MANCHESTER 1824

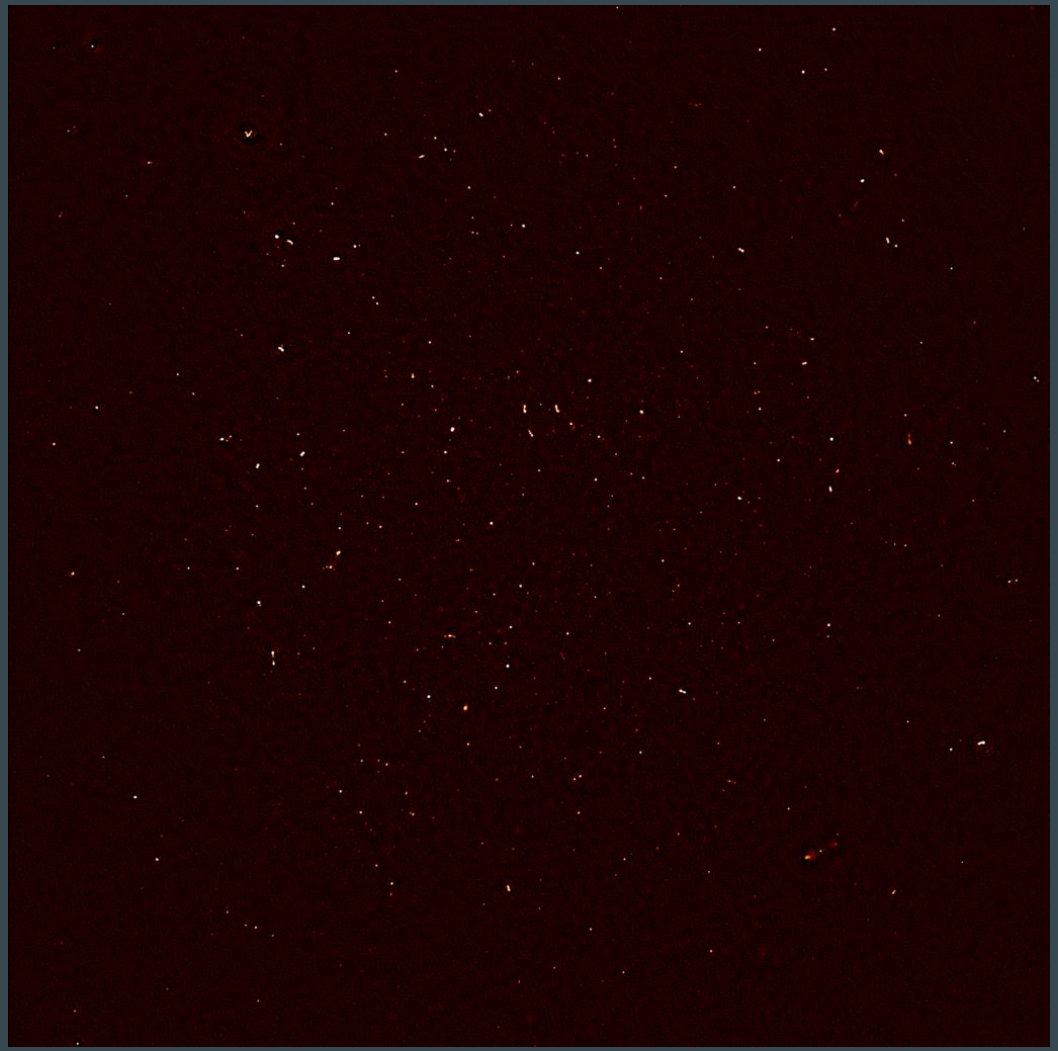
The University of Manchester

Image Based Galaxy Classification in Radio Astronomy

Hayden Rampadarath

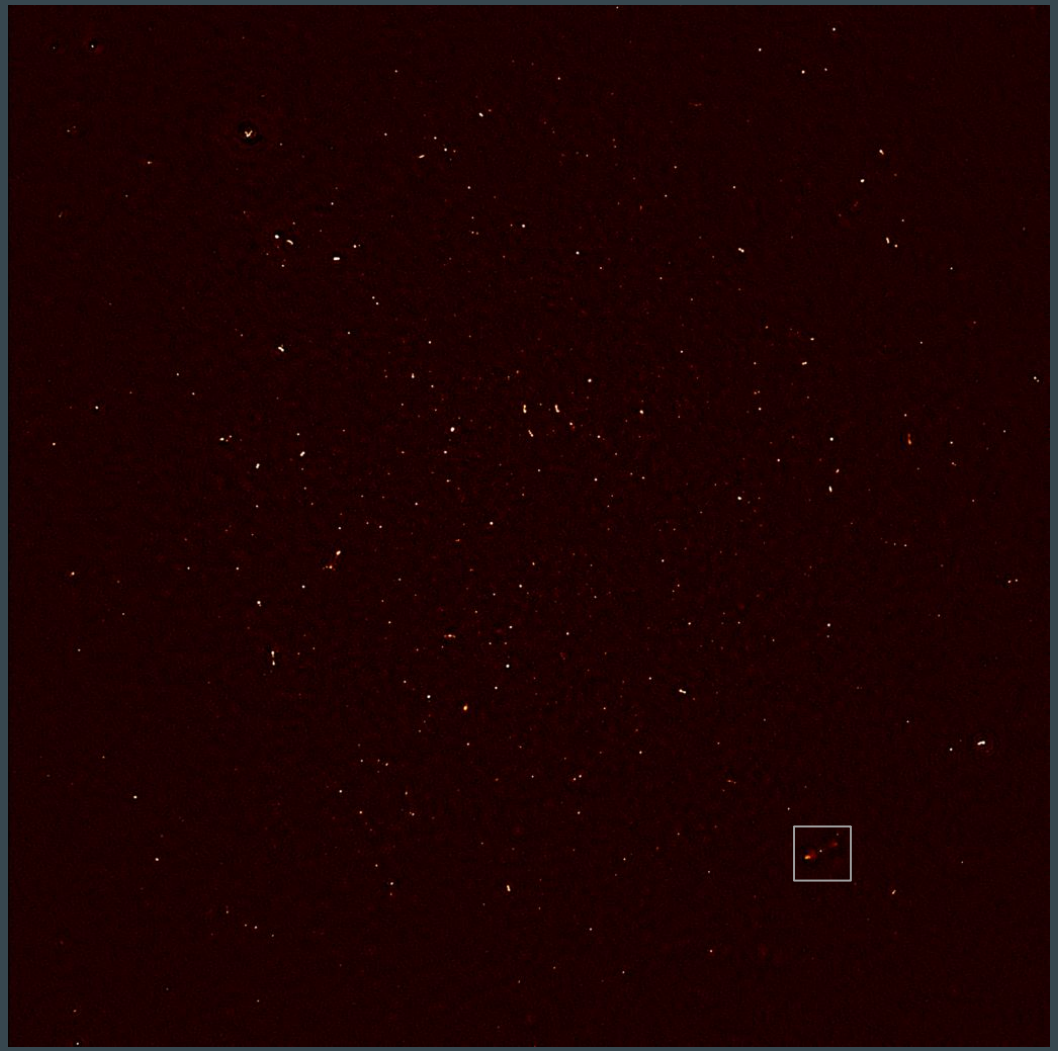
Nick Wrigley, Erselaan Kazmi (Nuffield Summer Student), & Tom Muxlow

4IR Workshop on Class Imbalance in Machine Learning Classification 22nd November 2017



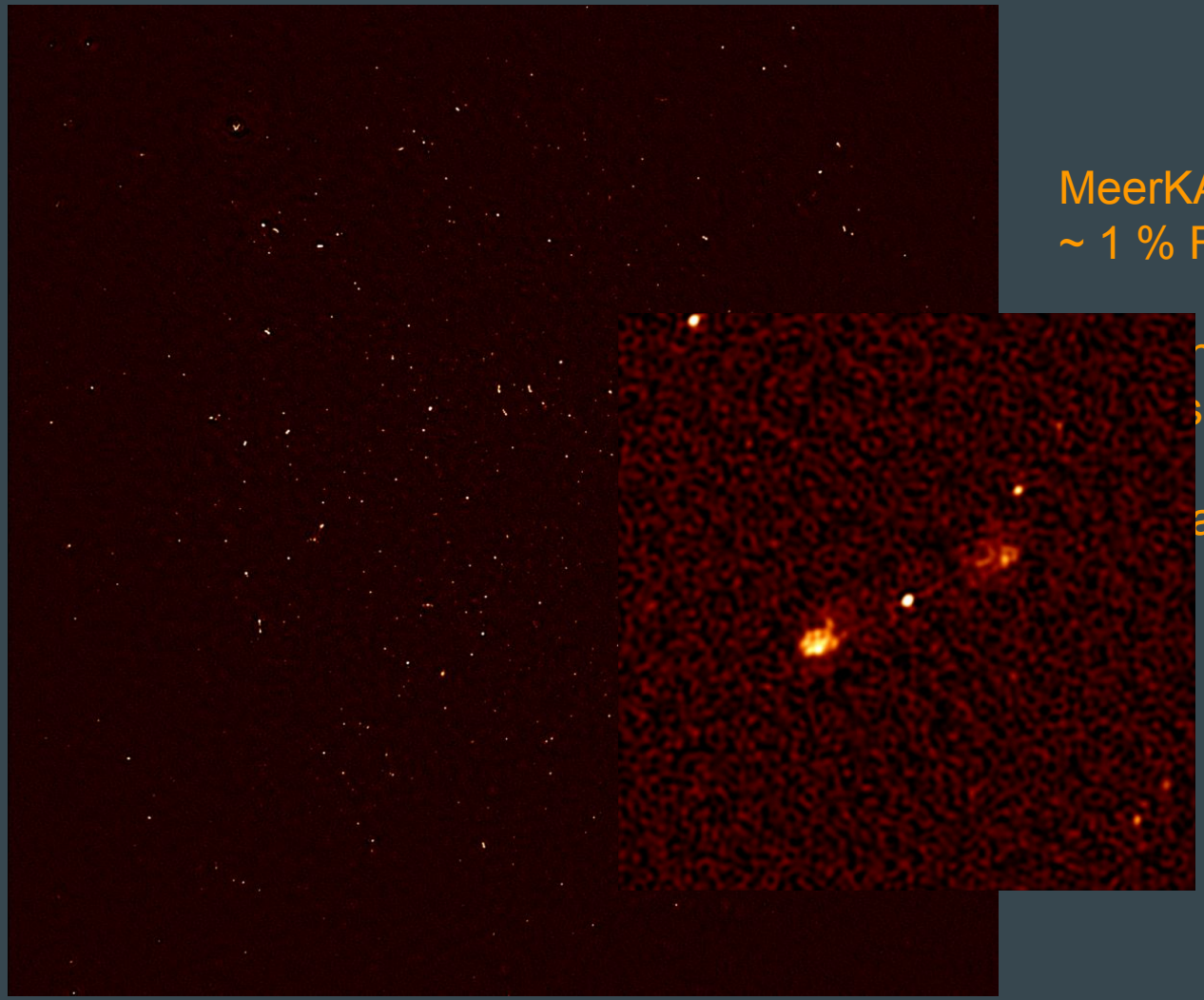
MeerKAT First Light image
~ 1 % Full MeerKAT.

More than 1300 individual objects – galaxies in the distant universe – are seen in this image.



MeerKAT First Light image.
~ 1 % Full MeerKAT

More than 1300 individual objects – galaxies in the distant universe – are seen in this image.

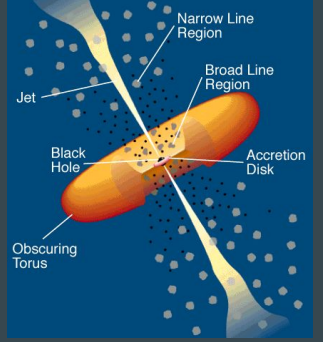


MeerKAT First Light image
~ 1 % Full MeerKAT

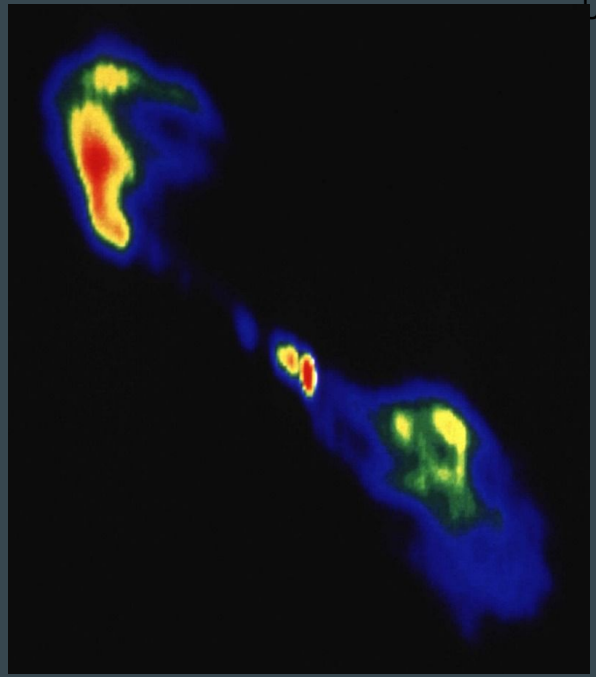
nan 1300 individual
s – galaxies in the
universe – are seen in
age.

Dominant radio emission process in Galaxies (1)

Supermassive Black Holes or AGNs (Active Galactic Nuclei) - non-thermal emission



Urry & Padovani 1995

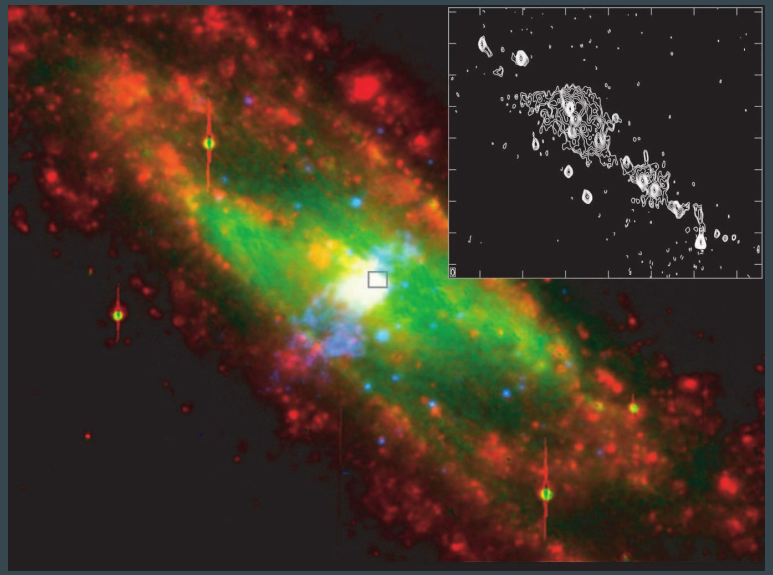


AGN Radio Galaxy Centaurus A

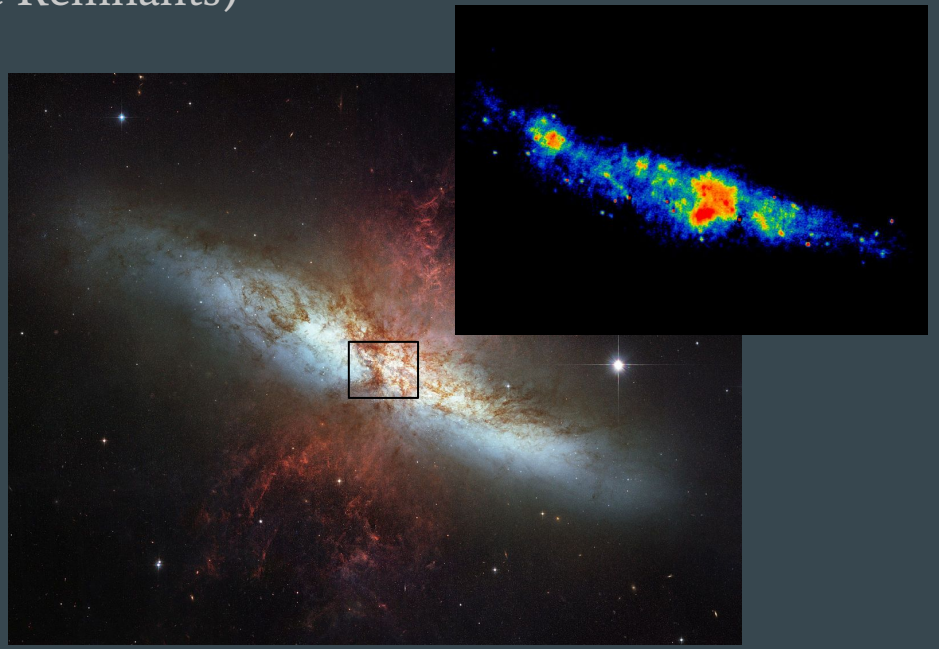
Image courtesy of NRAO/AUI - http://images.nrao.edu/10

Dominant radio emission process in Galaxies (2)

Star-Formation Processes (i.e. Starburst) - Thermal (ionised Hydrogen regions) and non-thermal (Supernovae and Supernovae Remnants)



Starburst galaxy NGC253. Image by E. Lenc



Starburst galaxy M82. Image by T. Muxlow

Separating AGN from Star formation

1. A reliable cosmic census of starburst and AGN populations
    a. radio sources are dominated by AGN at flux densities above ~200μ Jy (e.g. Bonzini + 2013)
    b. star-forming sources increasingly important at lower fluxes (e.g. Muxlow et al. 2005, Padovanni+ 2016
2. AGN/host galaxy interaction
    a. Understand the relationship between star formation and AGN activity
3. Obscured AGN activity
    a. Radio wavelengths are not plagued (as much) by obscuration as other wavelengths!

Requirements: Need to be able to separate AGN galaxies from SF Galaxies

Need to resolve each galaxy - high resolution & population statistics

The University of Manchester
Jodrell Bank Observatory

MANCHESTER 1824

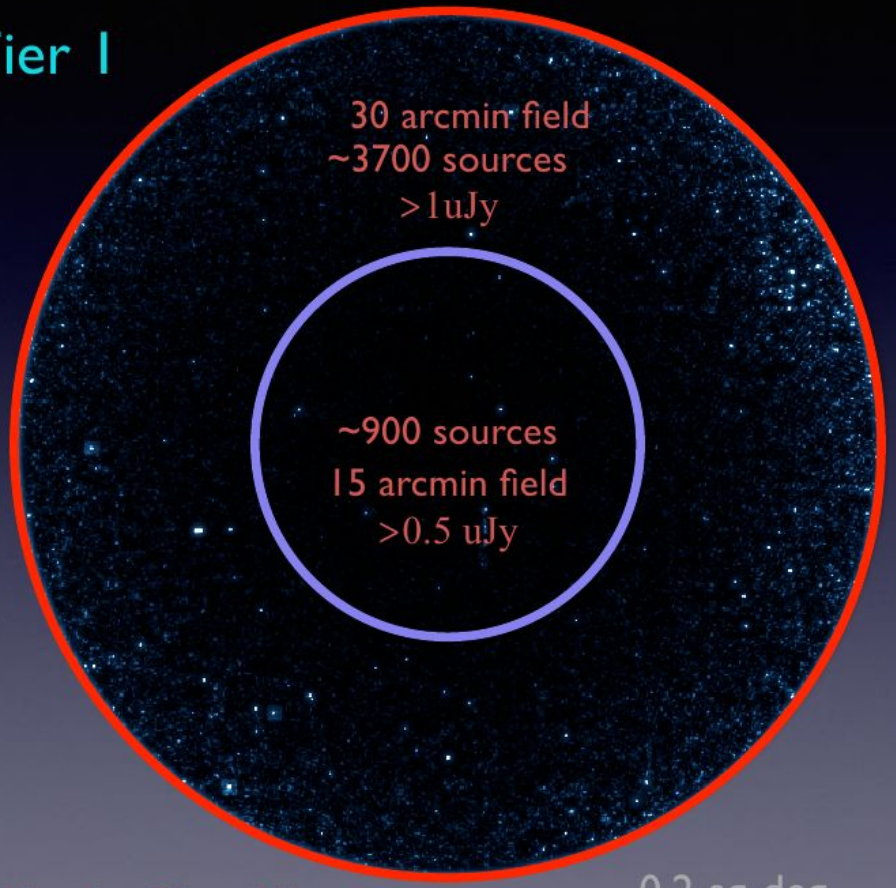
eMERGE

Tier 1

e-MERLIN Galaxy Evolution Survey

1.25 - 1.75 GHz
(L-band) >360 hrs
resolution~150mas

4.5 - 6.5 GHz
(C-band) > 250 hrs
resolution~50mas



0.2 sq deg

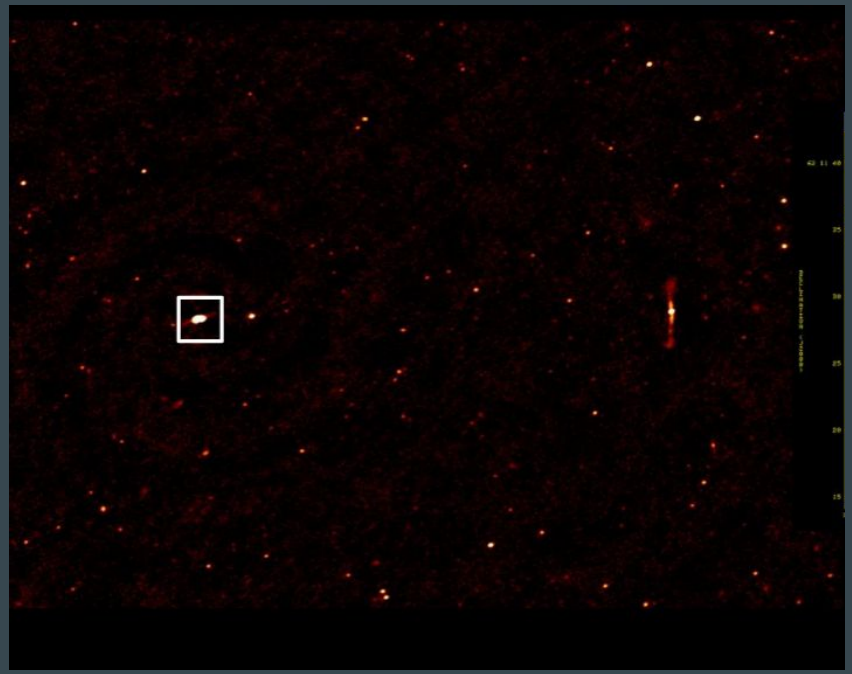
+>40 hrs Jansky Very Large Array L-band + C-band data

Classifying Galaxies in the Radio - Old School!

Classifying Galaxies in the Radio - Old School!

1. Select an object

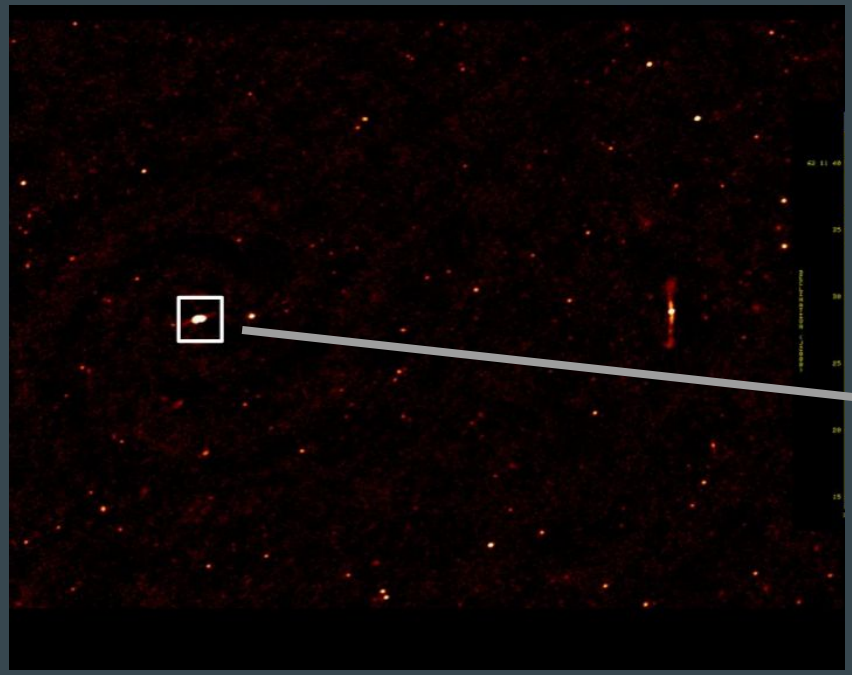
Images courtesy T. Muxlow

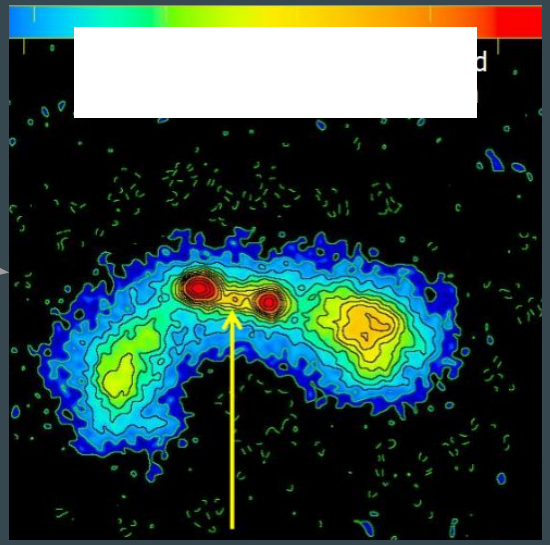


Classifying Galaxies in the Radio - Old School!

1. Select another object
2. Obtain its Morphology and spectrum (brightness vs frequency)

Images courtesy T. Muxlow

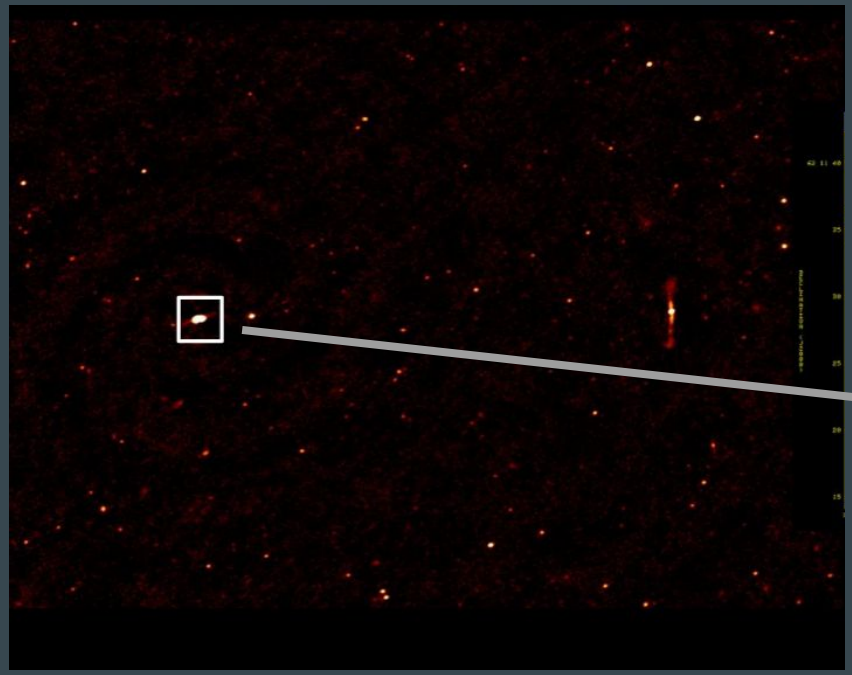


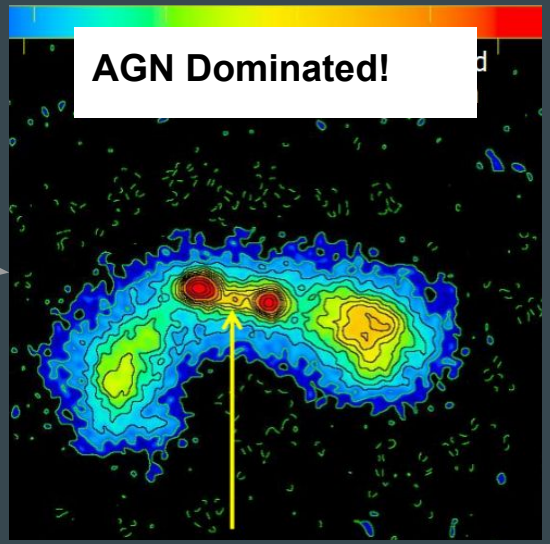


Classifying Galaxies in the Radio - Old School!

1. Select another object
2. Obtain its Morphology and spectrum (brightness vs frequency)
3. If spectrum is flat
4. & has a compact core with an extended feature (optional)
5. Confirm with other frequencies

Images courtesy T. Muxlow

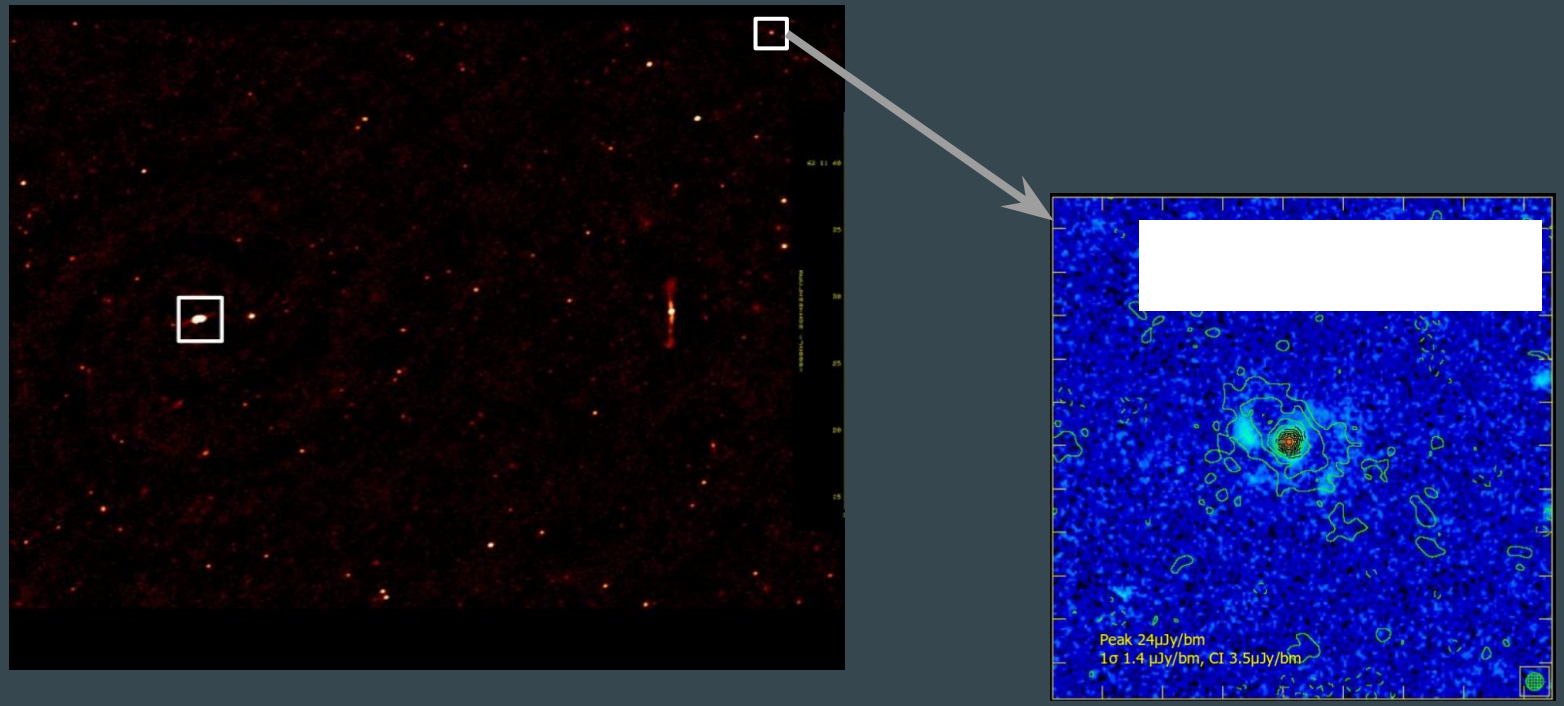




Classifying Galaxies in the Radio - Old School!

1. Select an object
2. Obtain its Morphology and spectrum (brightness vs frequency)

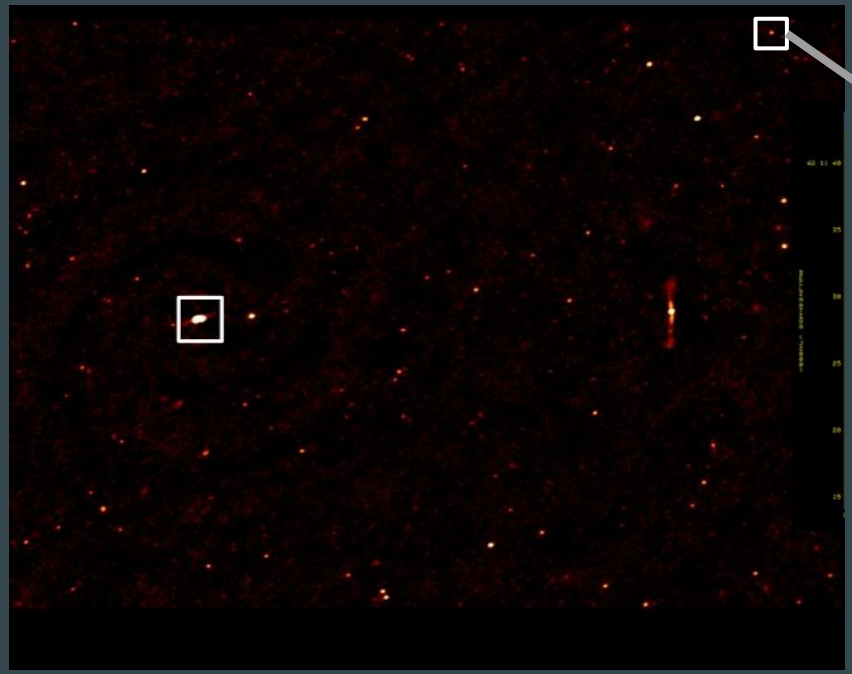
Images courtesy T. Muxlow

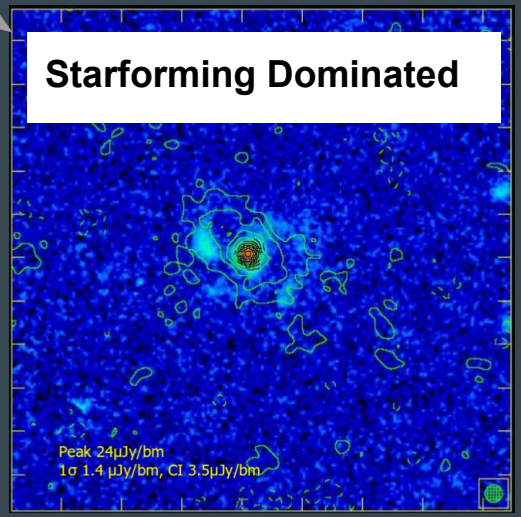


Classifying Galaxies in the Radio - Old School!

1. Select an object
2. Obtain its Morphology and spectrum (brightness vs frequency)
3. If Spectrum is steep (>-0.5)
4. & is extended on sub-galactic scales (i.e. fluffy)
5. Confirm with other frequencies

Images courtesy T. Muxlow



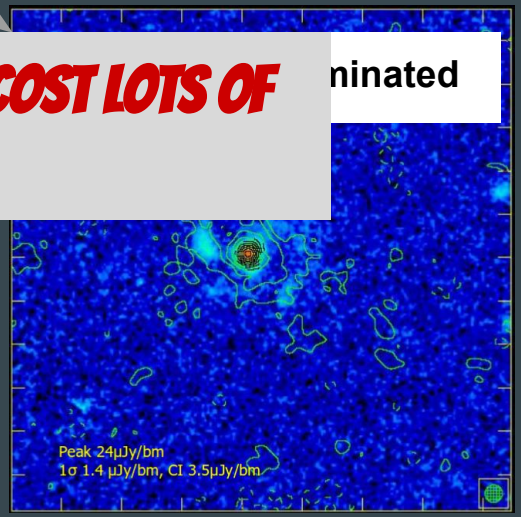


Classifying Galaxies in the Radio - Old School!

1. Select an object
2. Obtain its Morphology and spectrum (brightness vs frequency)
3. If Spectrum is steep (>-0.5)
4. & is extended on sub-galactic scales (i.e. fluffy)
5. Confirm with other frequencies

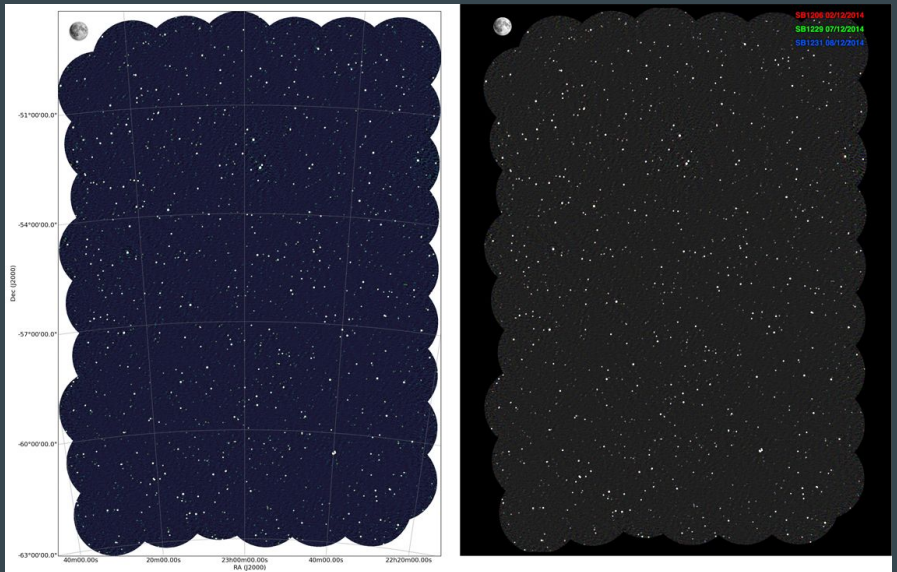
Images courtesy T. Muxlow

REPEAT FOR ALL GALAXIES – COST LOTS OF TIME



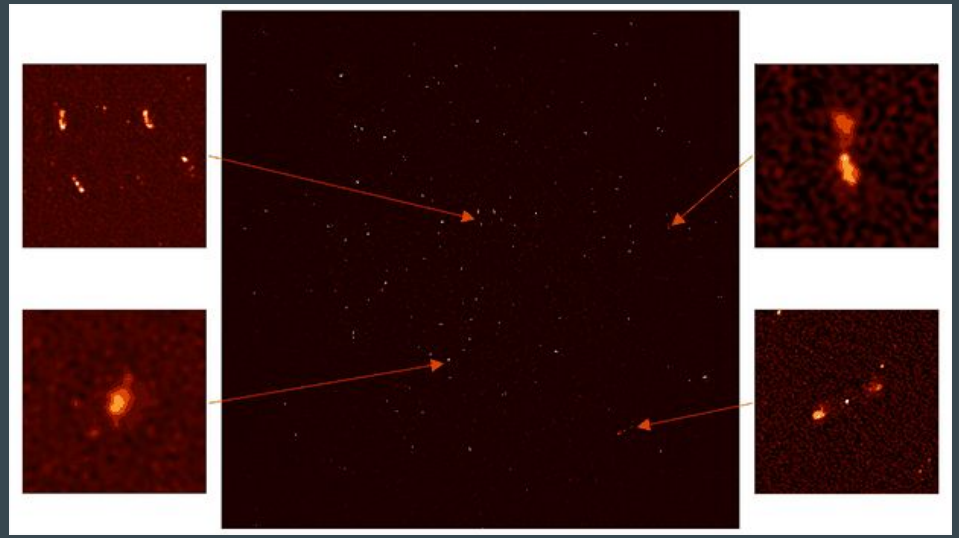
Next Generation of Instruments - Old school method Impossible!

ASKAP BETA Array



The latest image of the Tucana constellation was produced during ASKAP commissioning activities with the Boolardy Engineering Test Array (BETA) telescope. The resulting image is 150 square degrees (over three times the size of that previously produced), and contains around 2000 sources brighter than 5-sigma. The images shown above are two different versions of the same observation, on the right is the three colour RGB image referred to in the story text. The full moon is shown here for scale.
Image credit: Keith Bannister (observations), Ian Heywood (calibration and imaging) of CSIRO's ASKAP team.

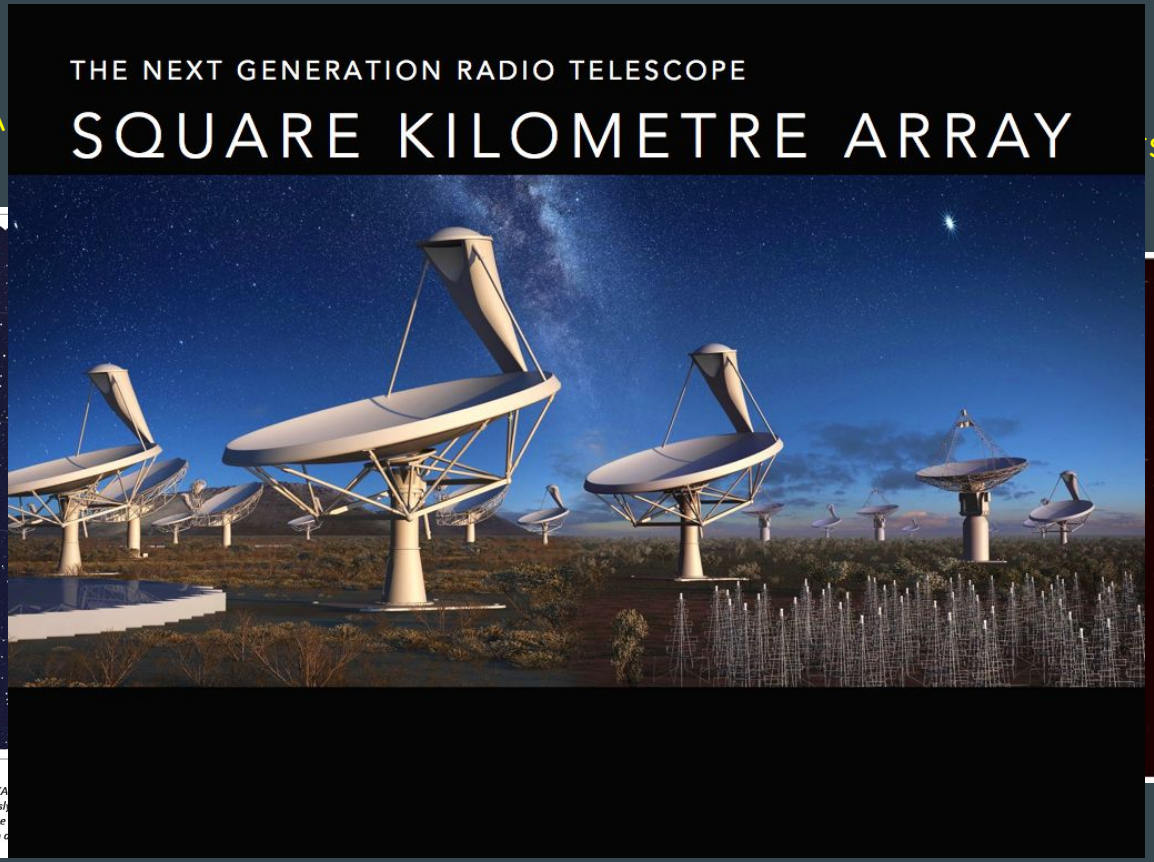
MeerKat radio telescope's First Light image



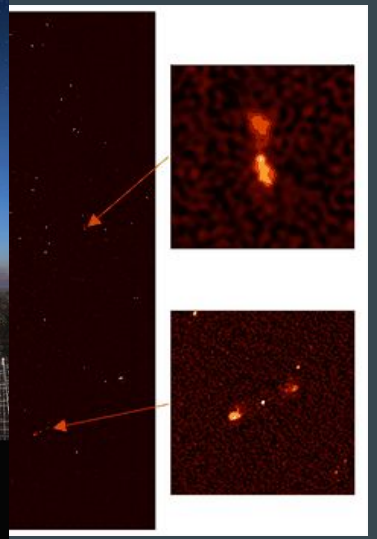
Next Generation of Instruments - Old school method Impossible!

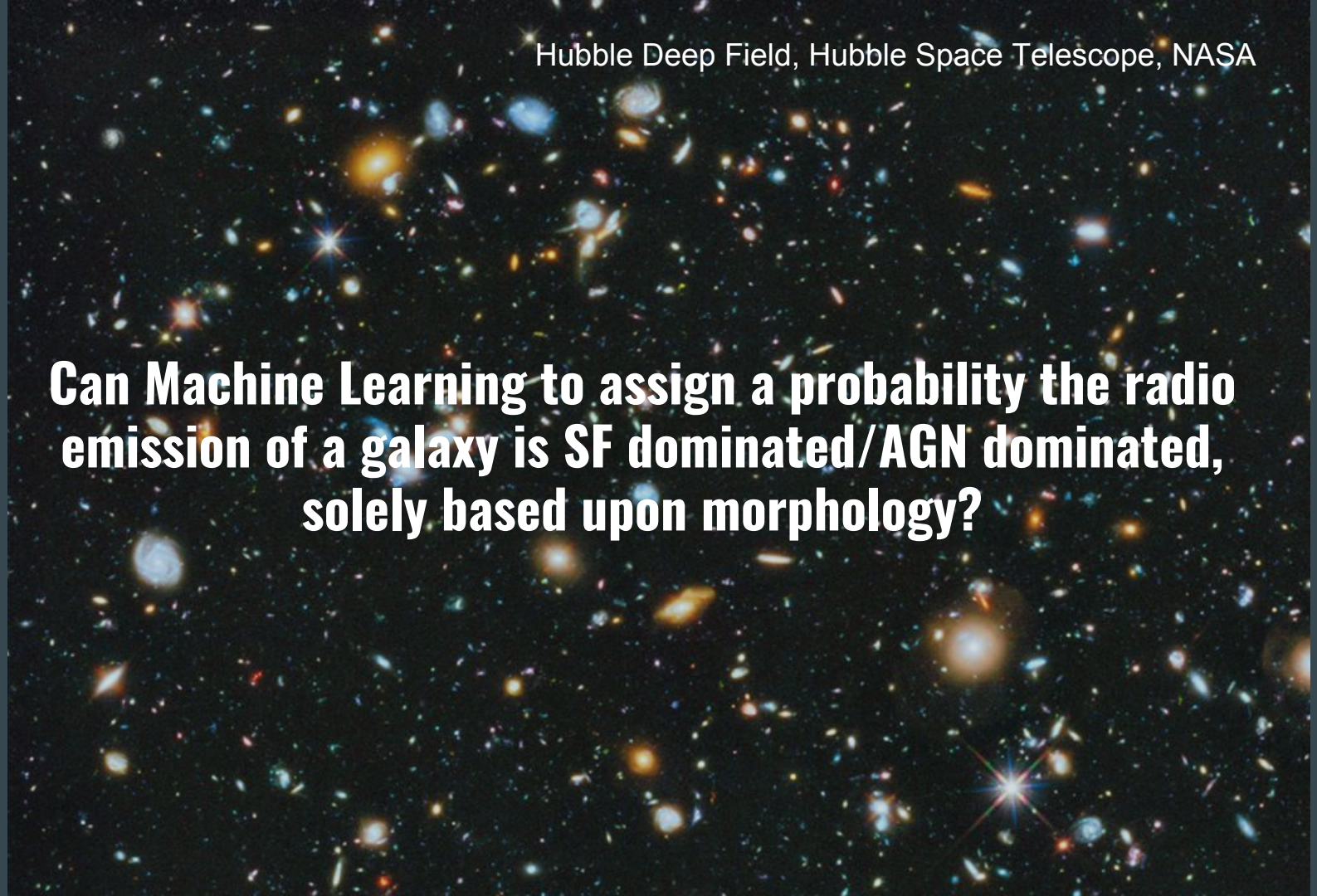
ASKA

The latest image of the Tucana constellation was produced during ASKA
image is 150 square degrees (over three times the size of that previously
two different versions of the same observation, on the right is the three
Image credit: Keith Bannister (observations), Ian Heywood (calibration a

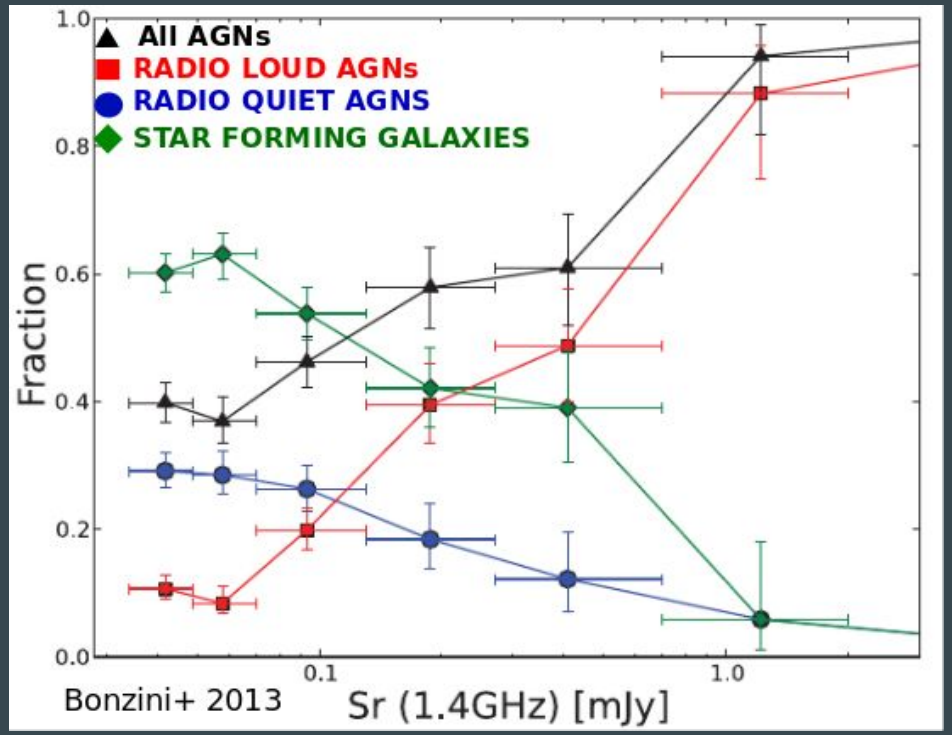


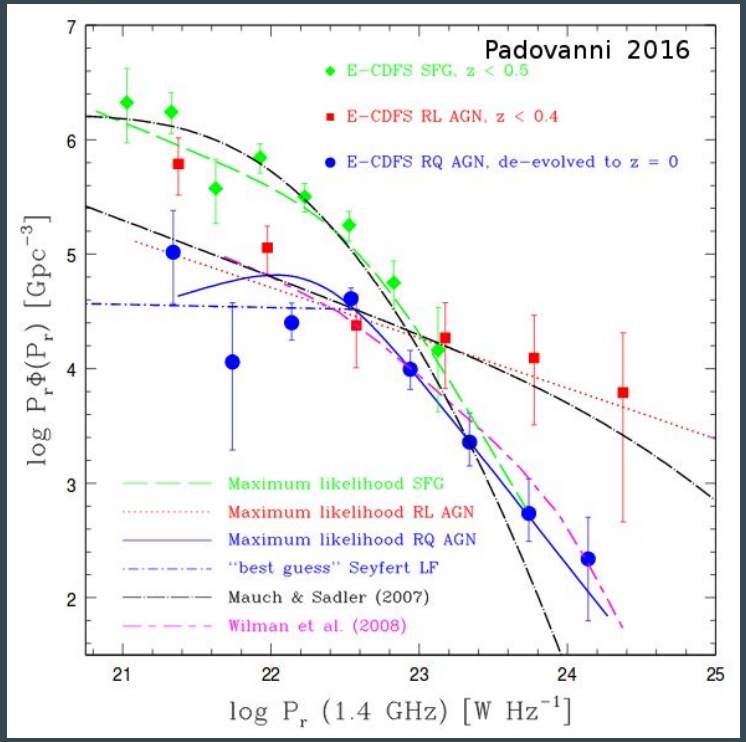
st Light image




Hubble Deep Field, Hubble Space Telescope, NASA
Can Machine Learning to assign a probability the radio emission of a galaxy is SF dominated/AGN dominated, solely based upon morphology?

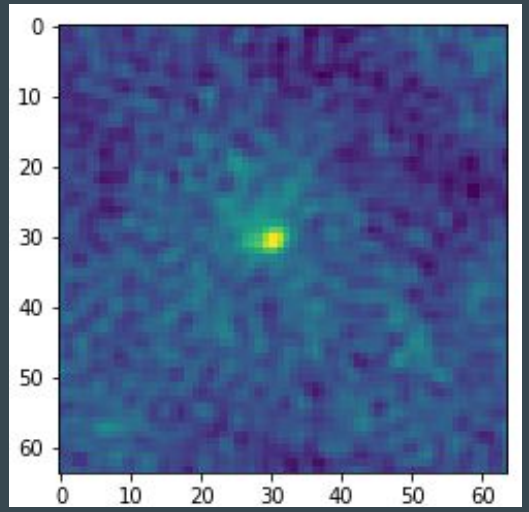
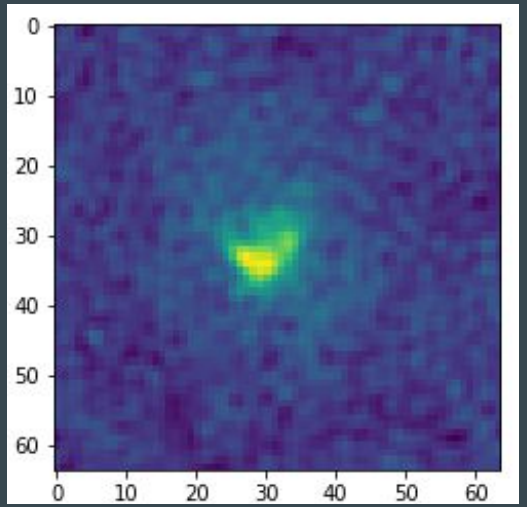
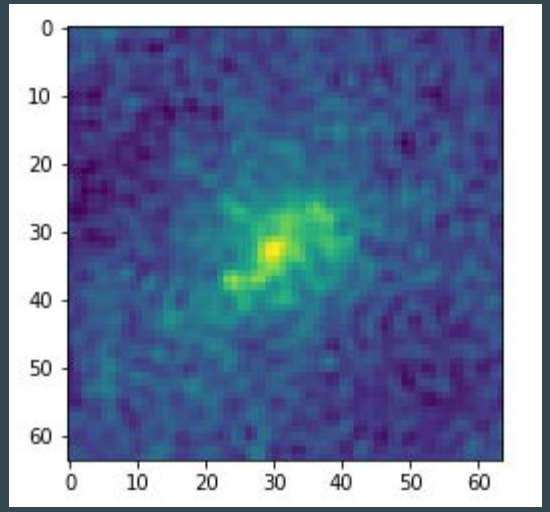
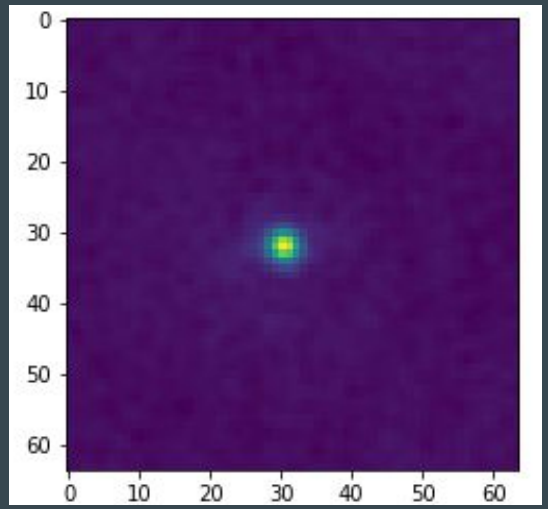
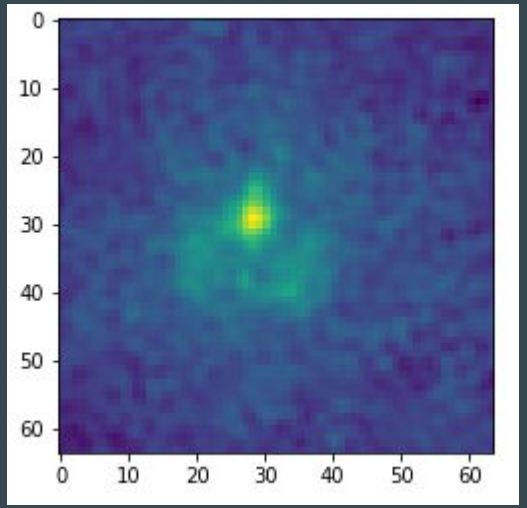
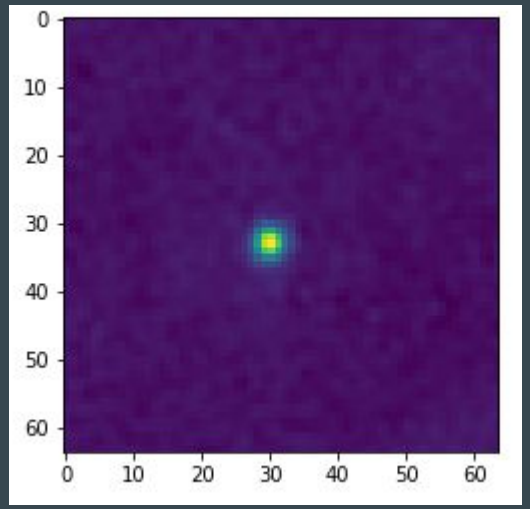
Imbalanced problem?





The Data

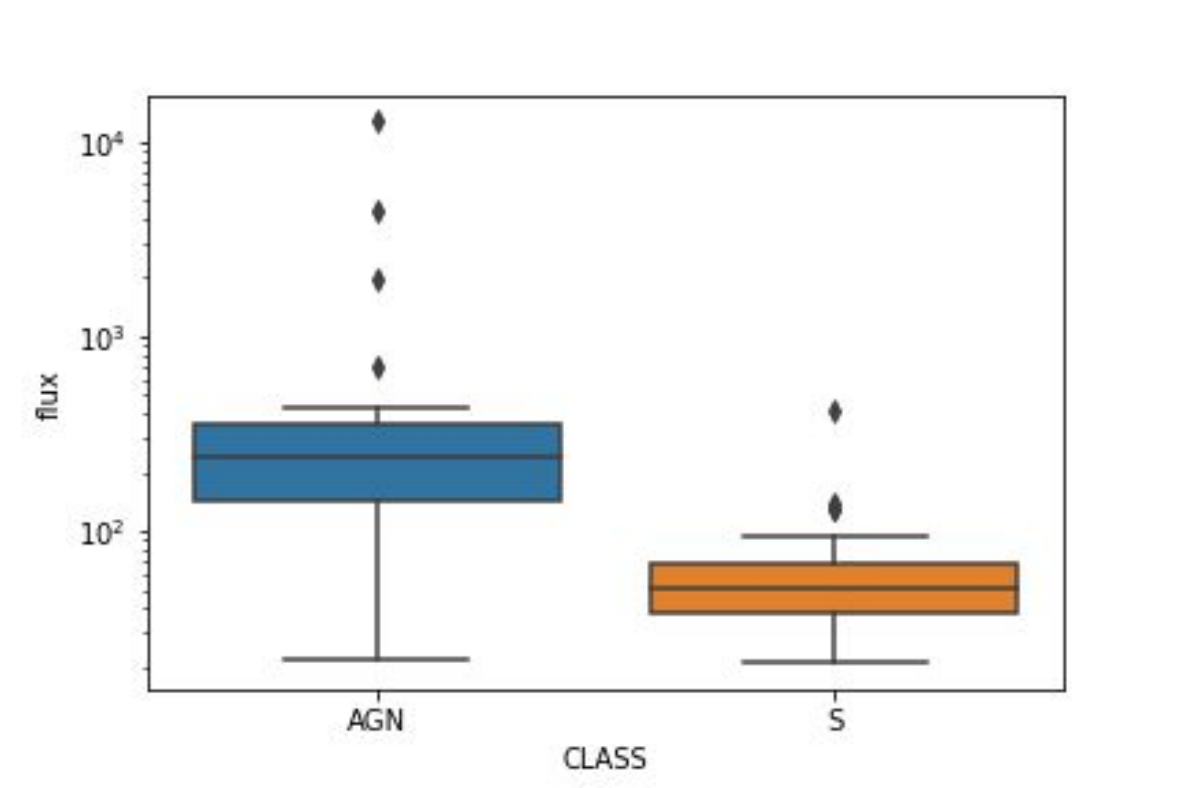
1. Images of 340 galaxies with the e-MERLIN array in the HDF
    a. 64 x 64 pixels images in FITS format
    b. Flatten to form the features - > 4096 features
2. 70 - previously classified by Muxlow+2005
    a. AGN/AGN candidate & Starburst/Starburst candidate
    b. 24 AGN dominated & 46 Starforming dominated galaxies.
3. 270 - Unclassified galaxies


0
10
20
30
40
50
60

0
10
20
30
40
50
60

0
10
20
30
40
50
60

0
10
20
30
40
50
60

0
10
20
30
40
50
60

0
10
20
30
40
50
60

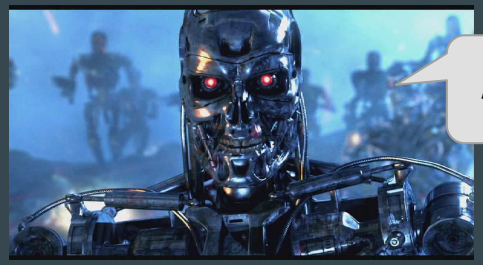
Data Preprocessing

- 70 classified galaxies -> impossible to train and test a model.
- Image augmentation - Simulate different observing conditions
  - Gaussian filters + rotation + noise -> randomise
  - Gaussian Blur using imagaug (OpenCV)
  - Total: ~264 AGN and ~276 Starforming augmented data
- The images were normalised to a peak brightness of 1
- Apply PCA to reduce the number of features that account for > 1% data variance ~ 17 components

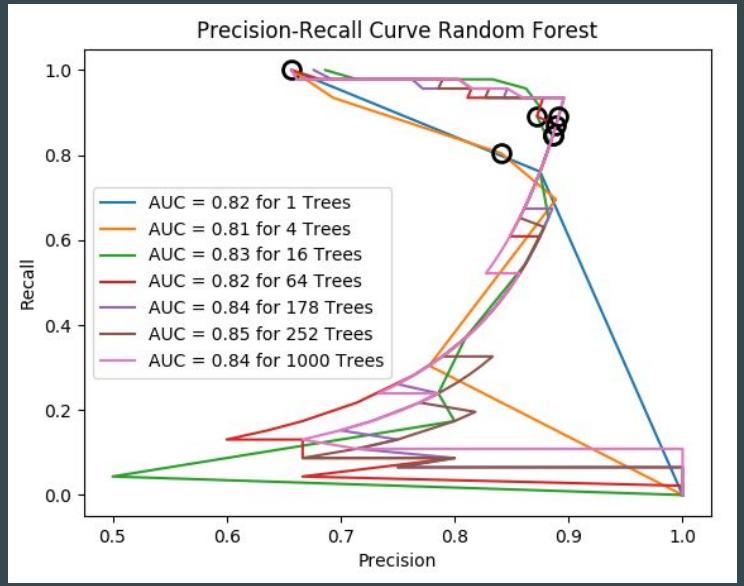
Data Preprocessing

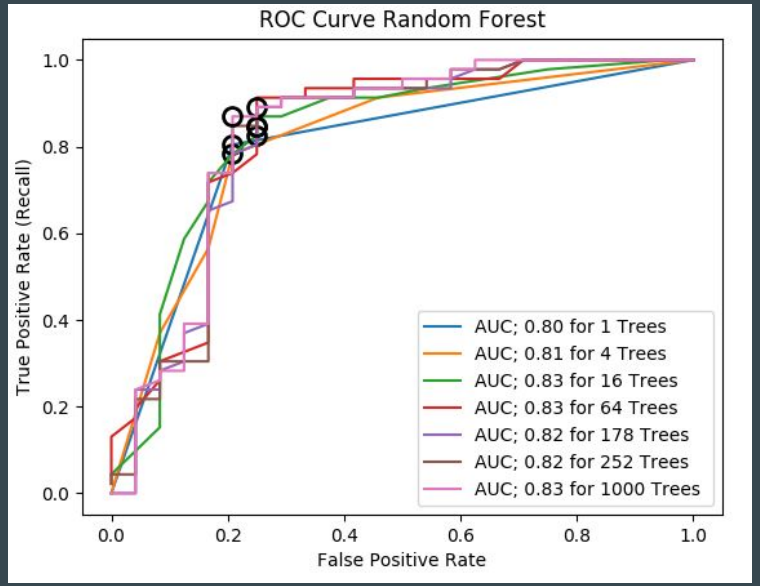


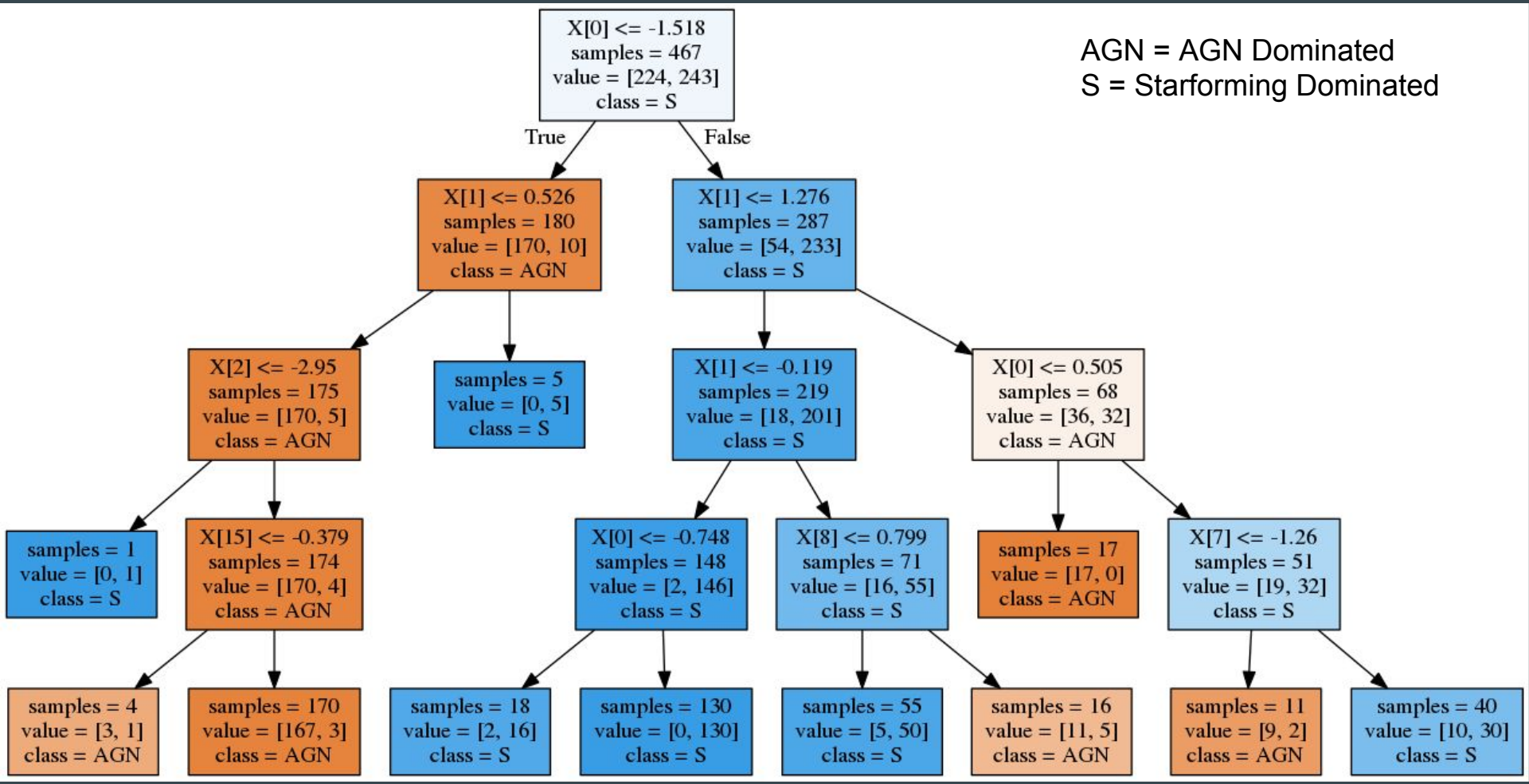
Classifier Training & Testing



- Training data -> Augmented data
- Test data -> Original data
- Tried - Random Forests, SVM, Gradient Boosted Trees and K-NN
- Settled with RF model of 300 Trees






AGN = AGN Dominated
S = Starforming Dominated
X[0] <= -1.518
samples = 467
value = [224, 243]
class = S
True
False
X[1] <= 0.526
samples = 180
value = [170, 10]
class = AGN
X[1] <= 1.276
samples = 287
value = [54, 233]
class = S
X[2] <= -2.95
samples = 175
value = [170, 5]
class = AGN
samples = 5
value = [0, 5]
class = S
X[1] <= -0.119
samples = 219
value = [18, 201]
class = S
X[0] <= 0.505
samples = 68
value = [36, 32]
class = AGN
samples = 1
value = [0, 1]
class = S
X[15] <= -0.379
samples = 174
value = [170, 4]
class = AGN
X[0] <= -0.748
samples = 148
value = [2, 146]
class = S
X[8] <= 0.799
samples = 71
value = [16, 55]
class = S
samples = 17
value = [17, 0]
class = AGN
X[7] <= -1.26
samples = 51
value = [19, 32]
class = S
samples = 4
value = [3, 1]
class = AGN
samples = 170
value = [167, 3]
class = AGN
samples = 18
value = [2, 16]
class = S
samples = 130
value = [0, 130]
class = S
samples = 55
value = [5, 50]
class = S
samples = 16
value = [11, 5]
class = AGN
samples = 11
value = [9, 2]
class = AGN
samples = 40
value = [10, 30]
class = S

Classifier Training & Testing

- Randomness of the data augmentation -> additional uncertainty
- The models were executed 10 times

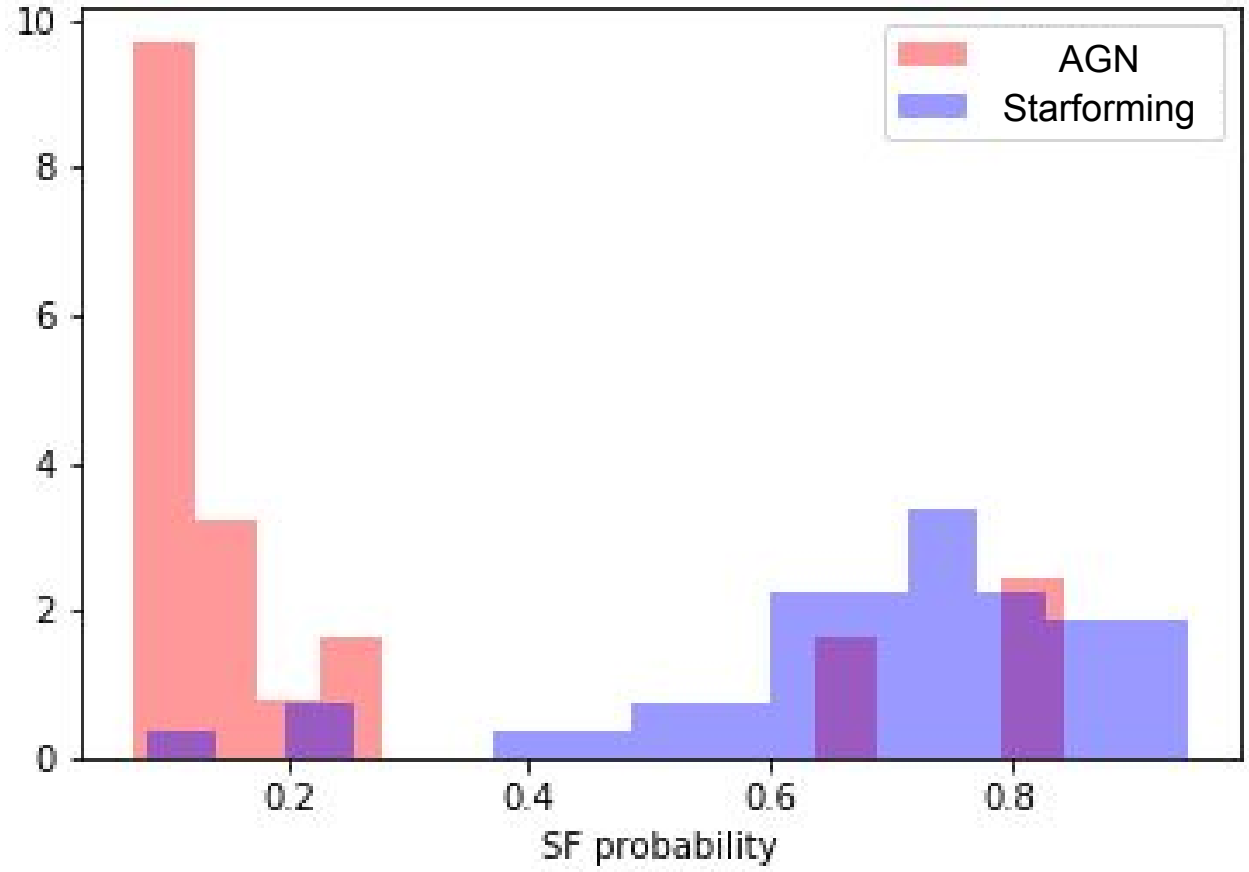
	Avg. Precision	Avg. Recall
AGN	0.79	0.79
Starforming	0.89	0.89

- Average P-R curve AUC = 0.85

Mod

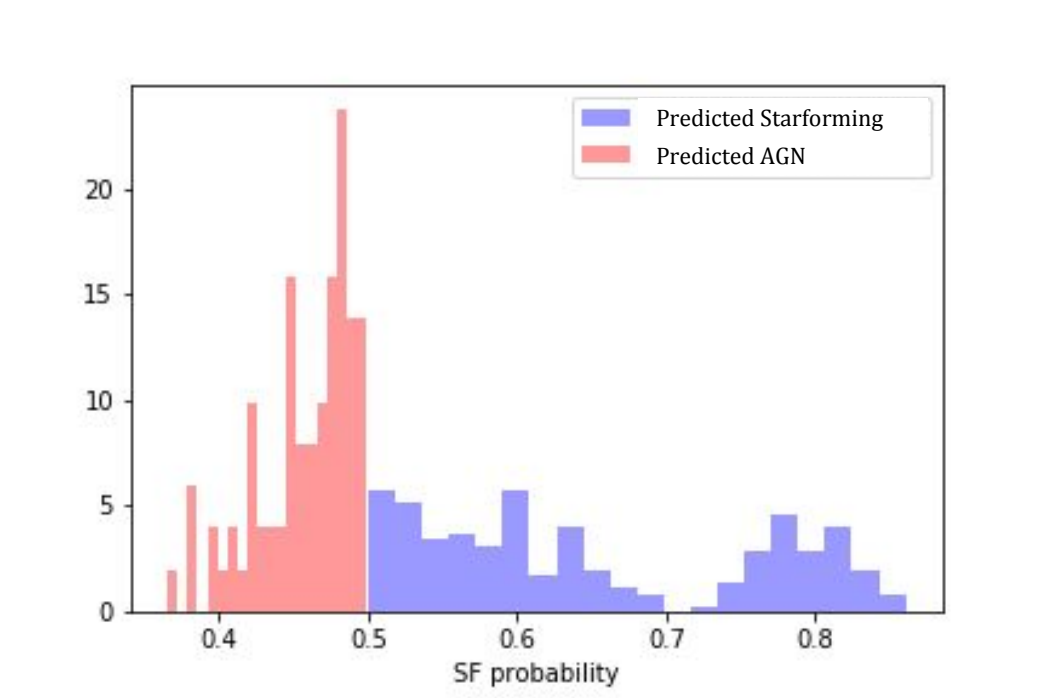
-

-



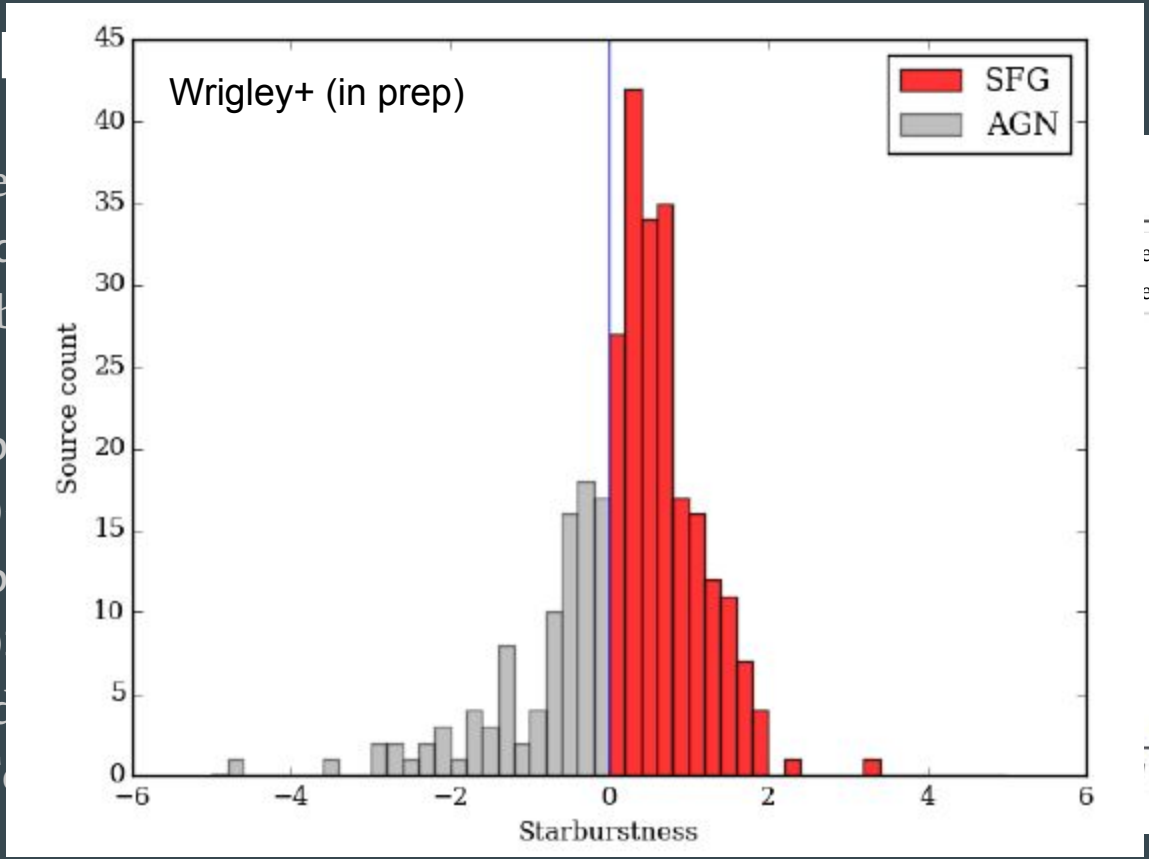
Apply to Unclass Data (270 Galaxies)

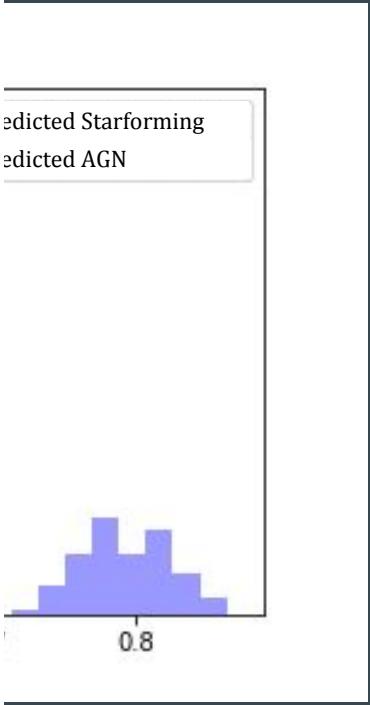
- The 10 models were applied to the unclassed data
- Average Probability Class 1 (SFG)
- Number of predicted galaxies with P(SFG)<0.5 = 76
  Number of predicted galaxies with P(SFG)>=0.5 = 194
- ~30% AGN dominated
- ~70 % Star Forming dominated



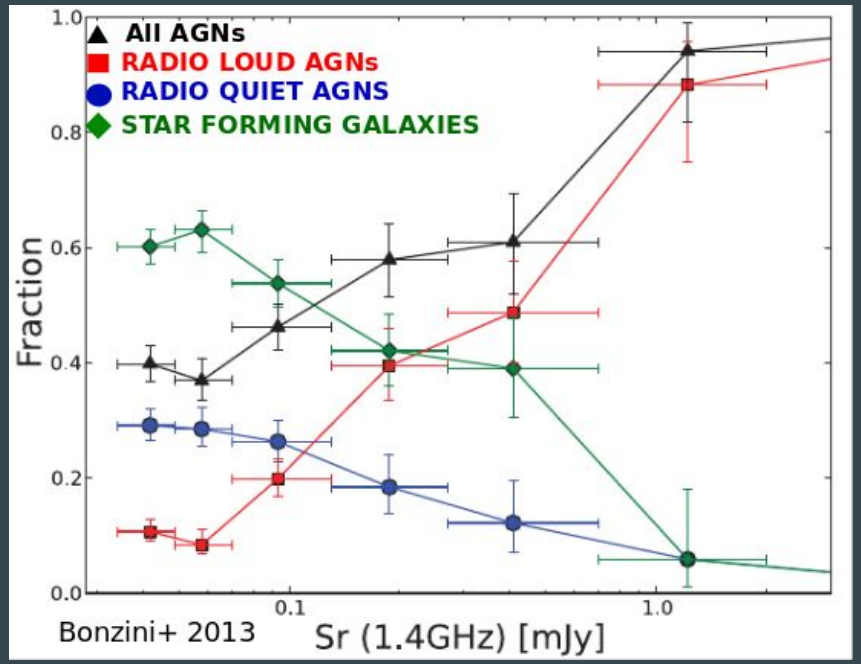
Apply to U

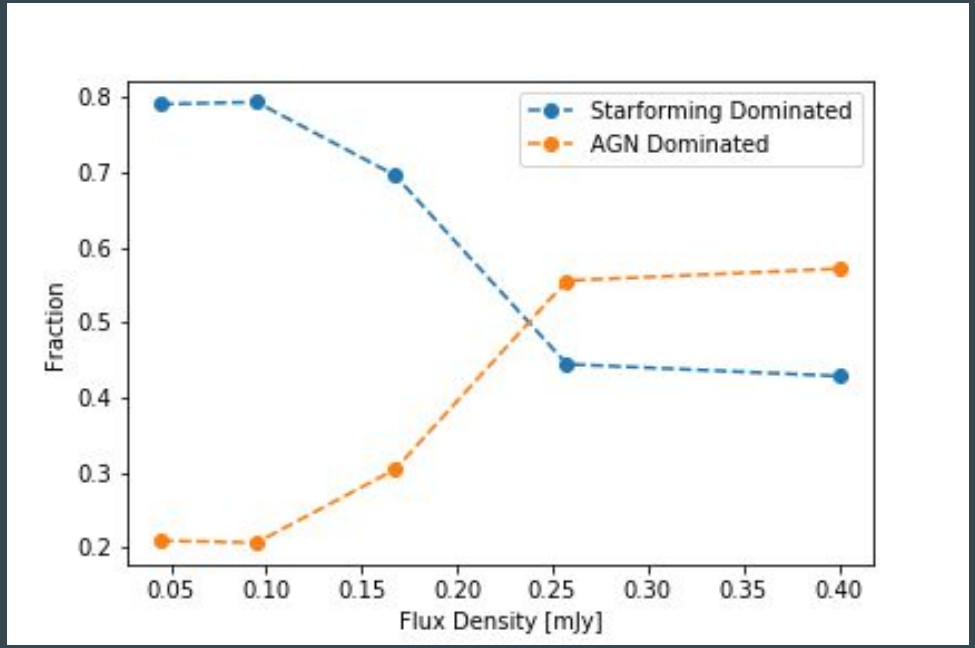
- The 10 mode
  the unclassed
- Average Prob
  (SFG)
- Number of p
  with P(SFG)
  Number of p
  with P(SFG)
- ~30% AGN c
- ~70 % Star F





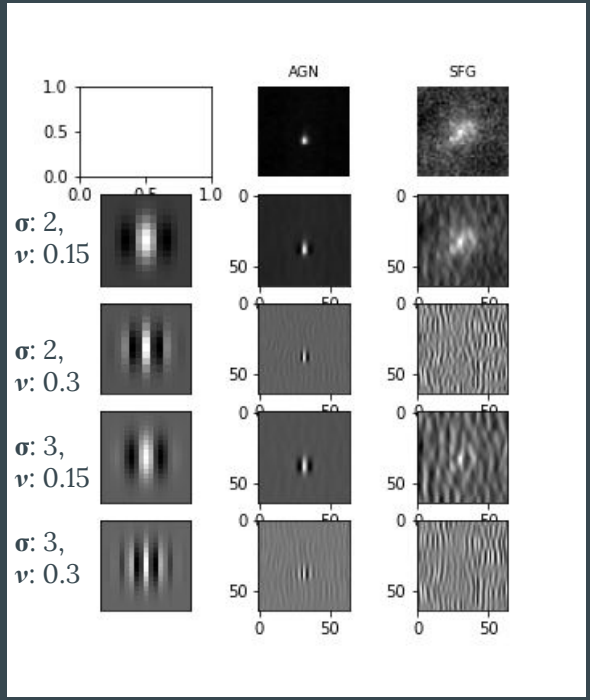
Comparisons with previous studies





Further Improvements

- To create a more general model, require more labelled data
  a. Observe more Galaxies (Time consuming)
  b. Simulate -> SKADS (Mphys/MSc student??)
- More sophisticated feature reduction:
  a. Gabor Filters - similar to human visual system
     - Applied to Gravitational Lens by Hartley et al (2017)
     - Can then be used to combine with other images



Summary

- Nuclei of (some) galaxies can be Star forming or AGN dominated
- Large statistical studies are beneficial to understand their connection
- Previous methods of classification are time consuming and expensive
- Next-generation of radio astronomy facilities increase the number of galaxies detected
- Can no longer use the “old ways” for classification
- Apply machine learning to the problem, results agrees with previous studies.
- Reduce the amount of time, manpower and cost of these studies