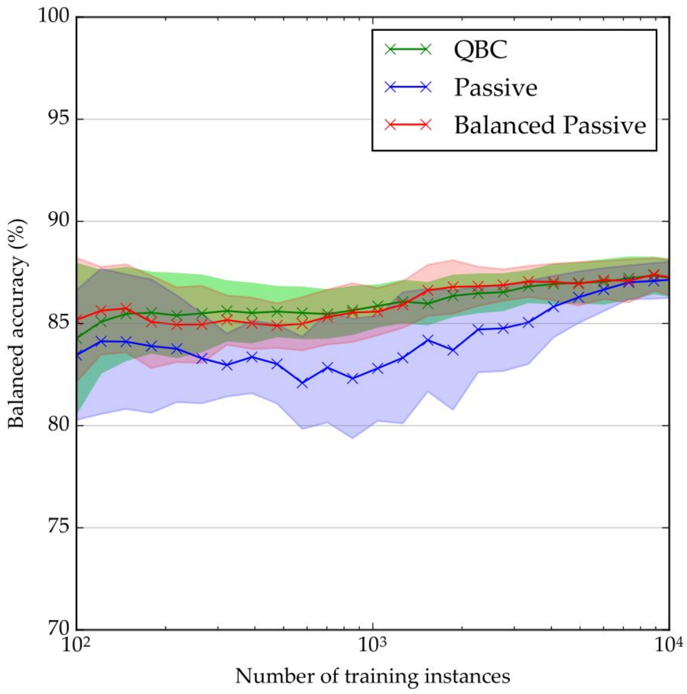
Query-by-committee (QBC)

- Train a number of classifiers on slightly different data
- “Committee” of classifiers votes on each example to label
- The less agreement the committee has on the label of a data point, the more informative that data point must be

- Intuitive
- Easy to implement

QBC on the galaxy classification task

- Committee of 20 logistic regression
  - Initialise with 100 random galaxies (with approximately 7% black holes)
  - Committee is allowed to choose subsequent data points to label
  - Labels are given by experts
- QBC outperforms random sampling
  - Evidence that QBC is effectively sampling balanced classes (i.e. equally sampling black holes and non-black holes)



Applications to cross-identification

- Could help us find “interesting” astronomical objects
- Could dramatically reduce the number of cross-identifications needed to accurately cross-identify all radio objects
- What if instead of experts, we queried citizen scientists?
    - Can we crowdsource active learning?
    - Can we use active learning with citizen science?
    - Are these different questions?

Can we apply active learning to crowdsourcing?

- Yes!
    - Yan et al. (2011) choose a data point *and* a labeller, modelling labellers like Raykar et al.
    - Mozafari et al. (2012) partition the data and only query specific labellers for each partition
    - Nguyen et al. (2015) combine experts with the crowd, choosing which to query each time
- But the crowd can get it wrong
    - Yan, Mozafari, and Nguyen try to incorporate labeller noise models
    - Can we re-request a data point? (e.g. Sheng et al. 2008 and Lin et al. 2016)

Can we apply active learning to citizen science?

- Maybe!
  - Treat citizen science as a crowdsourcing scenario
  - Apply your choice of Yan/Mozafari/Ngyuen
- Existing models assume that we can choose which labeller queries a point
  - Not the case in citizen science
  - Volunteers query *us* for data points to label, rather than us querying the volunteers
  - Can't choose who labels a specific data point in general
- Citizen science generally has many labellers
  - Crowdsourcing a task on Mechanical Turk might give you ~10 labellers
  - Radio Galaxy Zoo has >2000 labellers on just the 2600 radio objects I looked at
  - Existing algorithms are very slow or perform poorly on such large numbers
  - Label matrix is sparse — most labellers don't label most data points

Conclusion

- As we make bigger and better telescopes and surveys, radio cross-identification is growing to be an intractable problem
- I cast the radio cross-identification problem as a machine learning task
    - The problem can now be expressed in a machine learning context
    - I represented galaxies with features automatically extracted from radio images
- I developed a classifier for the task which recovers good accuracy
    - I trained the classifier with a never-before-used dataset (RGZ)
    - Fully naïve (i.e. not astronomical model-based) approach
- I found preliminary results on active learning for astronomy
- I produced an open source implementation of everything discussed here

References

Banfield et al. 2015. *Radio Galaxy Zoo: host galaxies and radio morphologies derived from visual inspection*.

Raykar et al. 2010. *Learning from crowds.*

Norris et al. 2011. *EMU: Evolutionary Map of the Universe.*

Norris et al. 2006. *Deep ATLAS radio observations of the CDFS-SWIRE field.*

Yan et al. 2011. *Active learning from crowds.*

Angluin 1988. *Queries and concept learning.*

Dasgupta 2011. *Two faces of active learning.*

Lewis & Gale 1994. *A sequential algorithm for training text classifiers.*

Cohn et al. 1996. *Active learning with statistical models.*