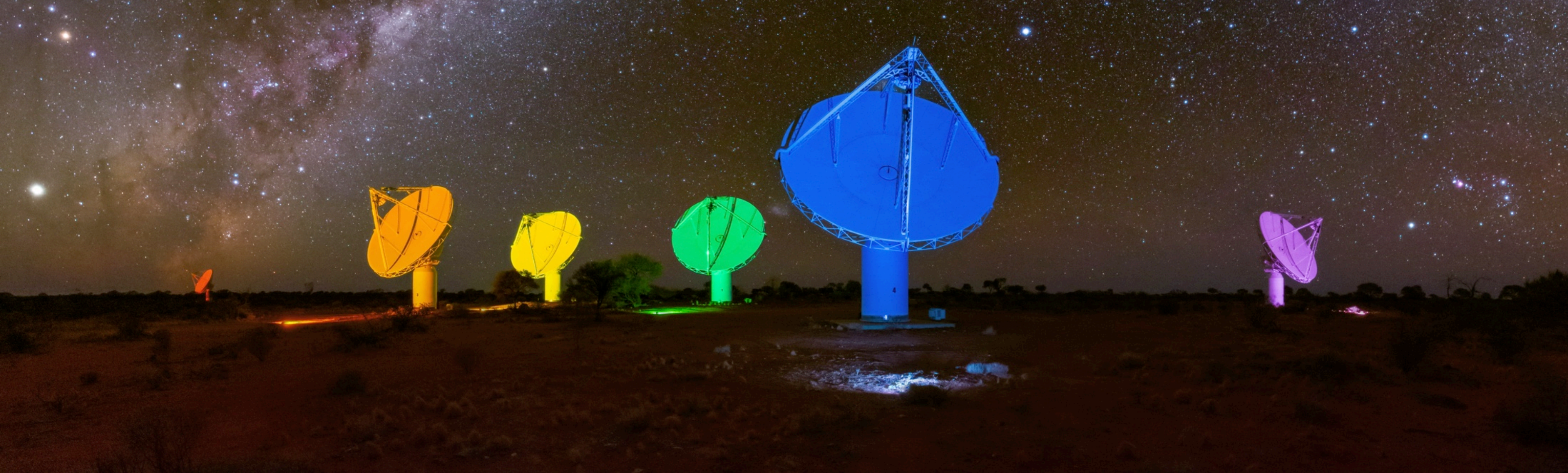
Self-organised Maps and the automatic classification of radio sources in the SKA era

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

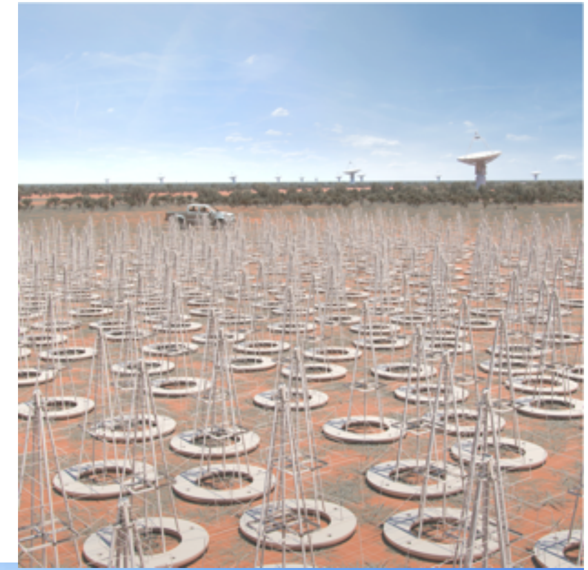


The Challenge

Collection of next-gen radio telescopes ☺

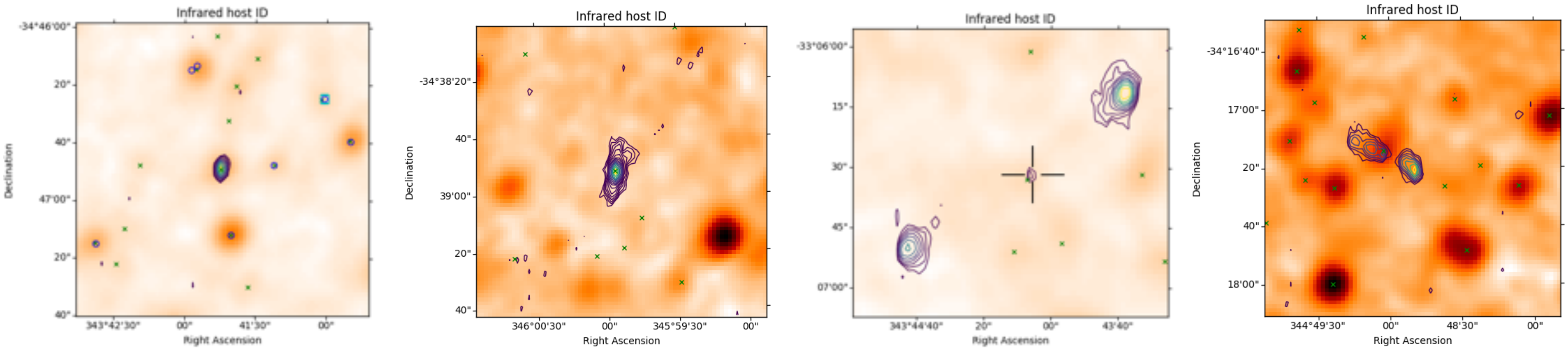
Soon to be overwhelmed by the incoming datageddon ☹

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



Aim of the game

GLASS ATCA + WISE



Produce a catalogue that describes galaxies and the set of resolved components they may have, including the host galaxy position in the case of resolved features

Solutions (?)

- Effort has been invested for many aspects of this problem
- Divide and conquer
  - Crowd sourcing through zoo-inverse platform, see “radio galaxy zoo”
- 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)
  - Labels – what to do?
    - New Universe, new sensitivities, new frequencies, new instruments …

What we *will* have

Set of images (from the instrument)

Set of source positions (from a source finder)

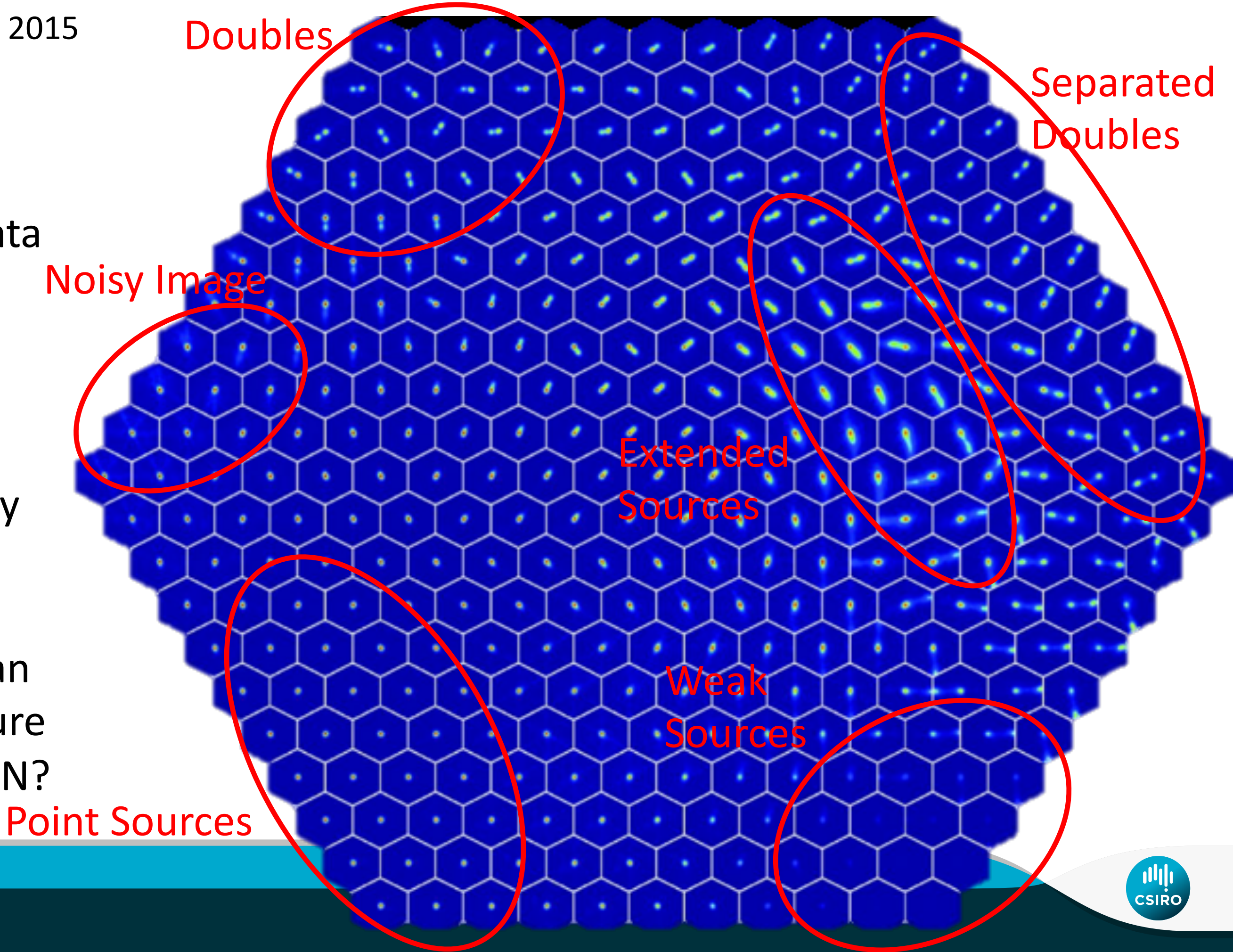
How far can we go with just these products?

PINK

Polsterer et al. 2015

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

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

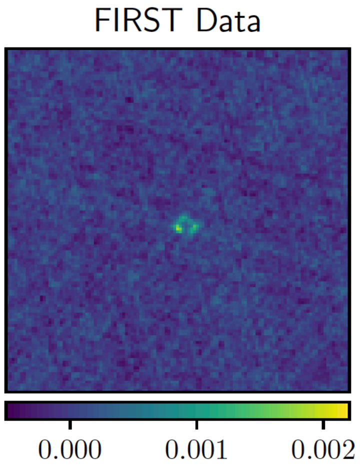


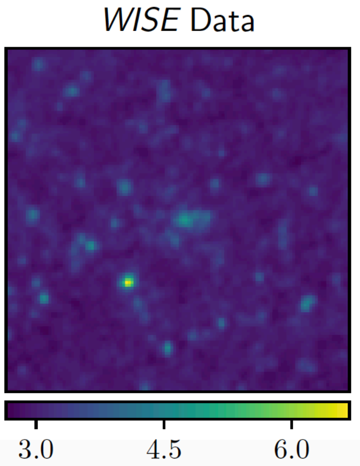
Lets start

- Started with the FIRST radio catalogue
  - ~950,000 source components (rows in the table)
  - No prior knowledge about how/which sources/rows are related – just a flat table

End goal – identify the related rows within the catalogue and add new information

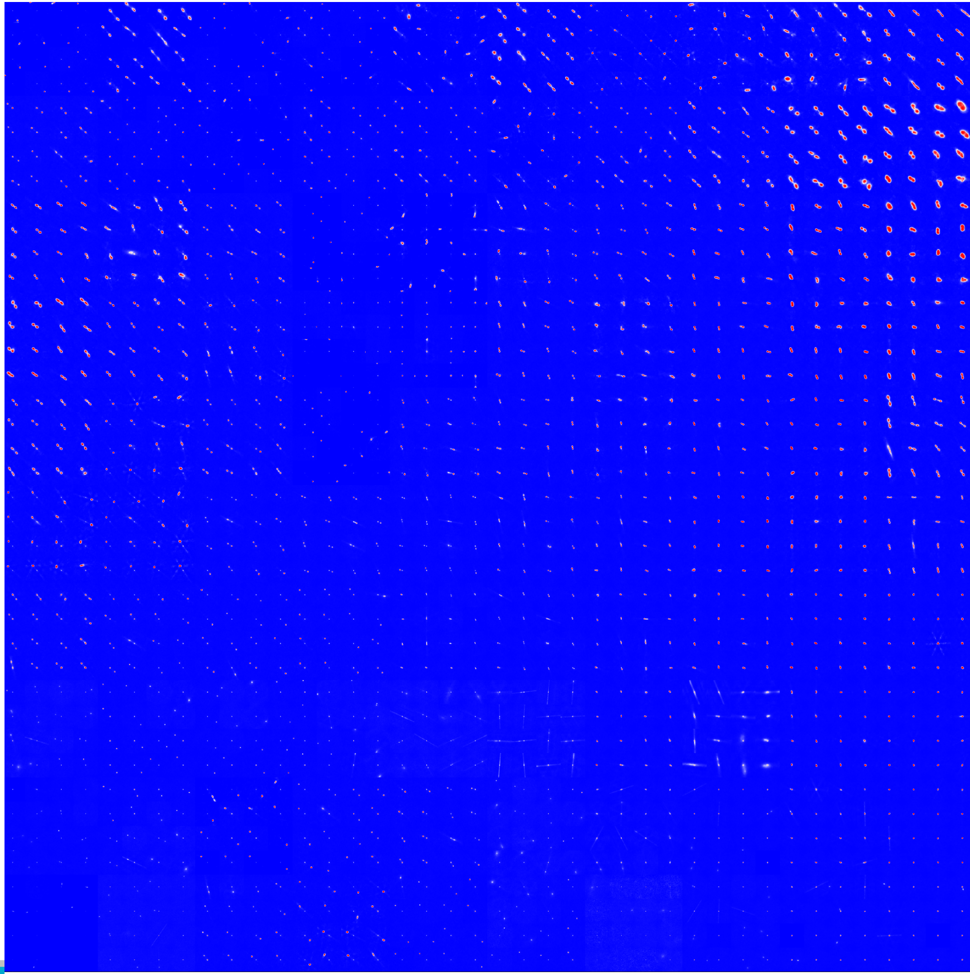
- Postage stamps images downloaded at the centered FIRST positions positions
  - IR gives information to infer object type
  - Images cubes were made





Train a SOM

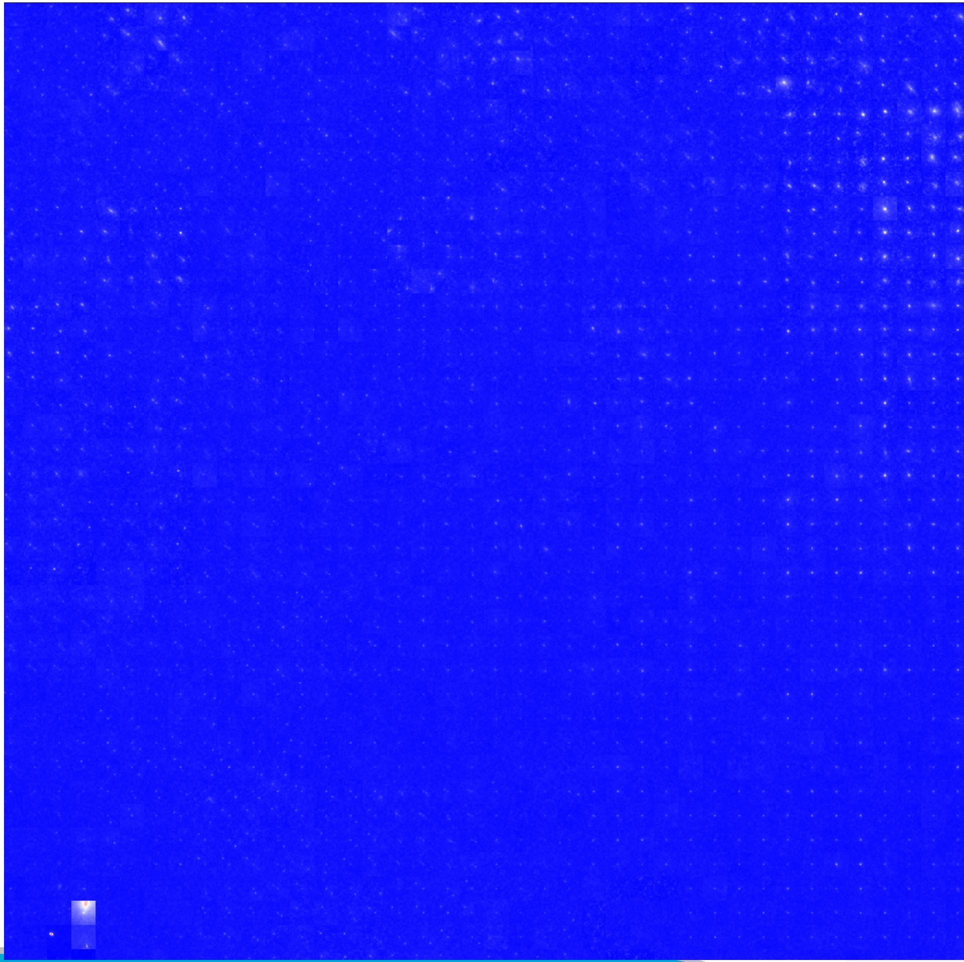
- A big one
- 40x40 neurons actually
- This is the radio channel



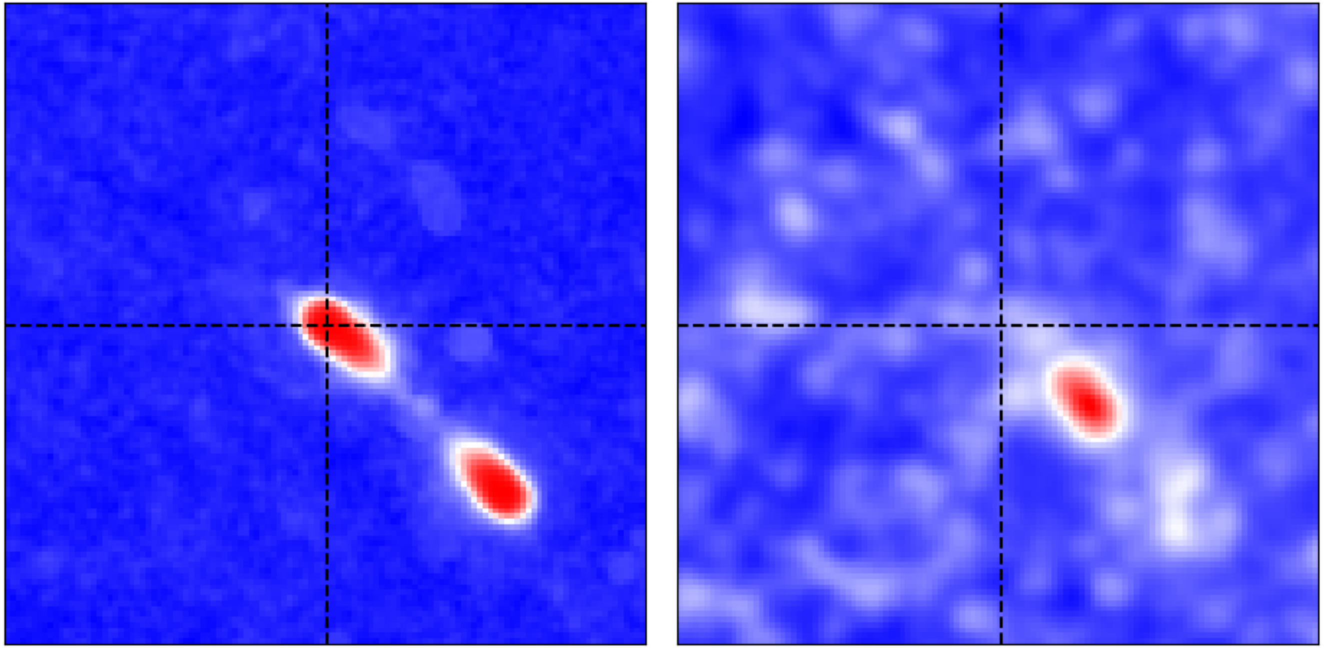
Train a SOM

- A big one
- 40x40 neurons actually
- This is the IR channel

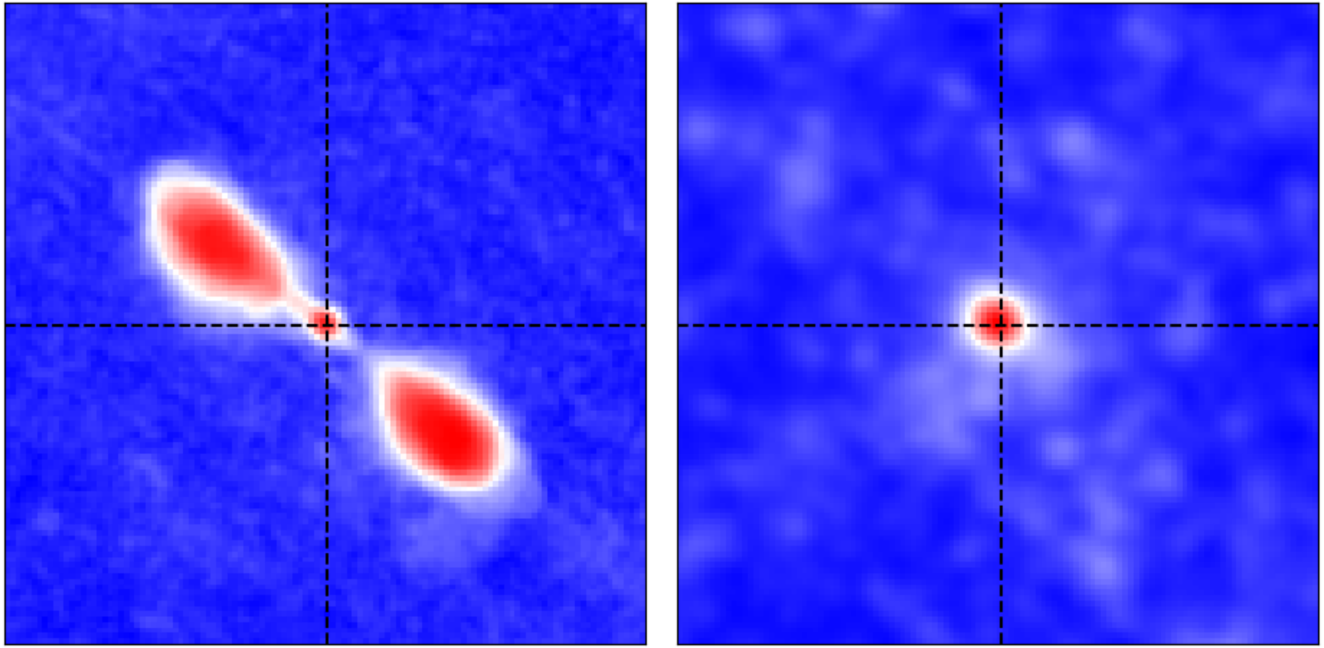
- Yes, very much aware not much detail can be seen

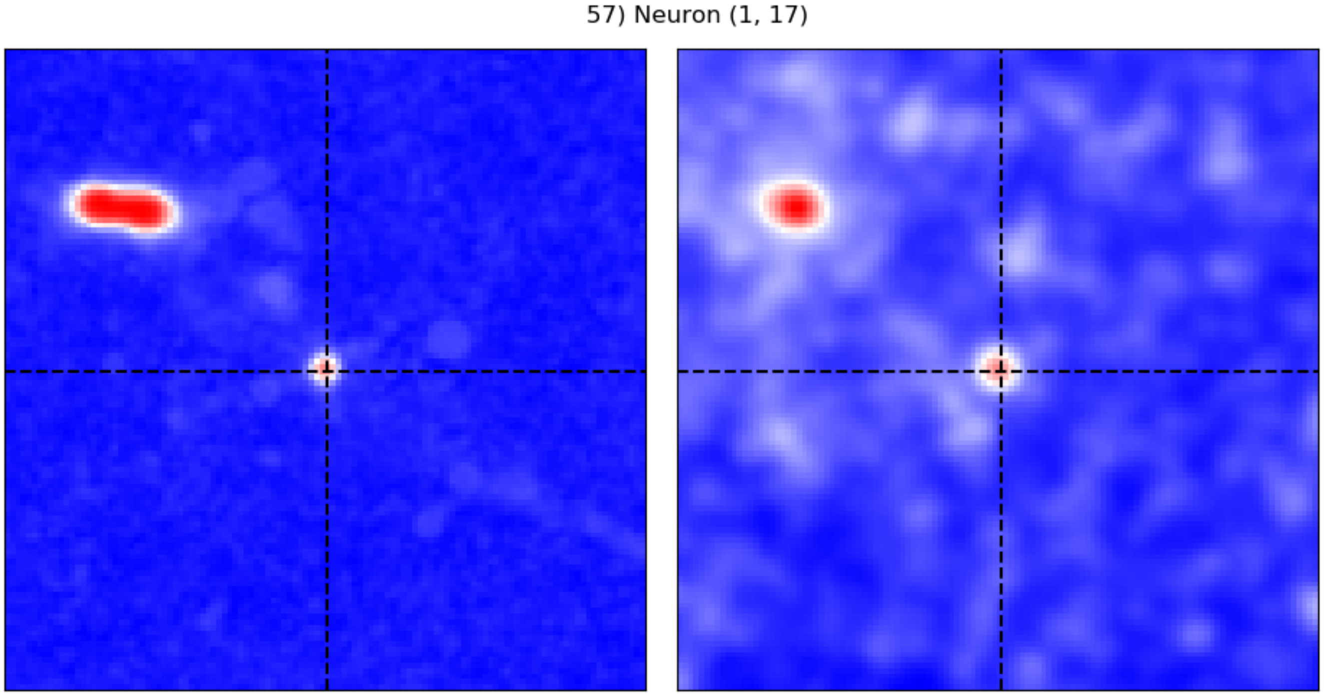


23) Neuron (0, 23)



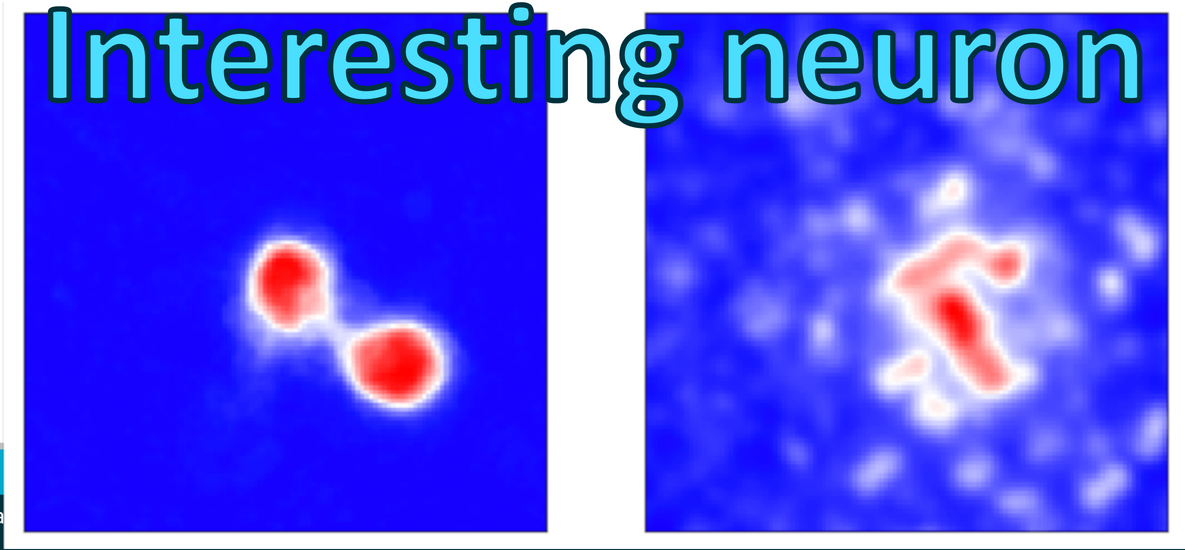
35) Neuron (0, 35)




57) Neuron (1, 17)

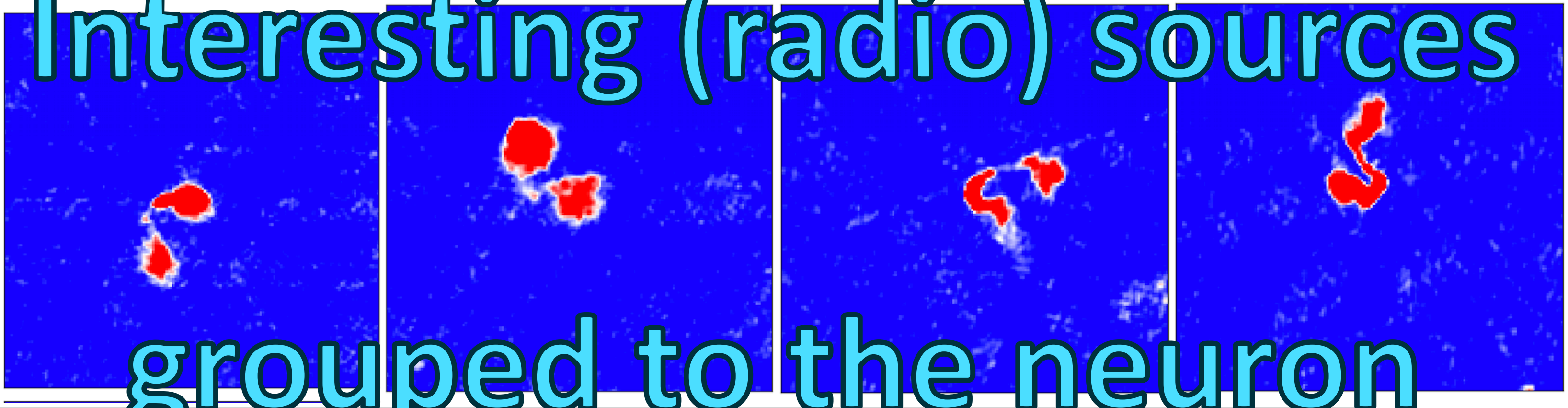
So what's the point?

- First off, we have given a framework to interactively explore the previous complex, unstructured image data
- Each individual image maps has a corresponding neuron
- Locate an interesting neuron, locating interesting sources



So what's the point?

- First off, we have given a framework to interactively explore the previous complex, unstructured image data
- Each individual image maps has a corresponding neuron
- Locate an interesting neuron, locating interesting sources

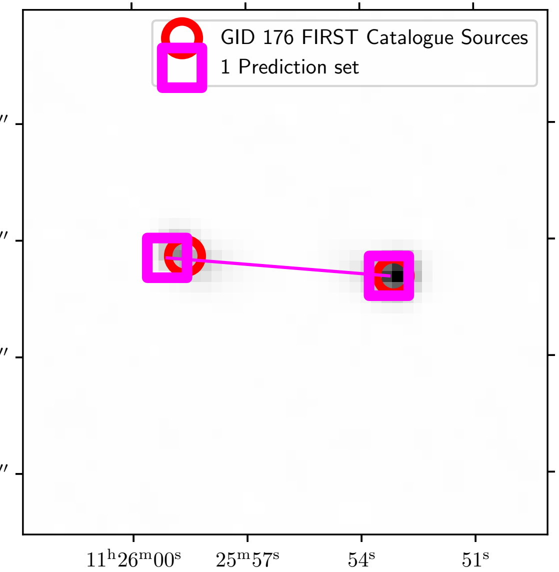


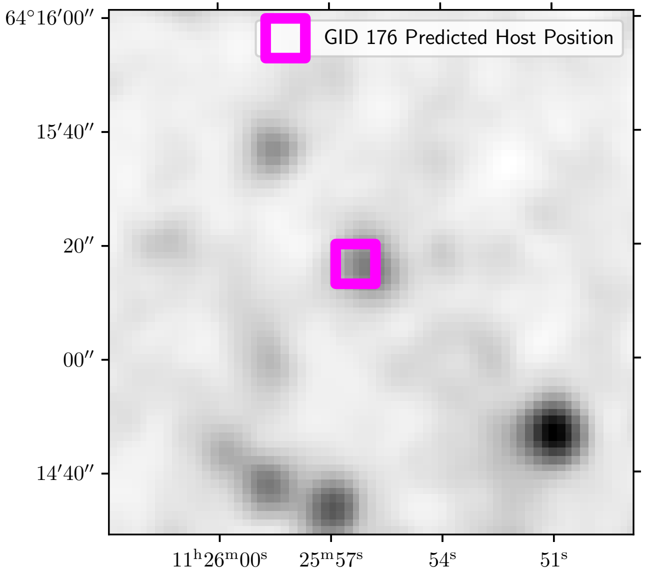
More importantly though

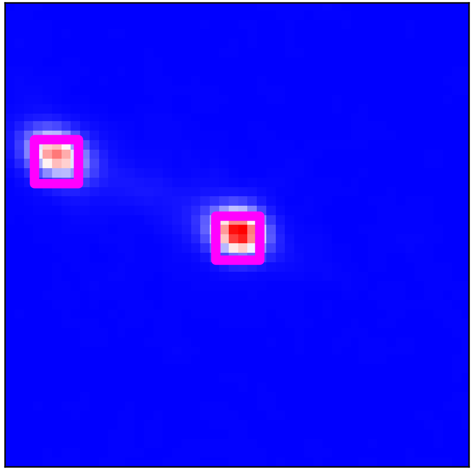
- PINK achieves rotational invariance through brute force
  - No ML, just good old fashion computation made possible with GPUs
- Lets label what a neuron contains **and** where it is located (pixels)
  - Transfer labels from neuron to objects that best match to them
  - Sky reference frame of source image + transform function + pixel locations = ***absolute sky positions***

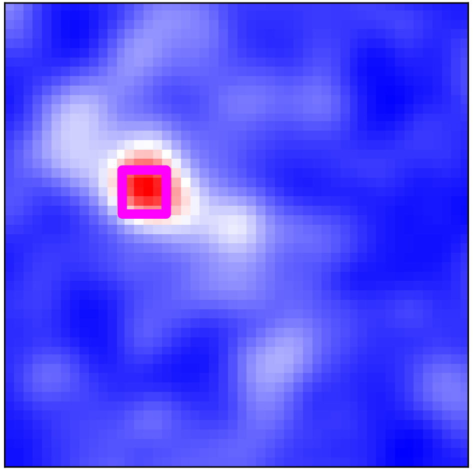
Figure 1. Both image transformations as they are applied to measure the similarity are shown exemplarily. The flipping (left) is shown on FIRSTJ075843.0+611936 and the rotation (right) is shown on FIRSTJ072529.5+614732.

Quick visual



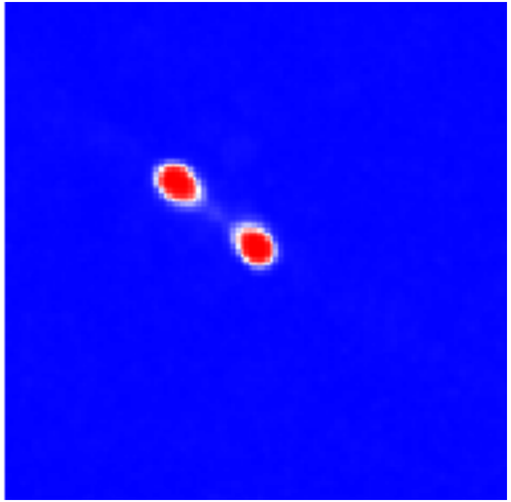




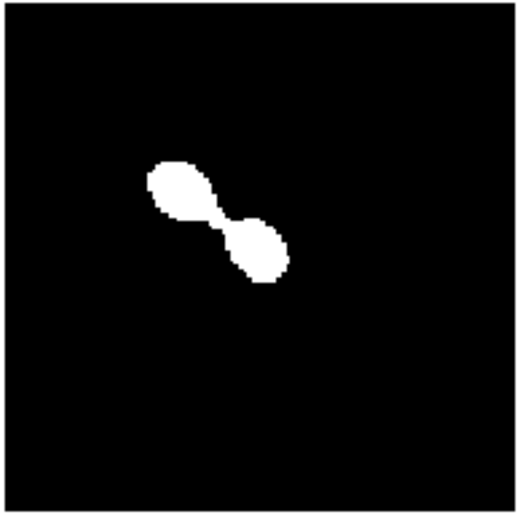


Generic filter shapes

- The neurons PINK constructs essentially represent a density of spatial intensities – a PDF
- Neurons can be treated as generic masks/filters

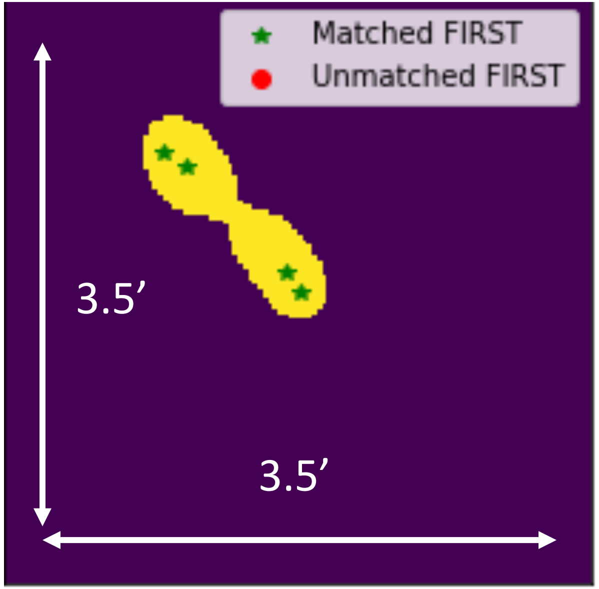


Thresholding/flood filling to isolate only related components described by annotations



Force these through catalogue space

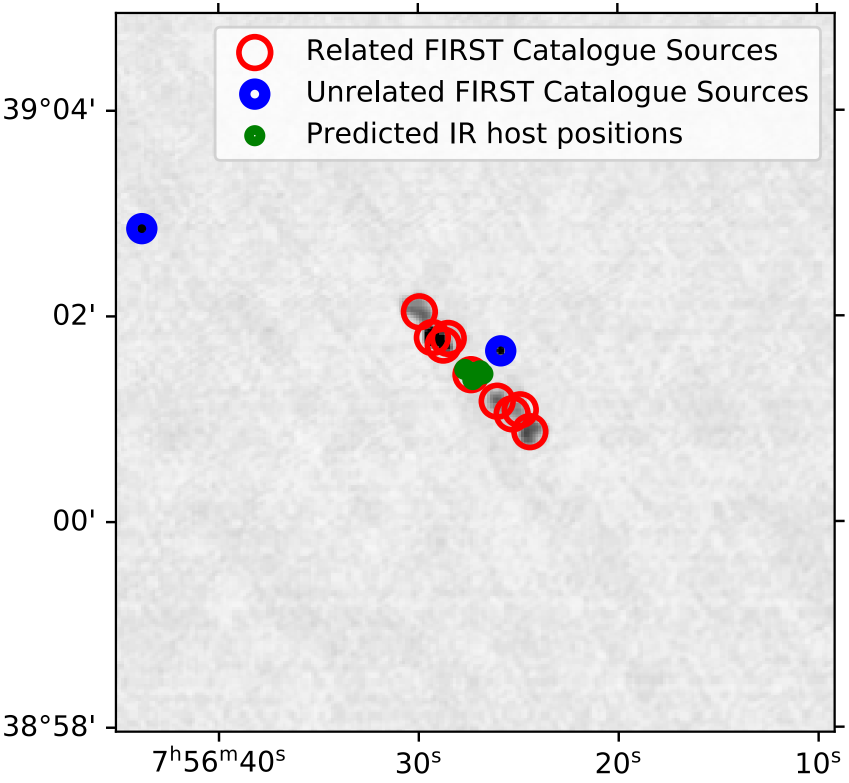
- For each source/row in catalogue
  - Identify best matching neuron
  - Obtain its filter
  - Force catalogue through it
  - Find related source components



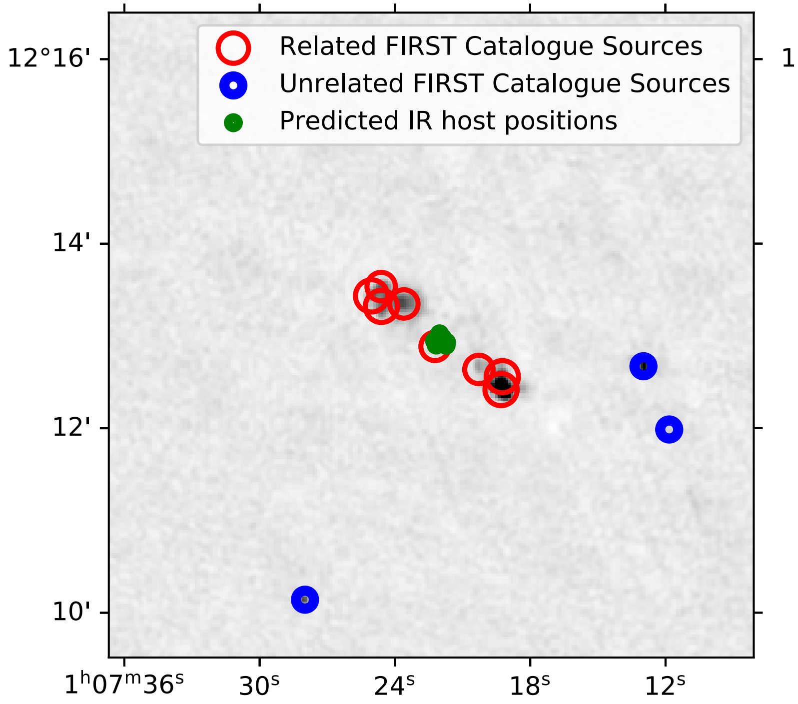
Each marker is a FIRST radio source. Those fallen within the filter are related

An example image

- Each circle is a FIRST source
- All red sources belong to a single intrinsic object, blue are unrelated near by objects
- Green mark represent IR host position
- Grouped together using *only* the products of an *unsupervised* algorithm



Another example image



- Note the size of the object
- No islands of contiguous pixels connected lobes to core
- Similar approach for WISE W1 catalogue

Conclusions

- Future radio surveys are going to be difficult
  - Data volumes too high too few people
- Machine learning obvious solution
  - Care needs to be taken – without proper labels how will we be biased?
- We have used an *unsupervised* method to group together objects
  - Exploited a dimensionality reduction tool to create meaningful classes which are then labelled -> very efficient and easily transferable
- *No* prior knowledge is required about the input data set
  - Should be applicable to any survey

Questions?