Machine Learning for Radio Astronomy: Everything is binary classification if you phrase it right

Matthew Alger

Slides: http://www.mso.anu.edu.au/~alger/radio-lunch-may

Australian National University

DATA 61

CSIRO

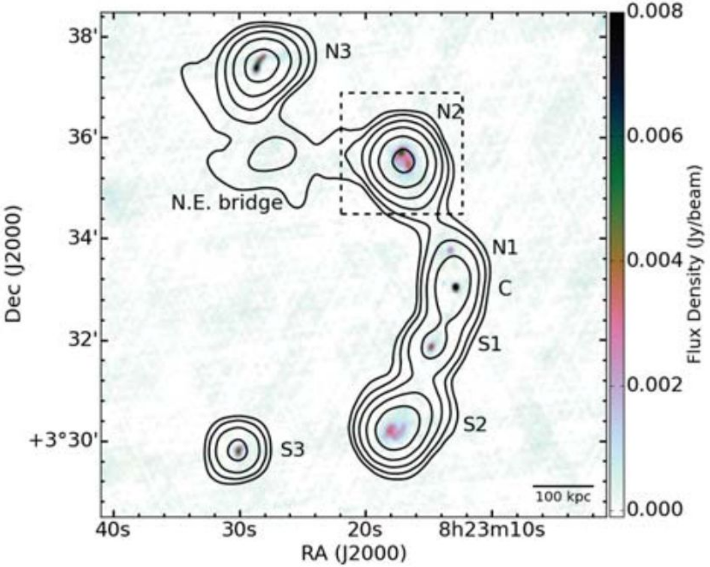
We have too much data

- Surveys like VLA-FIRST generate more data than we can look at
- Surveys like ASKAP-EMU generate more data than we can *store*

Australian SKA Pathfinder.
Image: CSIRO

The Very Large Array.
Image: NRAO

Lots of data hold lots of astrophysics

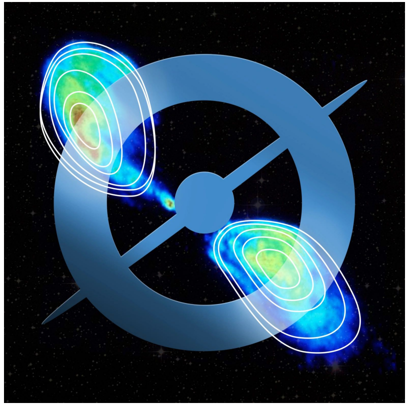


Bent, giant radio galaxy in NVSS/FIRST.
(~1 Mpc physical extent)
Image: Banfield+16

- Even 20-year old wide-area surveys like NVSS have lots of interesting astrophysics buried in them
- Much of this has come from manual inspection
- Plenty still to find

Strategies for data-at-scale

- Crowdsourcing (e.g. Radio Galaxy Zoo, Gravity Spy)
    - Fast: RGZ has classified 75000 galaxies in just 3 years
    - Serendipitous: Citizen scientists are endlessly curious
    - Noisy: Non-experts are not experts
- Asking students very nicely to look at all the data
    - Slower: Students are slow and grumpy
    - Opportunity cost: More fruitful things to do
    - Incomplete: We can't see *all* the data, so we miss things
- Machine learning
    - Fast: Computers are well-known to be quite speedy
    - Hard to interpret: Much state-of-the-art ML research is black magic
    - Unclear how to develop: Given a problem, how do we make ML work for it?



Obligatory xkcd

- “Machine learning has become alchemy.”
  — Ali Rahimi,
  NIPS 2017 Test-of-Time
- “[People] underestimate how much can be achieved with relatively crude systems”
  — François Chollet
- For useful science results, we need to understand what our methods are doing
  - Different to understanding how they are doing it!

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

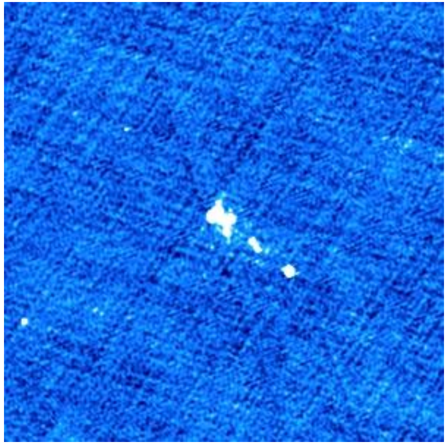
Image: xkcd

Machine learning in radio

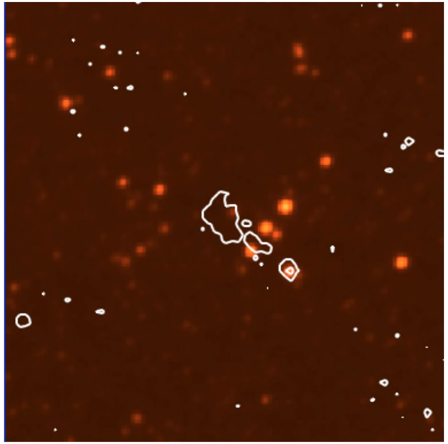
- Visualisation (Polsterer+15)
- Source classification (Aniyan+17)
- Component classification (Lukic+18)
- Host galaxy cross-identification (Alger+prep)
- Source identification (Wu+prep)

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

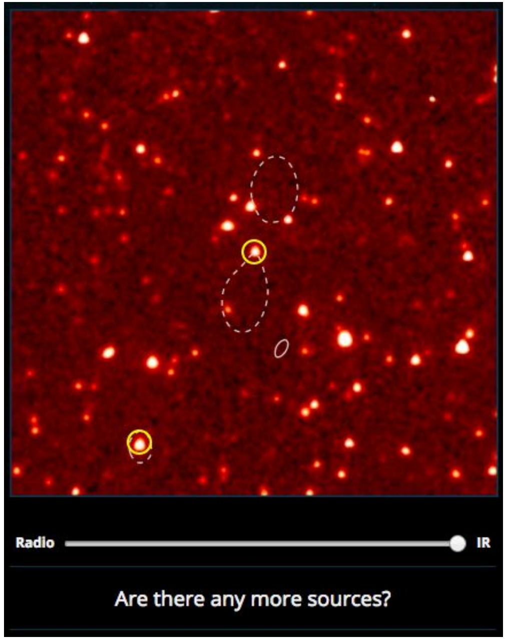


FIRSTJ023838.0+023450 at 1.4 GHz.
Image: FIRST



FIRSTJ023838.0+023450 in infrared.
Image: WISE

Machine learning can only answer some questions



Are there any more sources?

How do you turn an astrophysics question like “Where’s the host galaxy?” into a machine learning question like “Is this a 1 or a 0?”

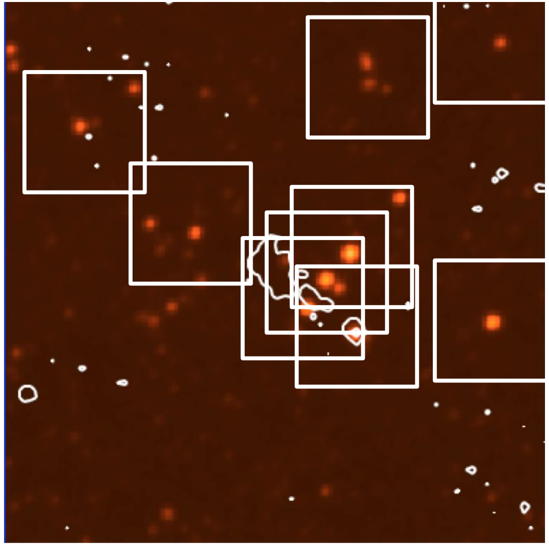
$$f : \textit{something} \rightarrow \{1, 0\}$$

$$\textit{host} : \textit{radio emission} \rightarrow \textit{IR host ?}$$

Learning to cross-identify radio emission

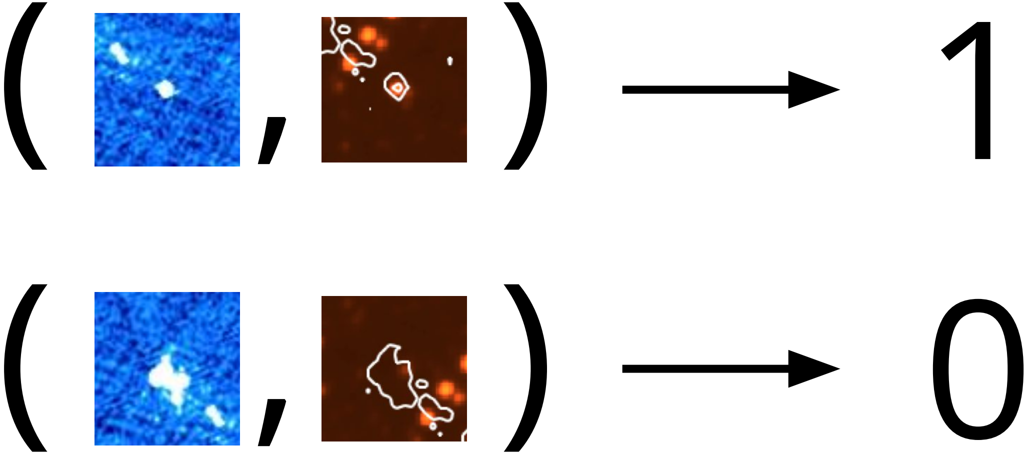
f : IR galaxy → {host, not host}

- Radio Galaxy Zoo = hosts
- Can use basic (and simple!) machine learning techniques to train and test models



Candidate host galaxies.
Image: FIRST/WISE

Cross-identification with binary classification

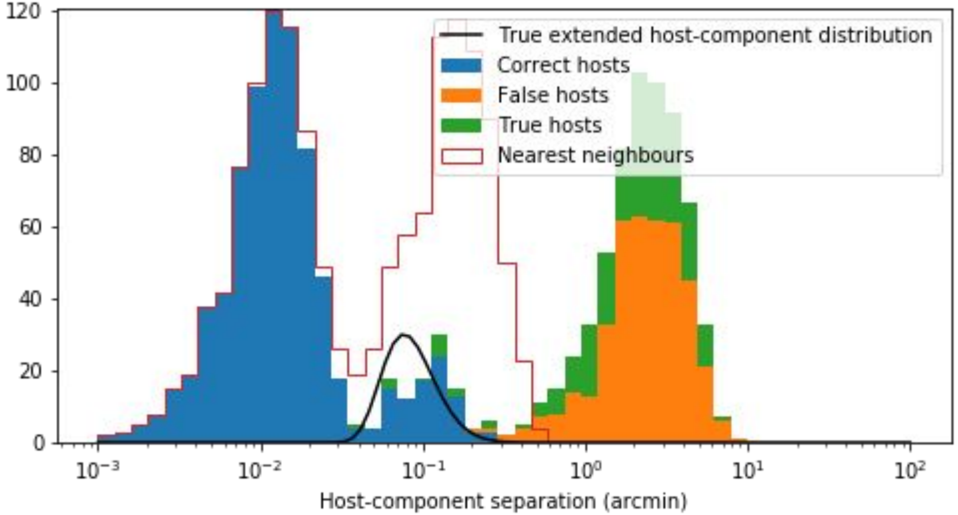


Representation of galaxy

Whether galaxy has an AGN

Understanding the link between ML and physics

- Since applying ML requires you force your problem into an ML framework, performance measures become confusing
  - e.g. classification accuracy to cross-identification accuracy
  - Uncertainties
- Evaluation and baselines important but underrated


True extended host-component distribution
Correct hosts
False hosts
True hosts
Nearest neighbours
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Host-component separation (arcmin)