Self-organised Maps and the automatic classification of radio sources in the SKA era

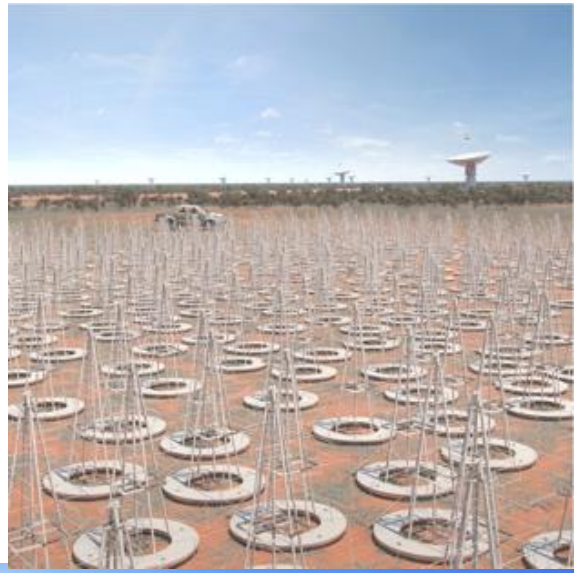
Tim Galvin | CSIRO Astronomy and Space Science
Minh Huynh (CASS), Ray Norris (CASS), Rosalind Wang (CSIRO Data61), Kai Polsterer (HiTS), Erica Hopkins (HiTS)

The Challenge

Collection of next-gen radio telescopes ☺

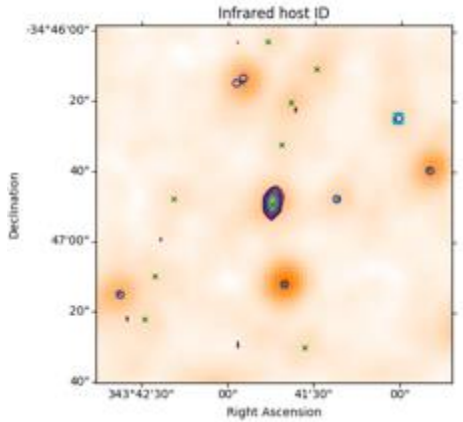
Soon to be overwhelmed by the incoming datageddon ☹

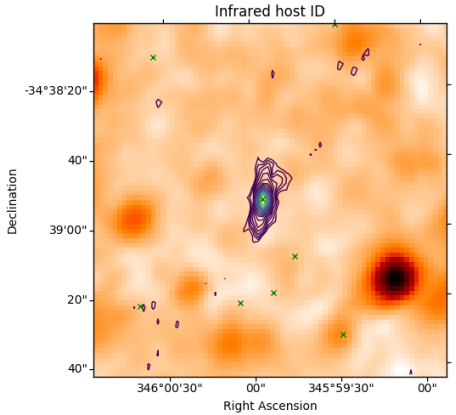
Limited set of humans to look at the important things ☹ ☹

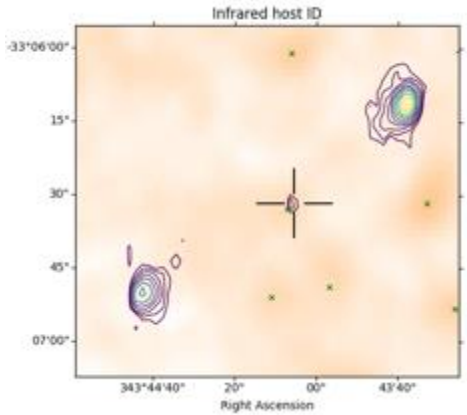


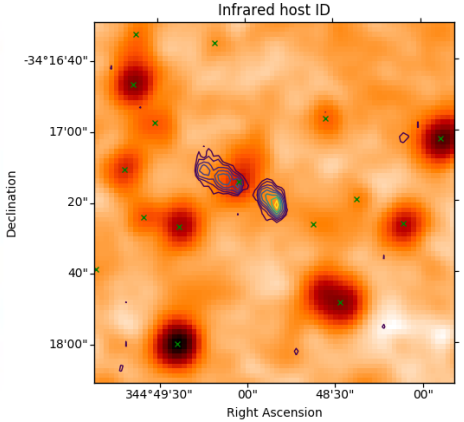
The Challenge

GLASS ATCA + WISE





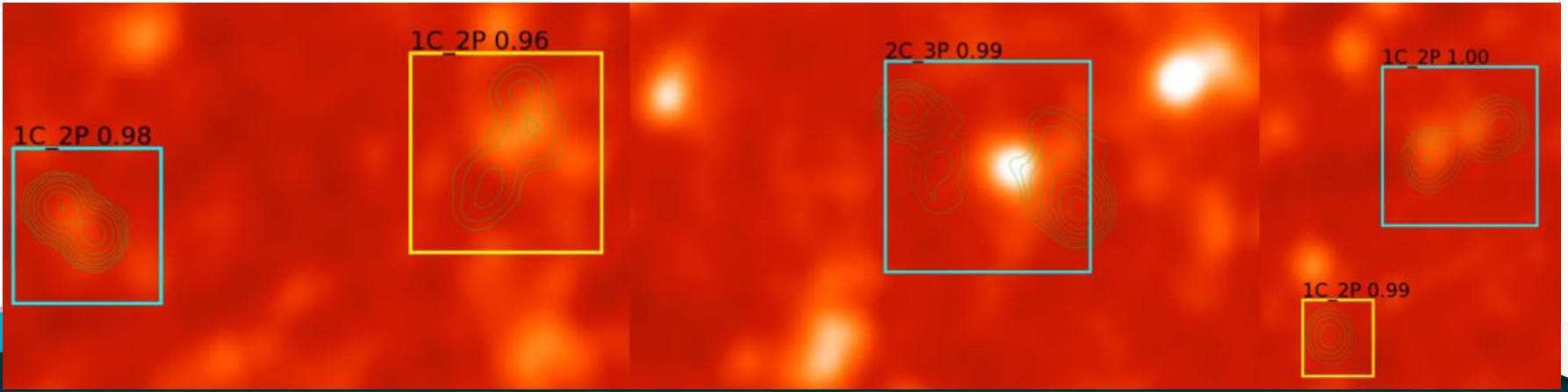




- Radio galaxies can be complex
- Often multi-wavelength information required (IR+Radio for example)
- So far, best classification and cross-matching is by human eye
- But too many sources in future (10s of millions!)

Machine Learning to the rescue (?)

- Effort has been invested for many aspects of this problem
- Convolution Neural Networks
  - Commonly applied to image data
  - Simple vs complex sources (Lukic et al. 2018)
  - Host galaxy identification (Alger et al. 2018)
  - CLARAN – source classifying (Wu et al. 2019)



Food for thought

- These methods require reliable labels to train
- Labels are expensive to obtain
- Need to consider
  - How transferable are labels?
    - Surveys will probe a different, younger Universe
    - Different frequencies = different physics
    - Different telescopes = different image characteristics
    - Different arrays configs = …
  - How transferable are trained models?
    - See above

Food for thought

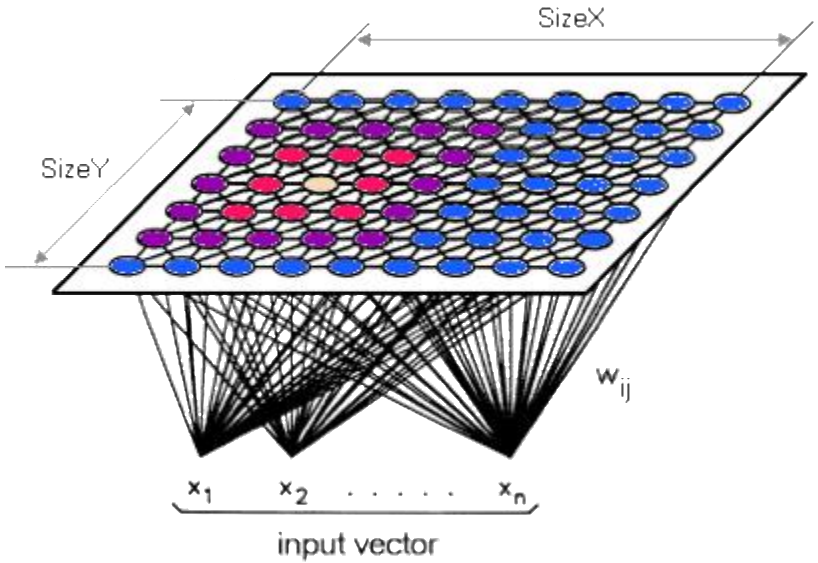
- Starting position for future surveys
  - Images
  - Positions of catalogue sources (components)
- Anything else is **extra** and not **guaranteed**
- How far can *unsupervised* methods take us?
  - No labels required
  - Training can begin immediately

Self-Organised Maps

An unsupervised method of clustering data

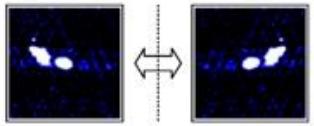
Simple idea:

- Start with an empty grid of weights (neurons) equal to shape of training data
- Select a random subject from training data
- Find best matching neuron from the grid
- Reward neuron by making it more similar to random subject
- Reward surrounding neurons, although not as much
- Repeat to completion, reducing area of influence throughout



PINK

- Parallelized rotation and flipping invariant Kohonen maps
  - Polsterer et al. 2015
- Radio galaxies:
  - Orientated differently
  - Generic SOM would see these as different
  - Bad SOM, a triangle is a triangle
- PINK brute forces problem
  - Produce all ‘realizations’ of single image
  - Compare all ‘realizations’ to map



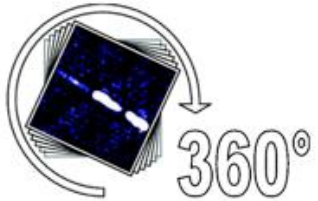


Figure 1. Both image transformations as they are applied to measure the similarity are shown exemplarily. The flipping (left) is shown on FIRSTJ075843.0+611936 and the rotation (right) is shown on FIRSTJ072529.5+614732.

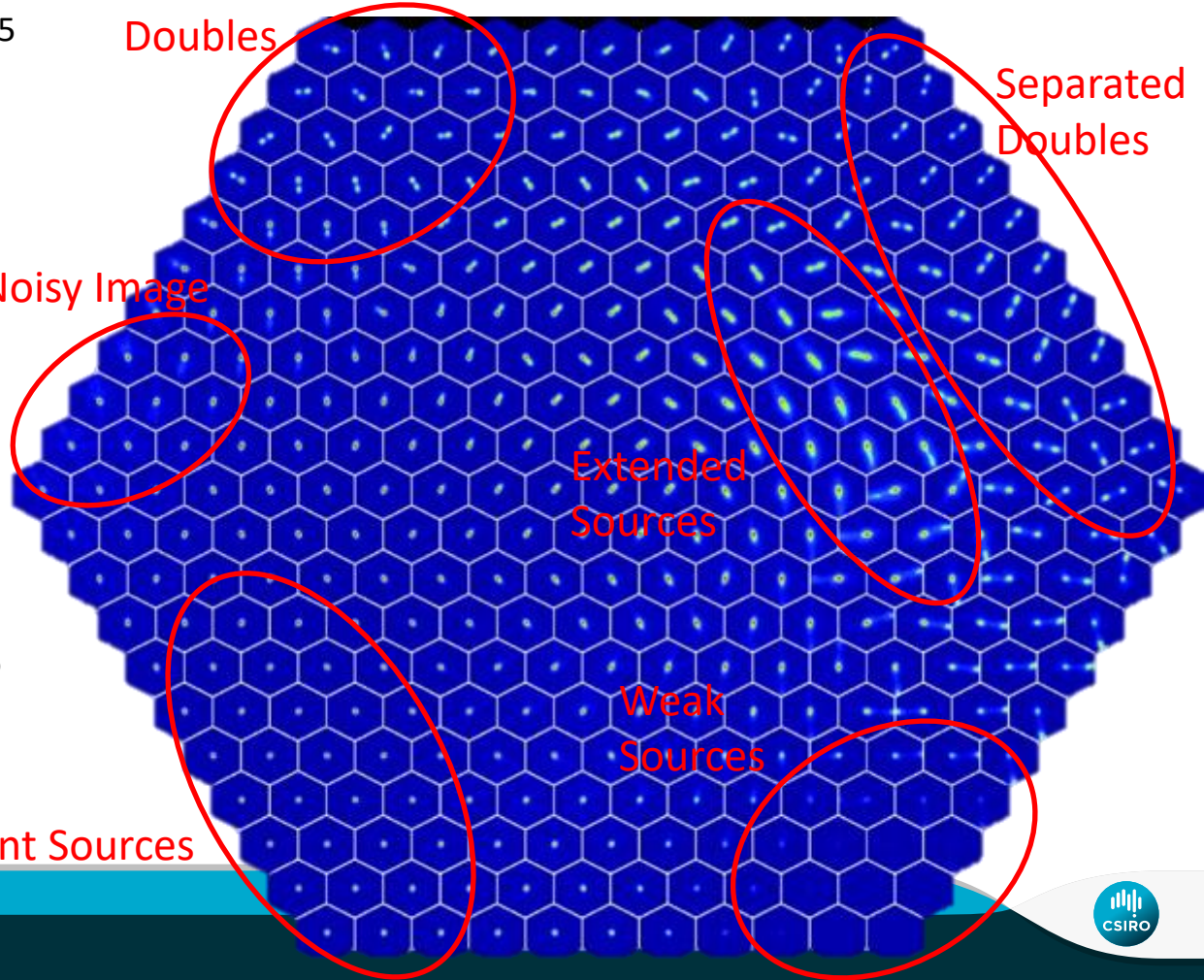
Polsterer et al. 2015

PINK

Polsterer et al. 2015

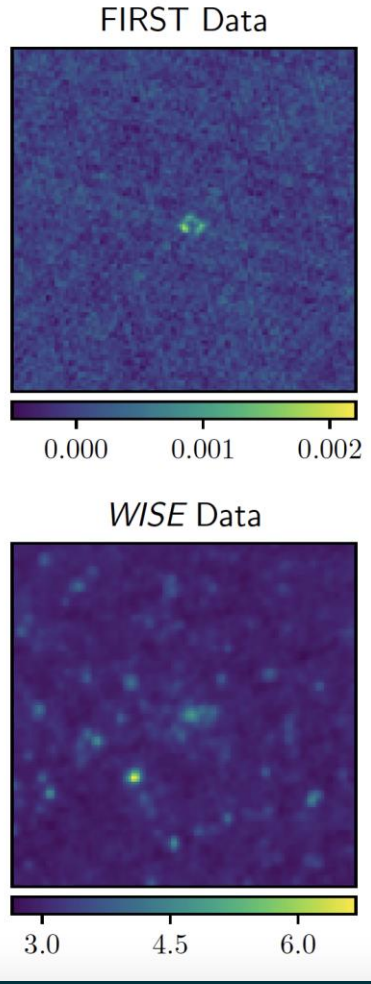
- Trained against 200,000 images from Radio Galaxy Zoo objects using FIRST data (Becker et al. 1994)
- Clustering of objects pretty obvious

- Can't infer much more than the SHAPE of radio structure
  - Two SFG or single AGN?



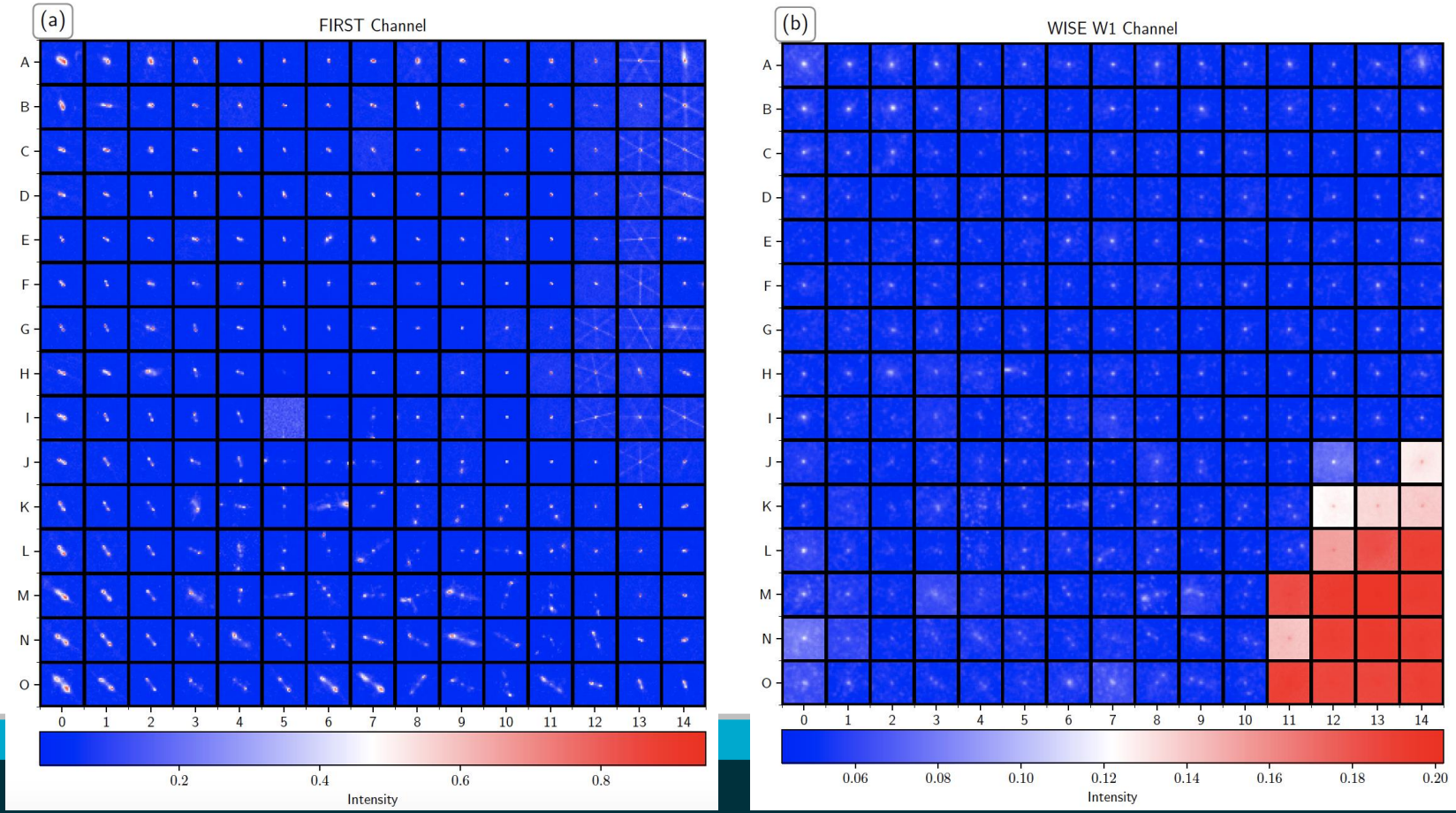
Our PINK Self-Organised Map

- Trained on both radio and infrared image
  - Combination can reveal physical meaning
- Positions taken from FIRST RA/Dec
- Radio images from Very Large Array FIRST survey data, WISE infrared imaging
  - Postage stamps images downloaded at the centered FIRST positions positions
  - Images cubes were made



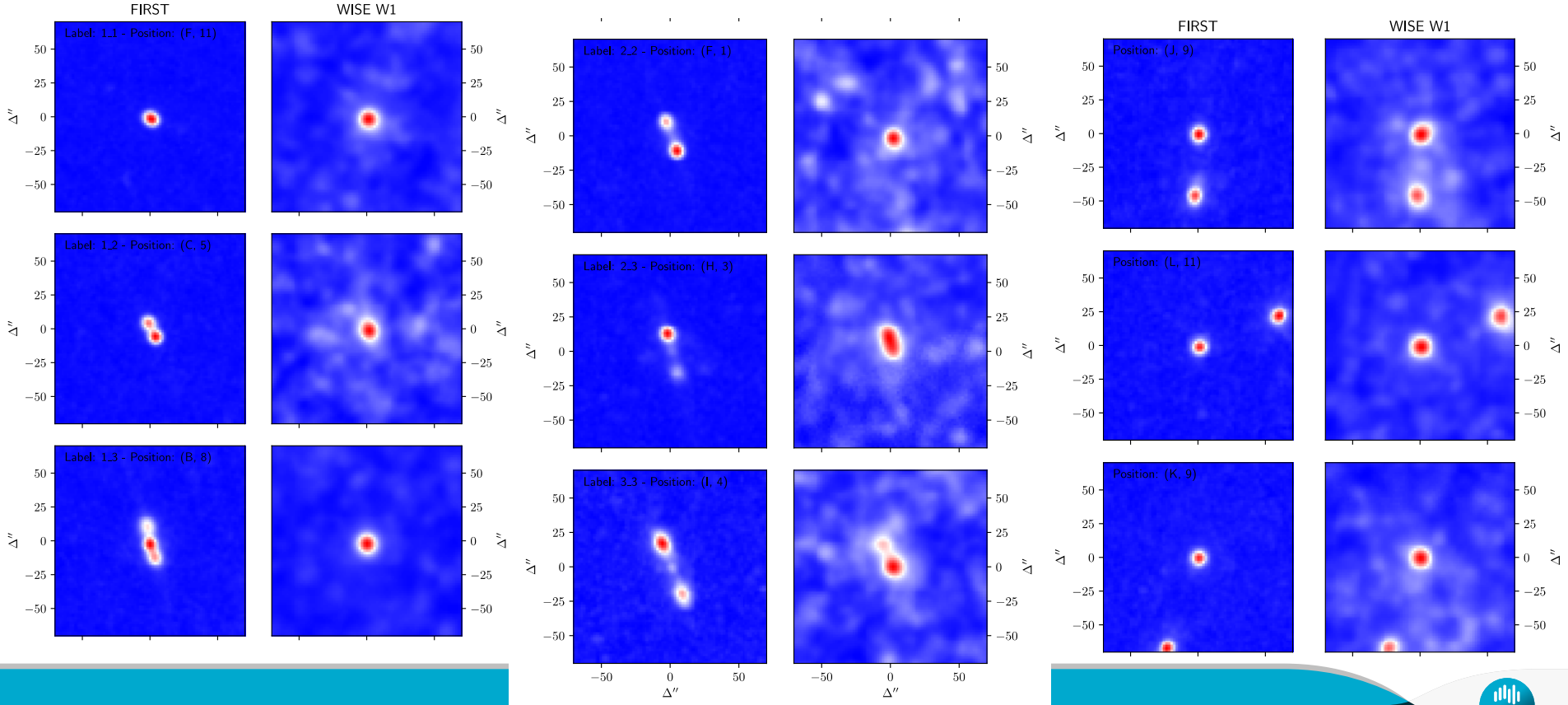
Our PINK Self-Organised Map

Galvin, Huynh et al. 2019



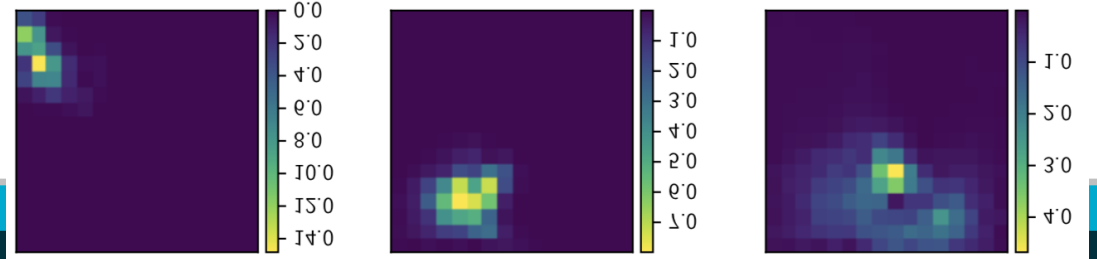
Example prototype objects

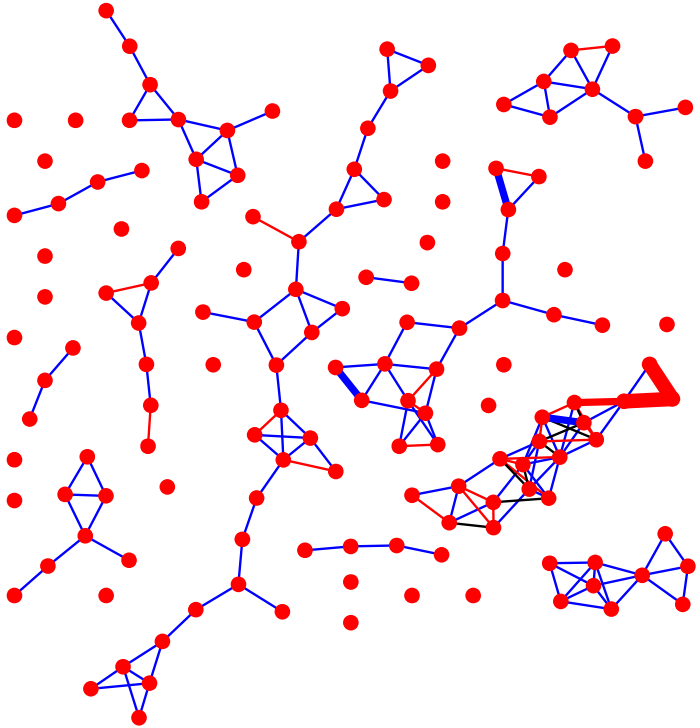
Galvin, Huynh et al. 2019



Clustering of SOM Neurons

- SOM makes data compete for representation on a limited space
  - Predominate features in the images dominate the SOM
  - Rare/complex object often ‘lost’ on the map
- Can’t just ‘go big’ training time grows
- Divide data into *related* segments and conquer
- CONNection VISualisation (CONNvis) method (Tasdemir & Merenyi 2009)
  - See how often neurons ‘activate’ to example sources in training set
  - Look for related neurons i.e. neurons that consistently activate at same time
  - Graph them and isolate unique sets

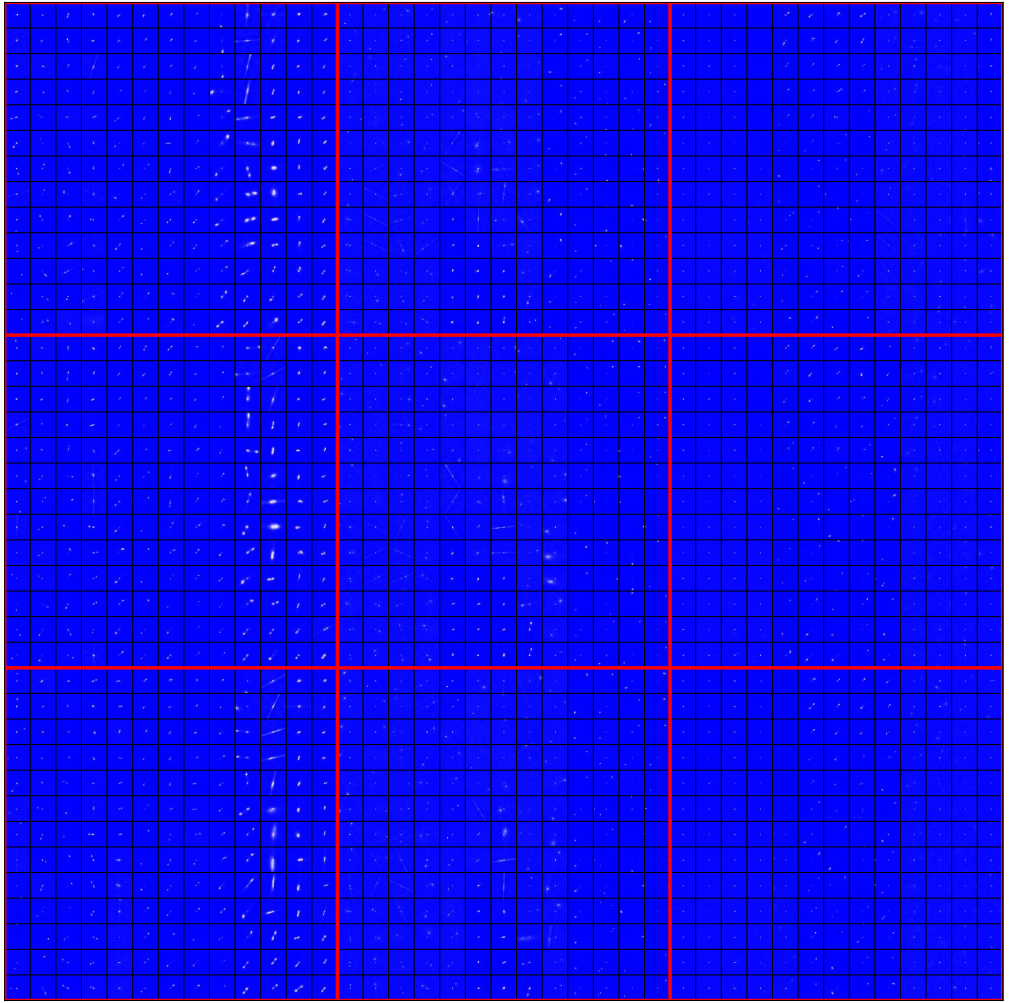




Galvin, Huynh et al. In prep

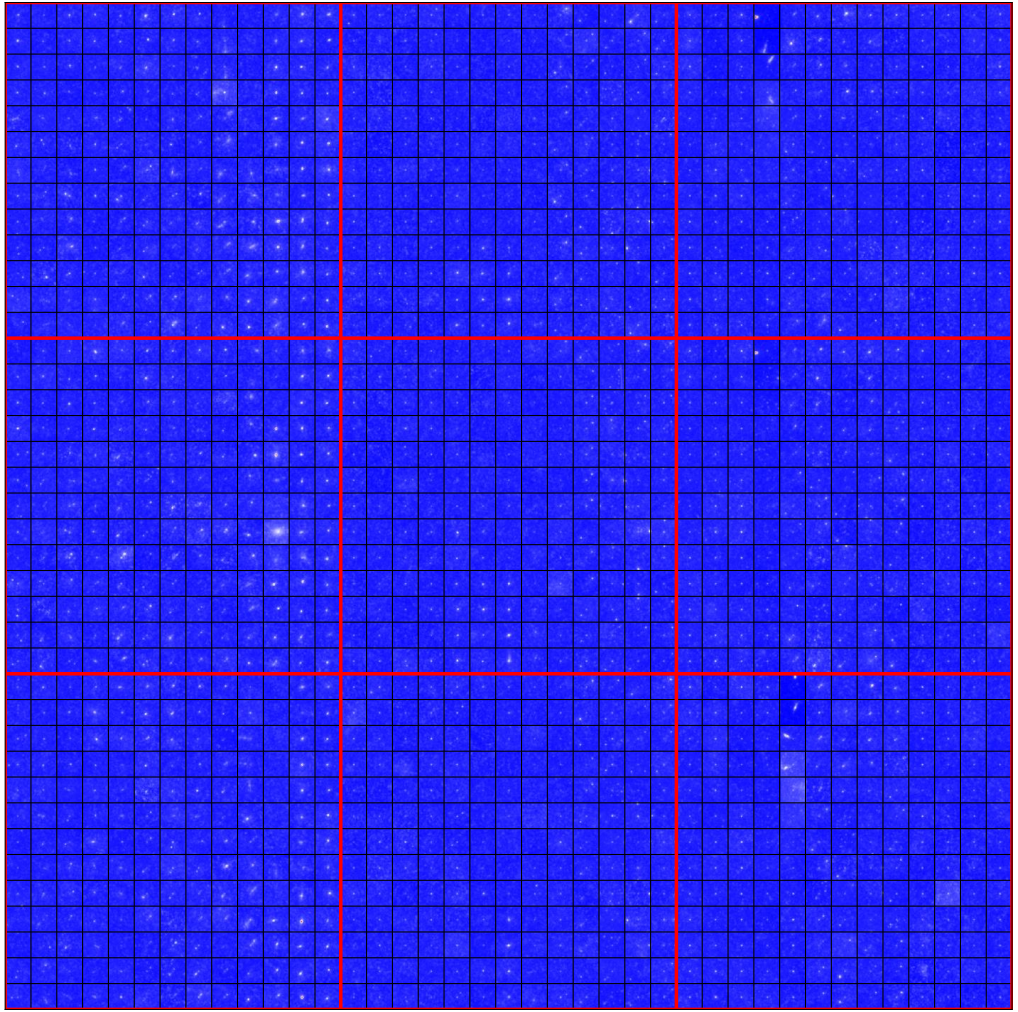
Mega SOM

- 9 individual source subsets
- 9 separately train SOMs
  - Each focusing on a particular feature within the data
  - Trained in parallel each on a subset of data speeds up considerably
  - 39x39 SOM in << day
  - To the right is the radio channel



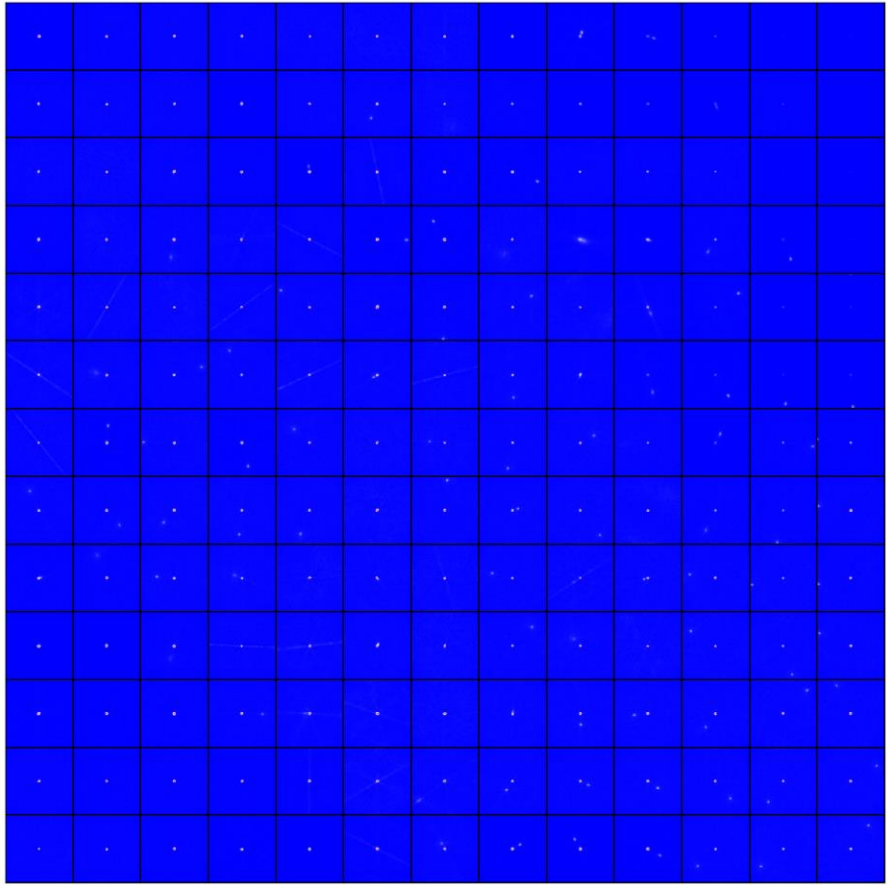
Mega SOM

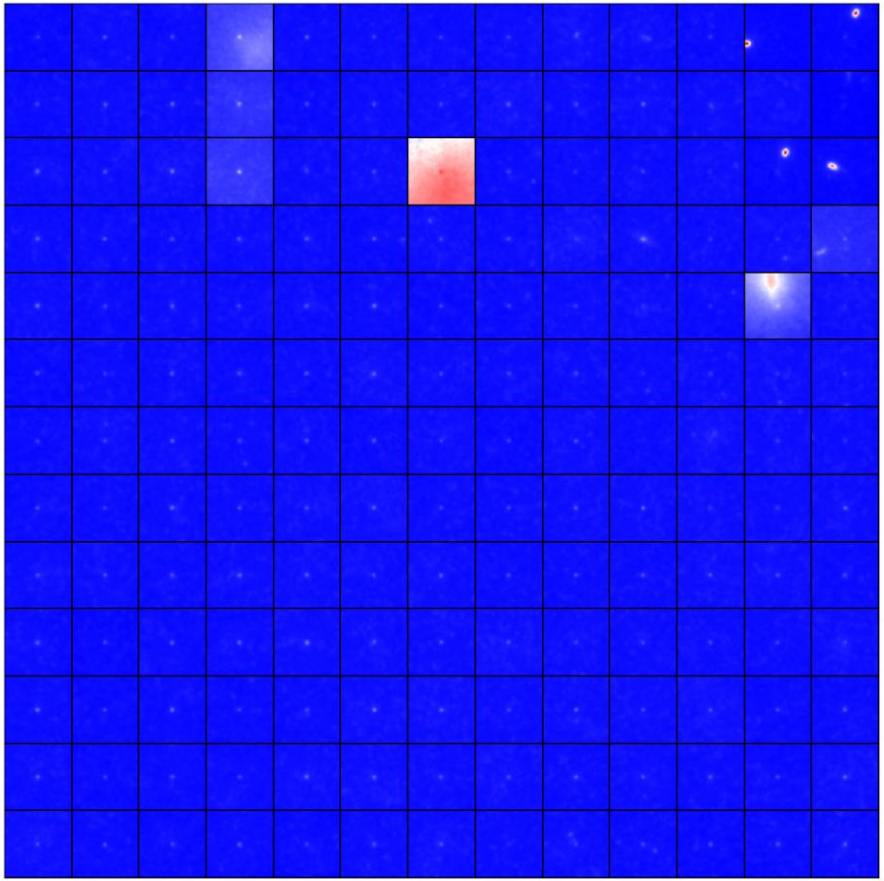
- 9 individual source subsets
- 9 separately train SOMs
  - Each focusing on a particular feature within the data
  - Trained in parallel speeds up considerably
  - 39x39 SOM in << day
  - To the right is the infrared channel



SOMs of Segmented Data

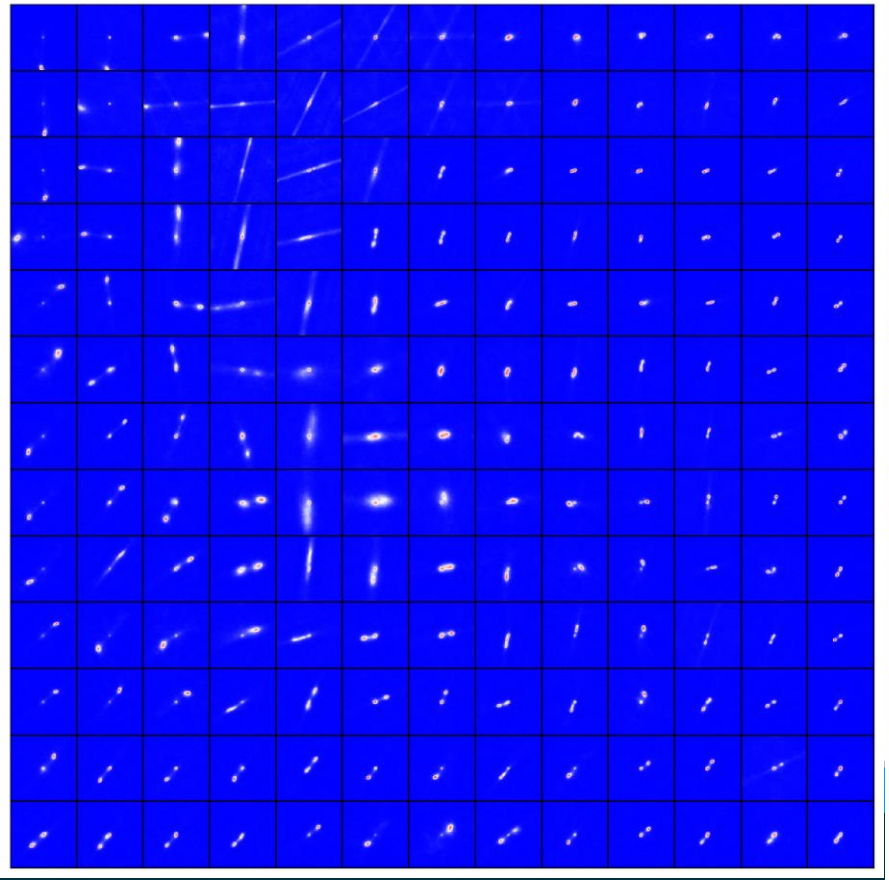
- Strong point sources - boring

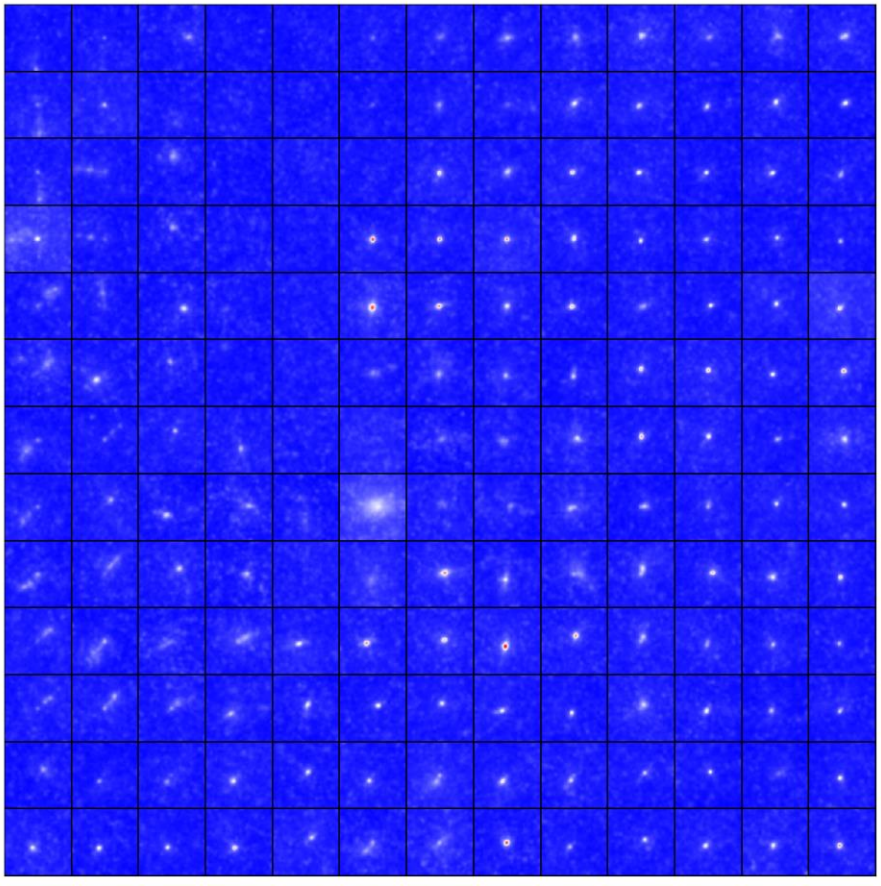




SOMs of Segmented Data

- Big, fluffy and wonderful things



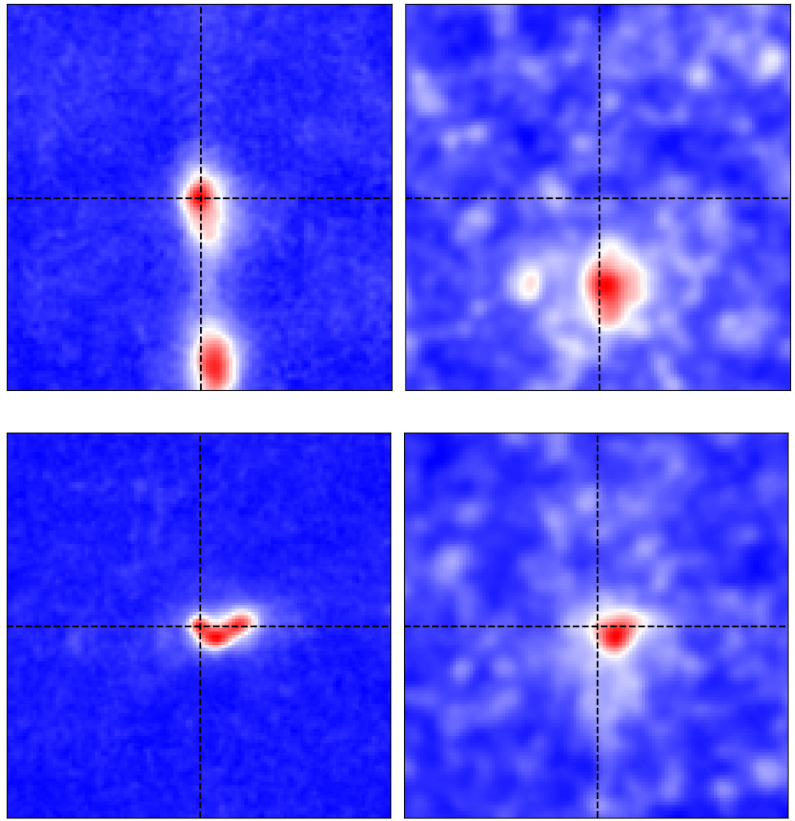


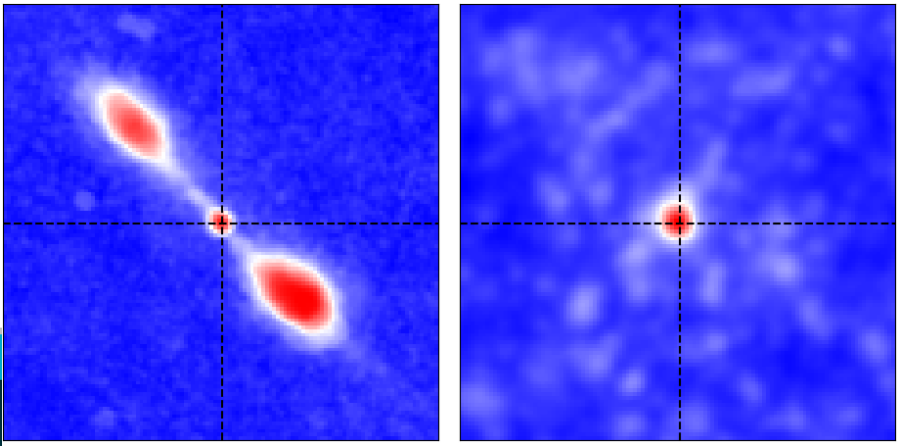
Complex sources

- Enlarging the SOM/segmenting allows for more complex features to be learnt
- Use new SOMs to identify populations of complex-shaped sources
  - Widely separated lobes (Giant Radio Galaxies?)
  - Wide-angle tailed galaxies

Example trained SOM neurons

FIRST Channel WISE Channel



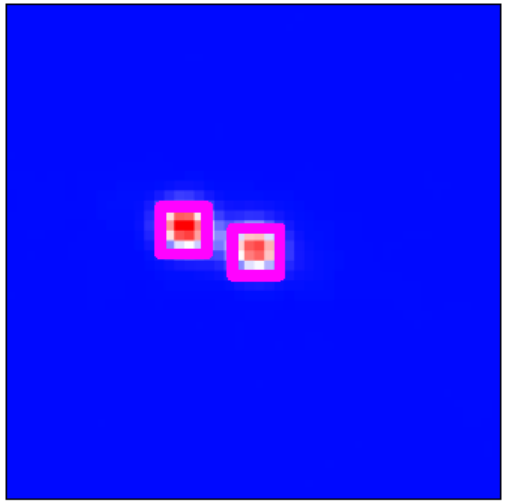


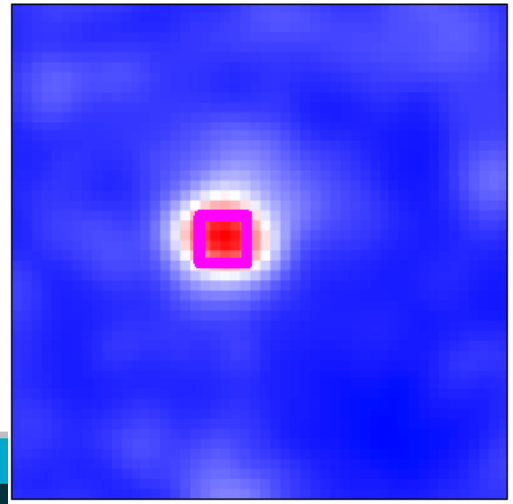
Using PINK – what's the point ?

- Simplify -> annotate -> transfer
- The SOM takes the unstructured, complex input image data
- Reduces the scope of the problem to a simpler, structured one
- Knowledge from simple SOM can be *transferred* back onto training data
  - A source will have a best matching neuron
  - Can create labels for a neuron (*what* and *where*)
  - 950k sources vs <<2k neurons
  - PINK brute forces rotation/flipping (i.e. transform)
    - Labels transferred from neuron to source (*what*)
    - Go from pixels on neuron to sky on source (*where*)

Transforming knowledge

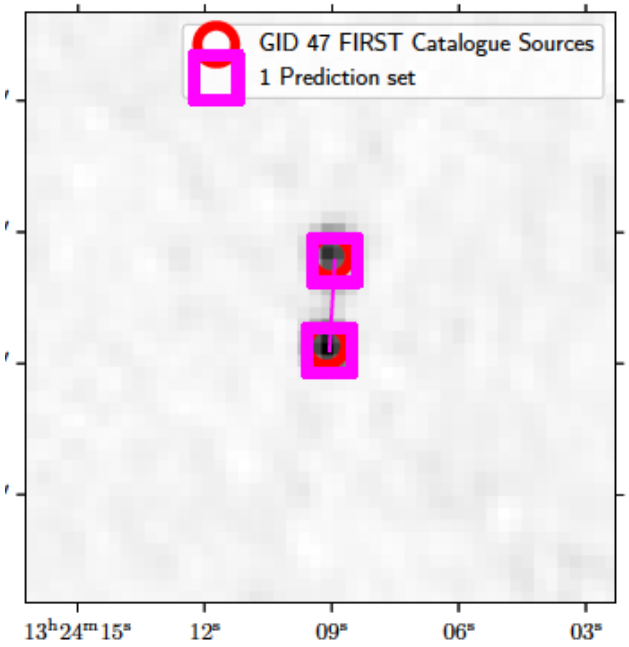
- Pixel level information from neurons can be *transformed* into sky positions
- We know
  - source sky reference frame
  - transform matrix
  - pixel level position of features (relative to center)
- Pink open boxes represent *related* features to the centered radio feature that have been annotated by me

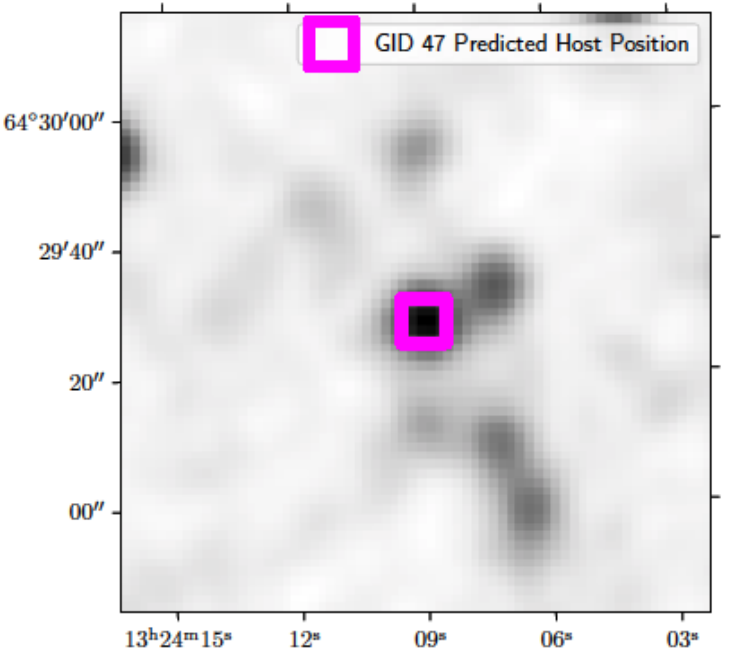




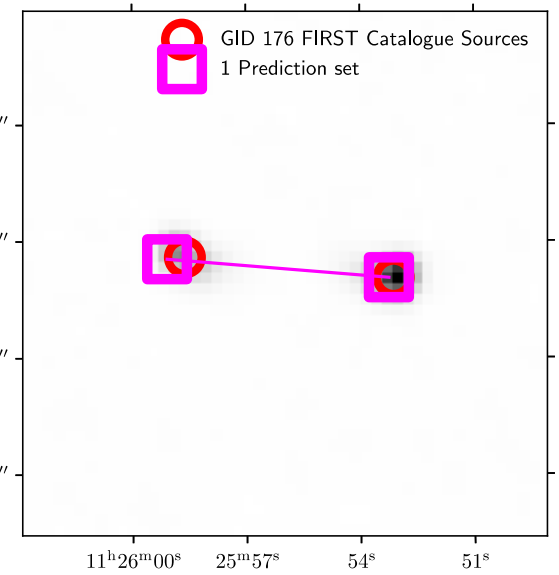
Applying to a source

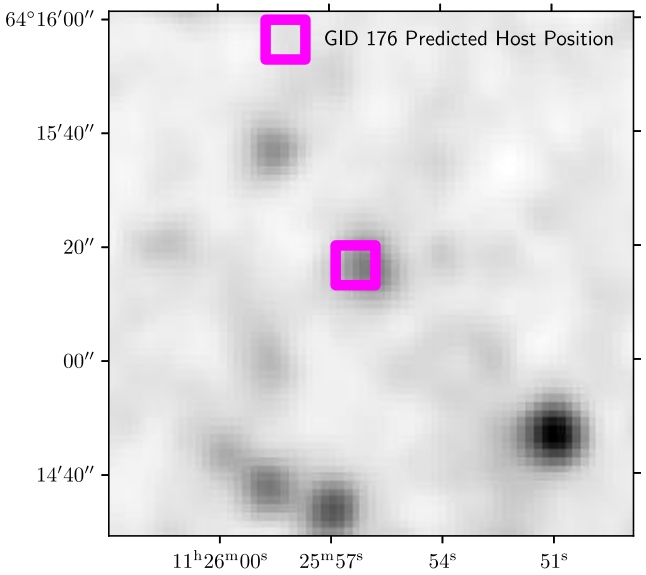
- FIRST source
- Pixel positions transformed to real sky positions
- Mechanism to identify related radio components and IR host
  - Radio Galaxy Zoo

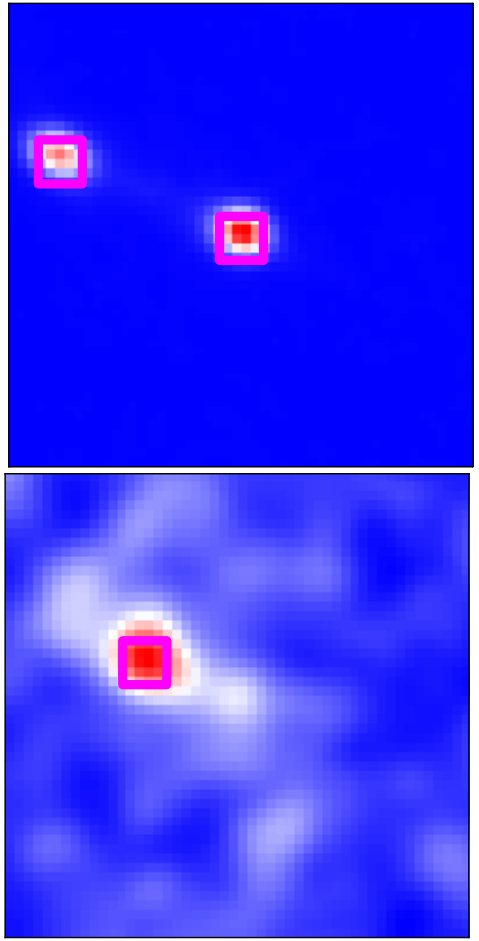




And another

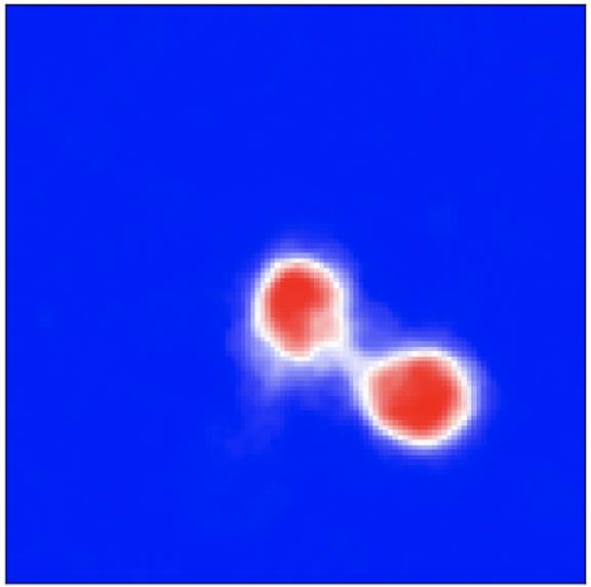

GID 176 FIRST Catalogue Sources
1 Prediction set
11h26m00s
25m57s
54s
51s

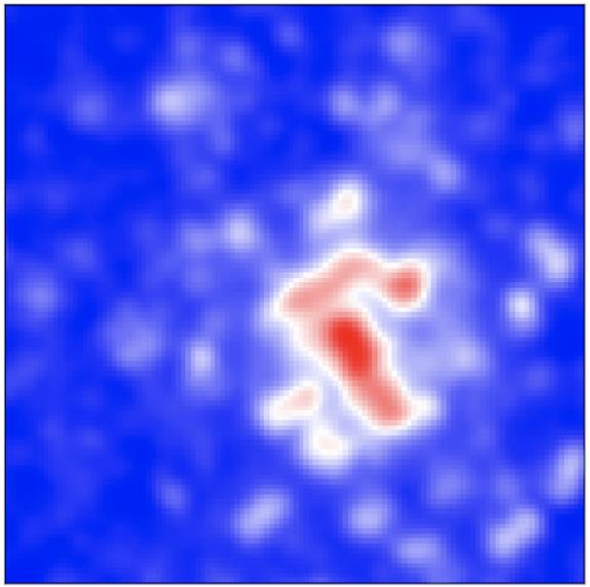

GID 176 Predicted Host Position
64°16′00″
15′40″
20″
00″
14′40″
11h26m00s
25m57s
54s
51s



Structured the data

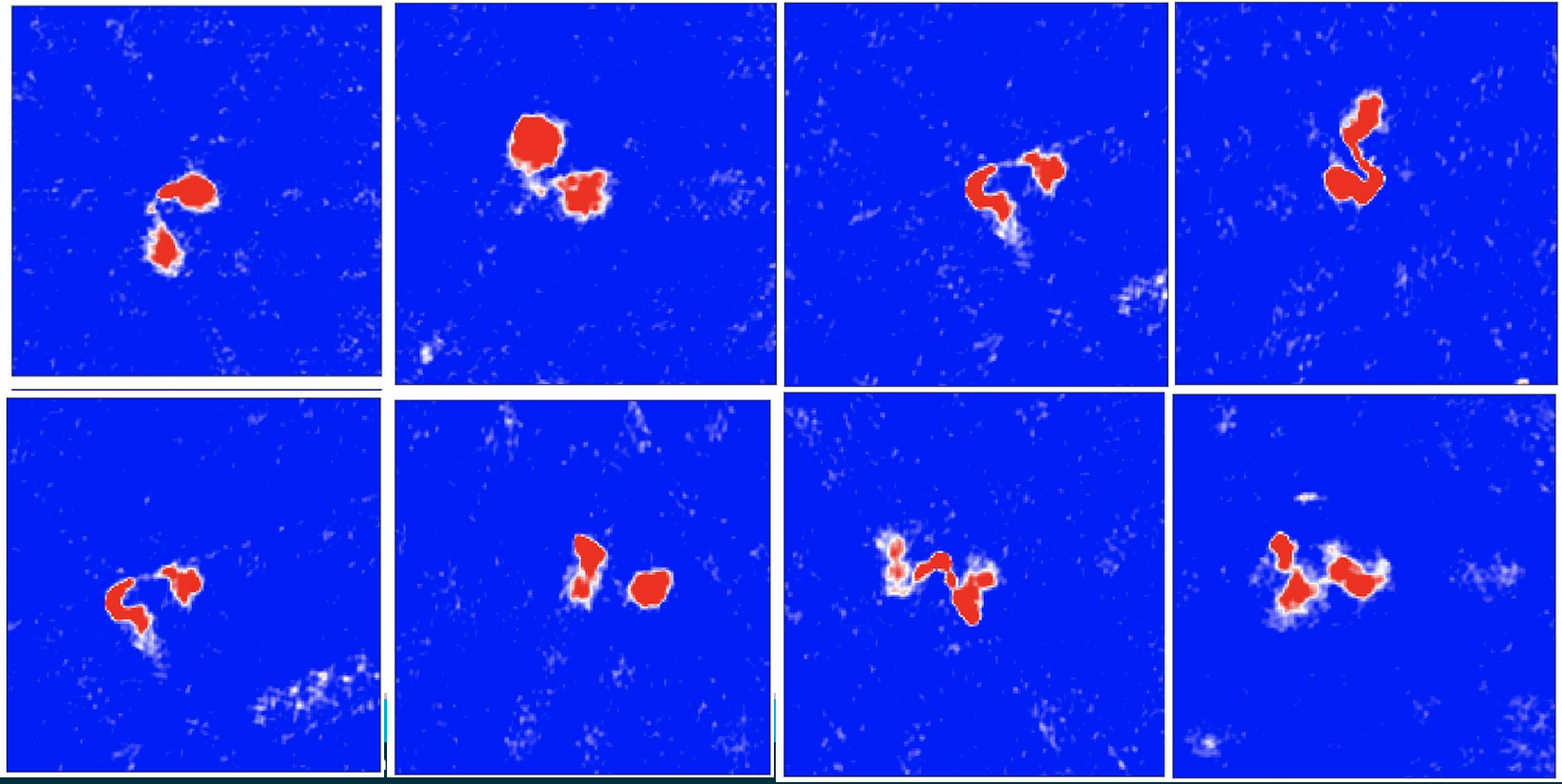
- Framework to efficiently explore the previously unstructured data
- Still need to *understand* it
- Lobes of this neuron are pretty circular, IR is very noisy…





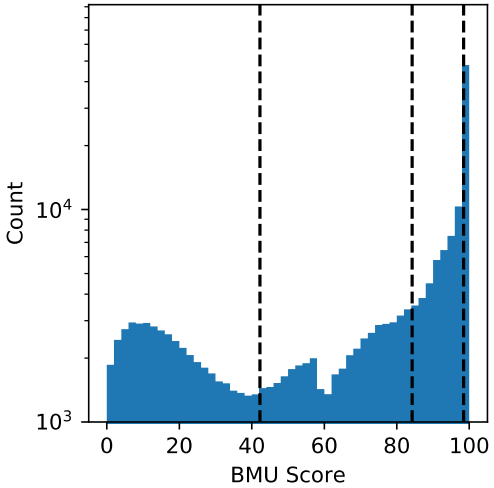
Structured the data

- To find the really interesting things (FRII galaxies for astronomers)

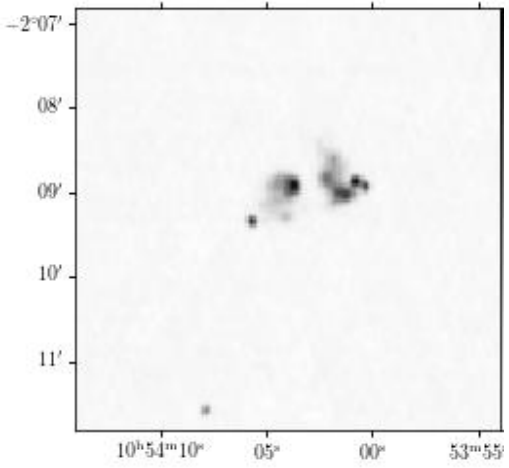


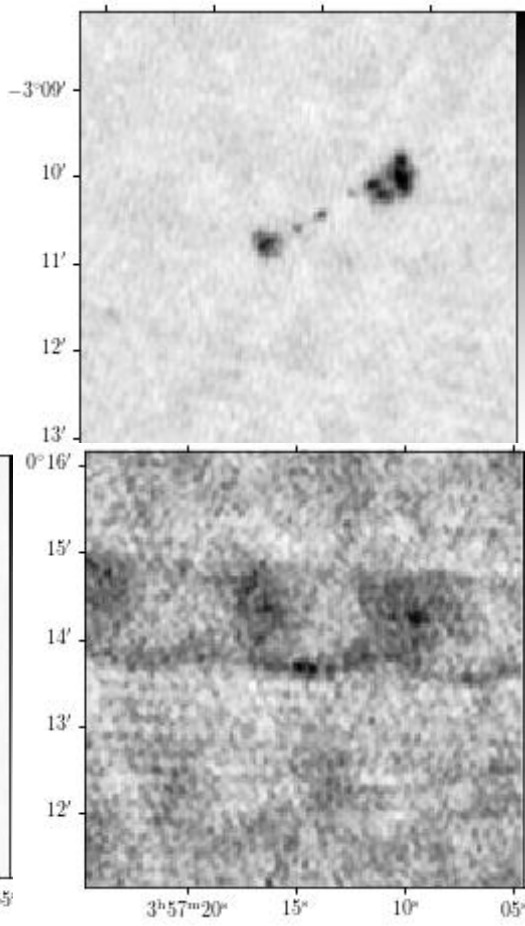
Outliers

- Each source will match up to a neuron
- Finding the poor matches will reveal interesting objects!



- Score near 100 is excellent match of a source to neuron
- Vertical dash lines represent quartiles in the distribution





Future Work

- Still very much a WIP, but approach appears to be sound
- Polish up python code+module
  - Most work is on github, but again WIP: https://github.com/tjgalvin/pink_e3
  - Tools to visualize performance of SOM/data
- More than just pixel position
  - Shapes encode information lost by a single pixel position
- Finish applying to FIRST+WISE
  - Produced a collated catalogue
    - Related radio components identified and link
    - Likely IR host position estimated
    - Measure of ‘goodness’/trust
- Apply to other SKA precursor surveys!
  - Anything/everything
  - Not necessarily just radio+ir -> stokes I + spectral index maps could be used

Summary

- Machine learning methods will be necessary in SKA era to catalogue and classify radio galaxies
- Self organised maps (SOMs) are a simple unsupervised method of clustering
  - No *prior* knowledge is needed
  - Can be applied to *any* SKA survey
- SOMs can be used to transfer knowledge:
  - Simplify -> annotate -> transfer
  - Labels applied to neurons
    - What a neuron/source contains
  - Labelling where things are on a neuron
    - Consolidate multi-component sources into single radio galaxy
    - Locate likely IR/optical host galaxy
  - Identify the really interesting things that we should actually look at (as humans)