

# I trained a classifier and now I don't know what to do with it

Matthew Alger

Slides: <http://www.mso.anu.edu.au/~alger/mso17>



Australian  
National  
University



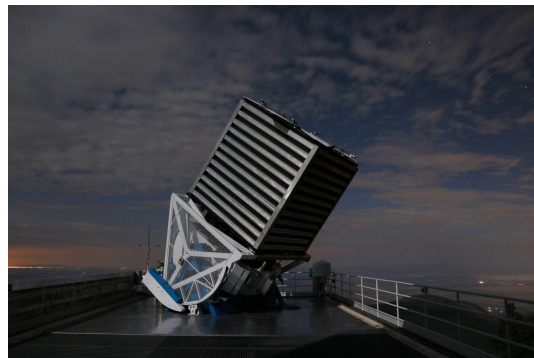
# We have too much data

- Surveys like SDSS and FIRST generate more data than we can look at
- Surveys like EMU generate more data than we can *store*
- How do we look through it all?



Australian SKA Pathfinder.

*Image: CSIRO*

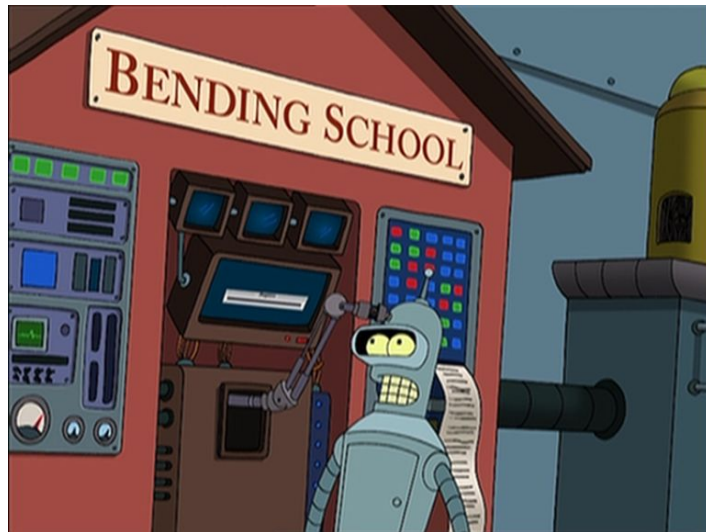


The SDSS Telescope.

*Image: Patrick Galume*

# Let's use a classifier

- A **classifier** is a function  $f: \mathbb{R}^d \rightarrow [0, 1]$
- Plenty of applications:
  - Galaxy morphology classification
  - Transient detection
  - Artefact removal



Machine learning.

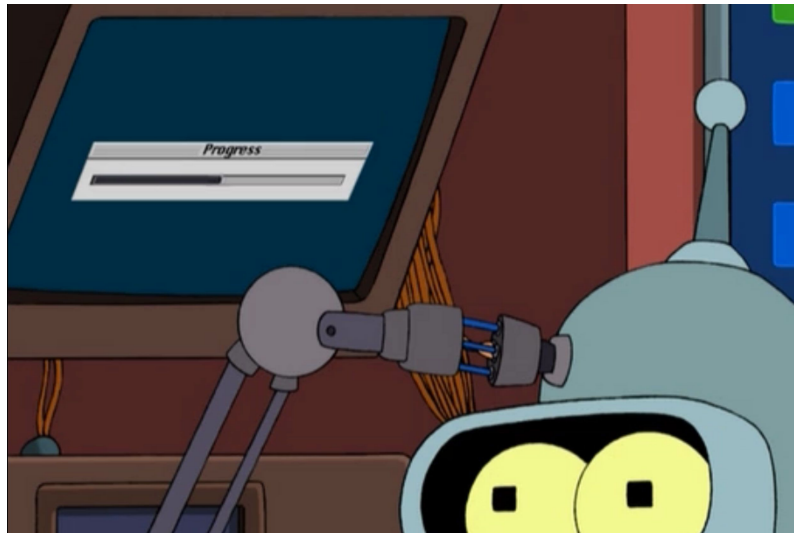
*Image: Groening et al./Twentieth Century Fox*

# Training a classifier

- Standard approach: parametrise  $f$  and find good parameters
- e.g.

$$f(\mathbf{x}; \mathbf{w}) = (1 + \exp(\mathbf{w} \cdot \mathbf{x}))^{-1}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \varepsilon \nabla L_{\mathbf{w}}$$

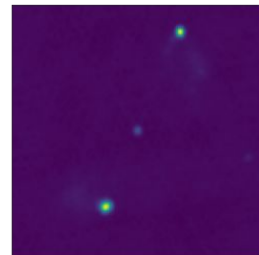
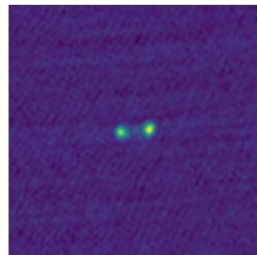
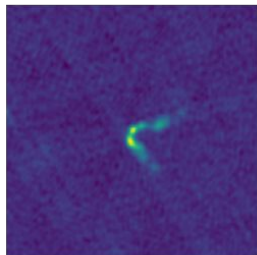
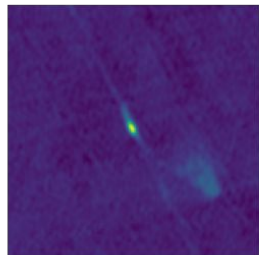
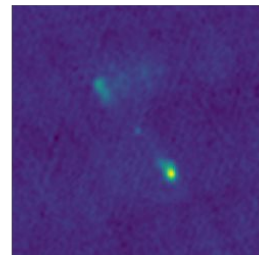
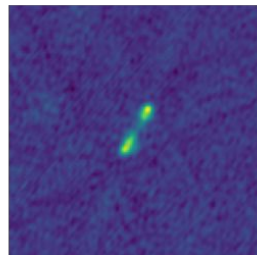
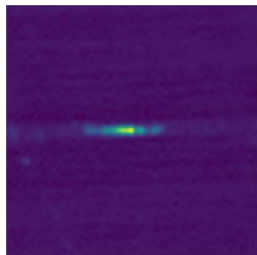
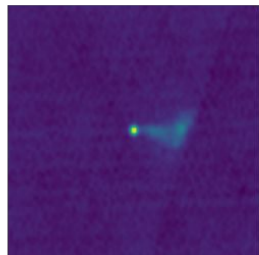


Many tasks can be learned through gradient descent.

*Image: Groening et al./Twentieth Century Fox*

# Classifying radio galaxies

- Task: Classify radio galaxies as *Fanaroff-Riley Type I* or *Fanaroff-Riley Type II*
- $f(\mathbf{x})$  outputs a number closer to 0 for FR-I and 1 for FR-II



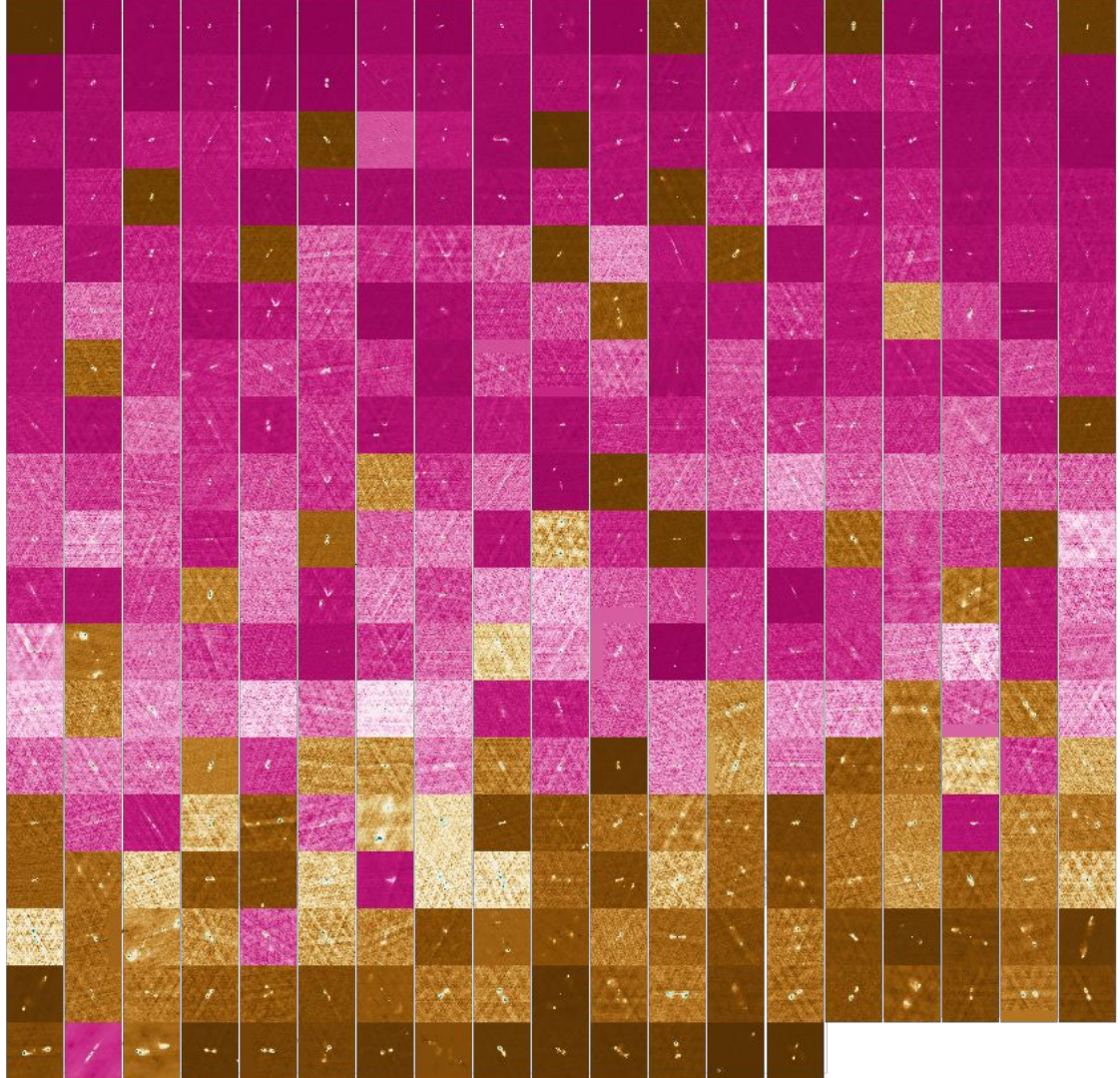


# Sorting by $f(\mathbf{x})$

- We can sort the radio galaxies by the output of  $f(\mathbf{x})$
- What do different parts of the list tell us?

Radio galaxies sorted by  $f(\mathbf{x})$ . Orange galaxies are FR-IIs, while pink galaxies are FR-Is.

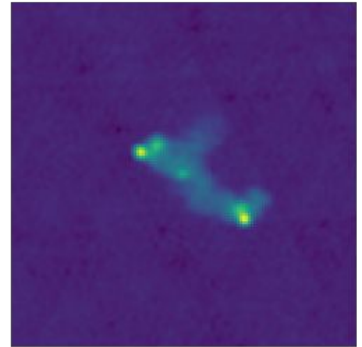
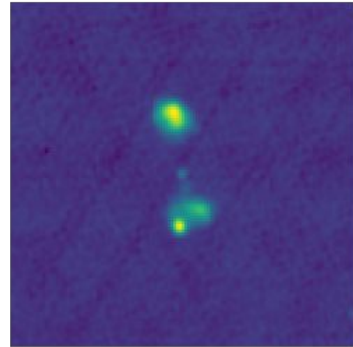
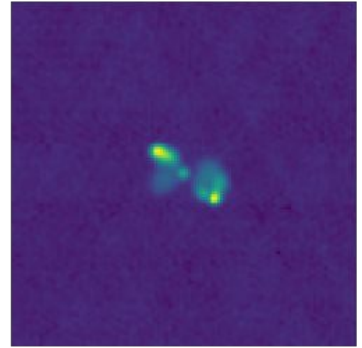
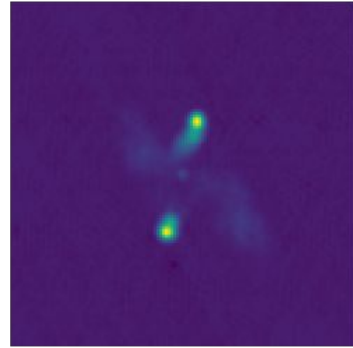
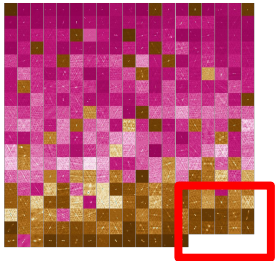
*Images: FIRST*



# The top end

- Maximum  $f(x) \rightarrow$  most like an FR-II (or least like an FR-I?)

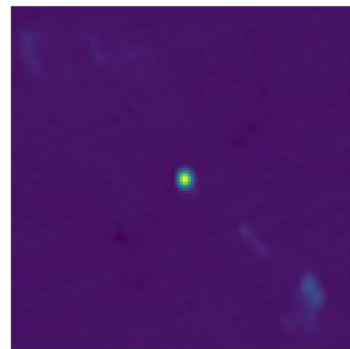
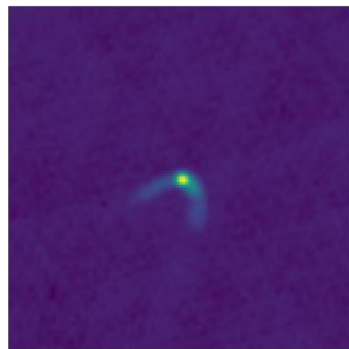
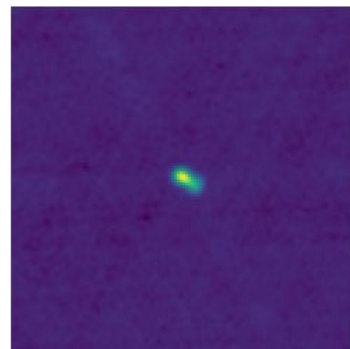
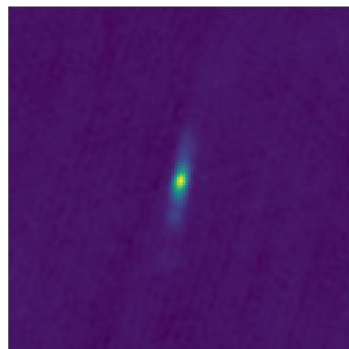
Maximally  
Fanaroff-Riley II



# The bottom end

- Minimum  $f(\mathbf{x}) \rightarrow$  most like an FR-I (or least like an FR-II?)

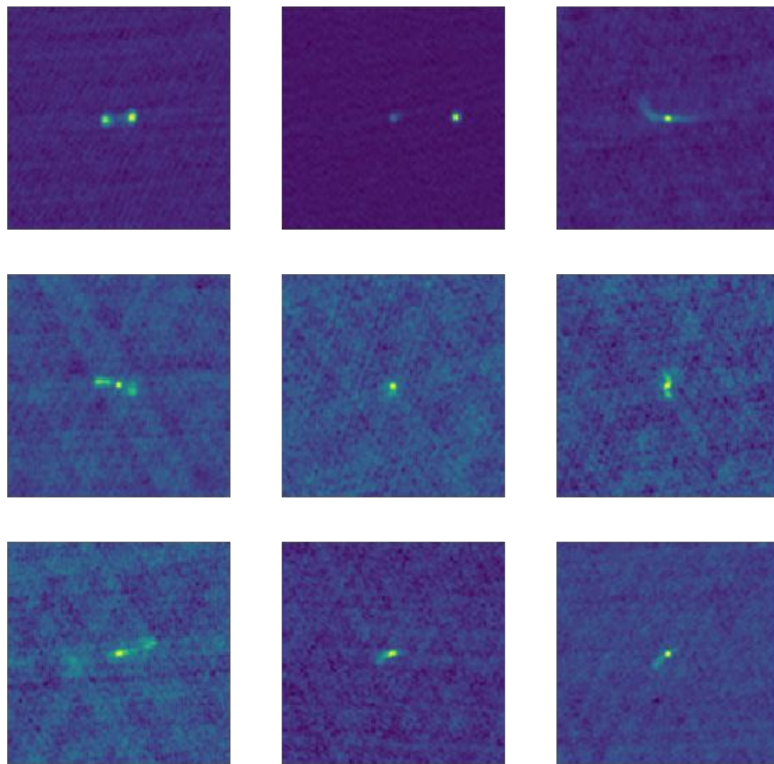
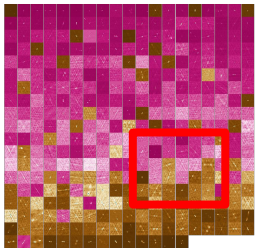
Maximally  
Fanaroff-Riley I





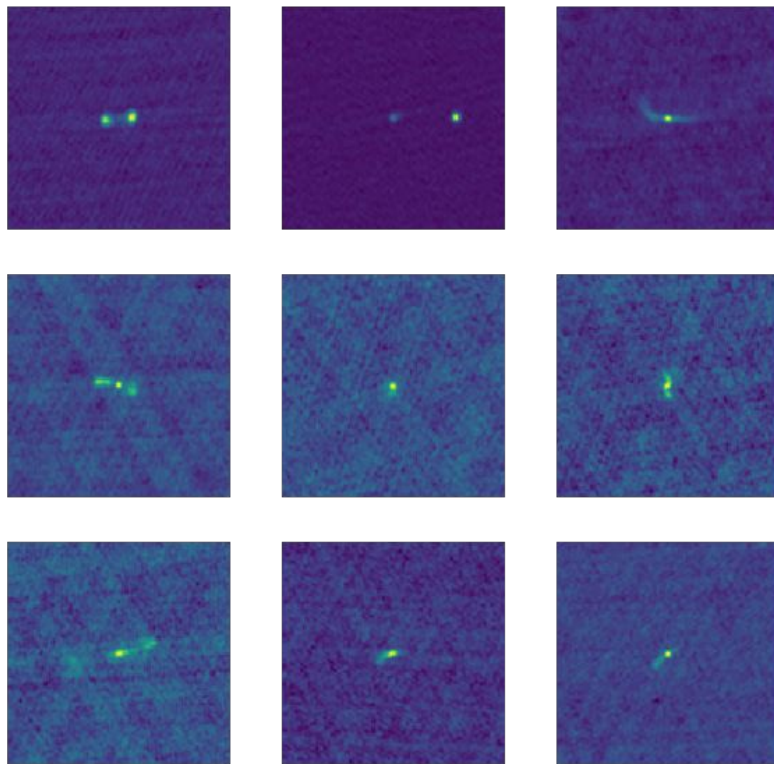
# The middle

- Uncertain objects
  - Not quite like an FR-I
  - Not quite like an FR-II

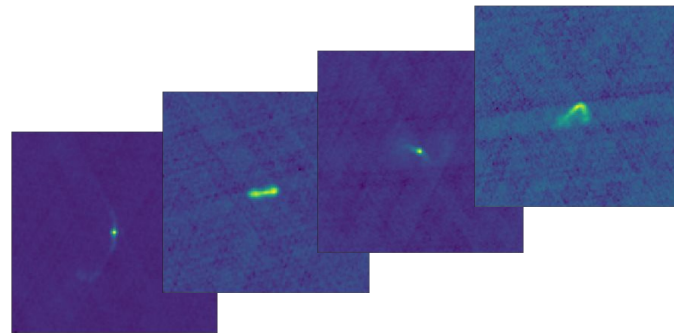
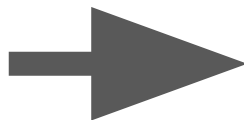
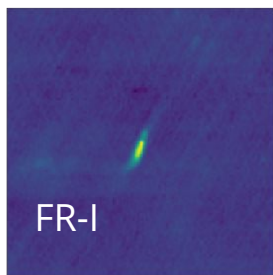
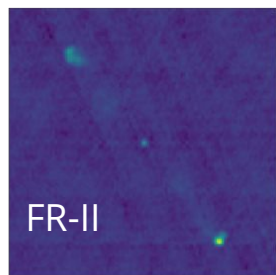


# The middle

- Physically interesting?
  - Are these really halfway between an FR-I and an FR-II?
  - Are these uncertain for some meaningful reason?
- Interesting for learning?
  - Does something about these objects confuse the classifier?
  - If training focused more on objects like this, would we get a better classifier?

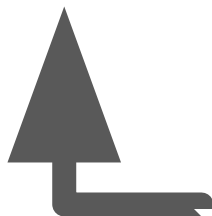


# Building a better classifier

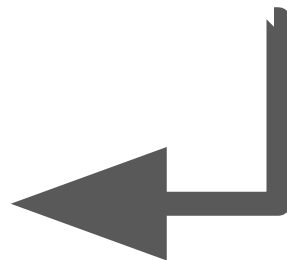
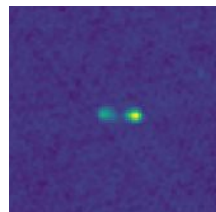


Train a classifier with  
labelled data

Classify unlabelled data

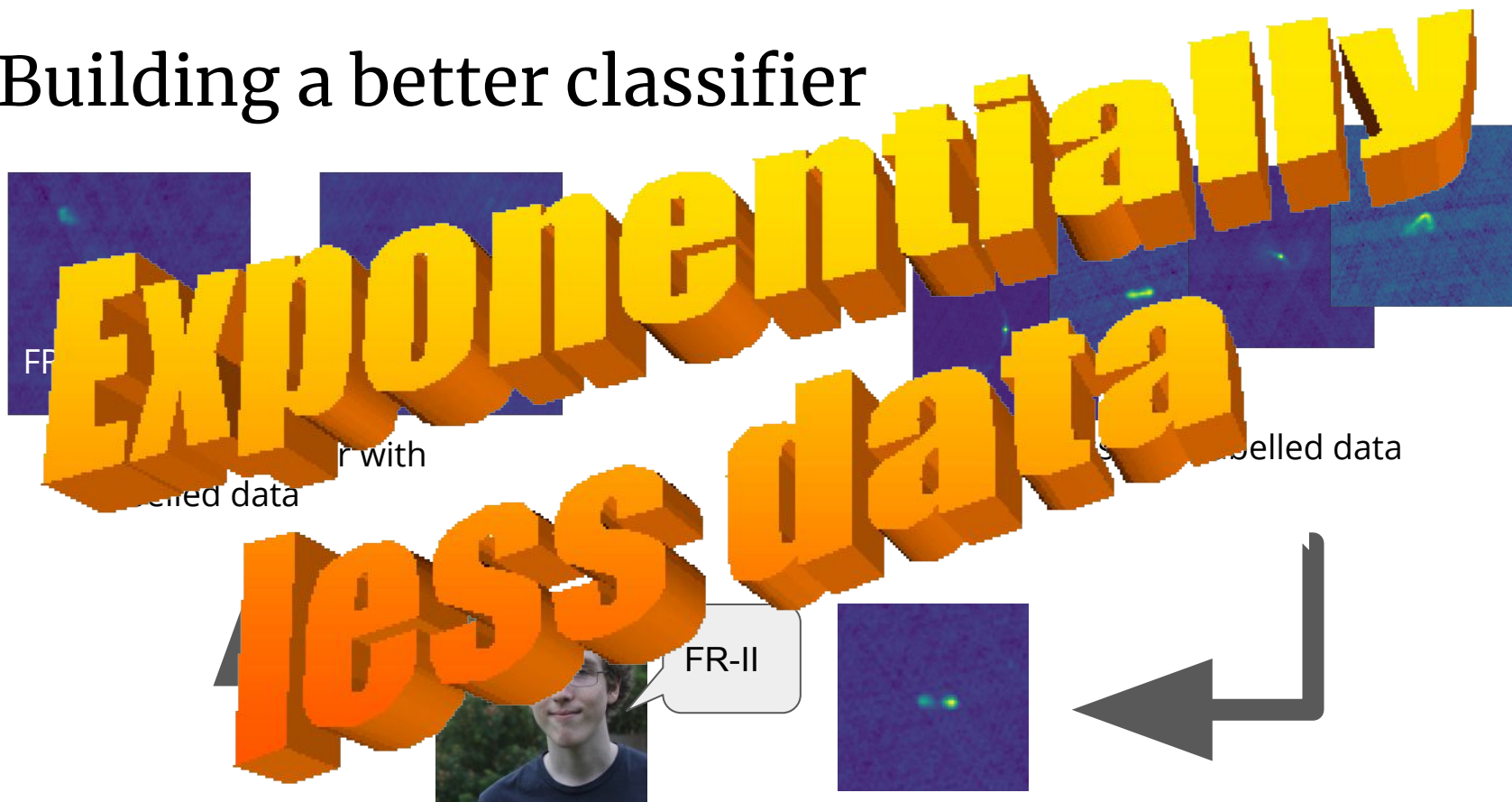


FR-II



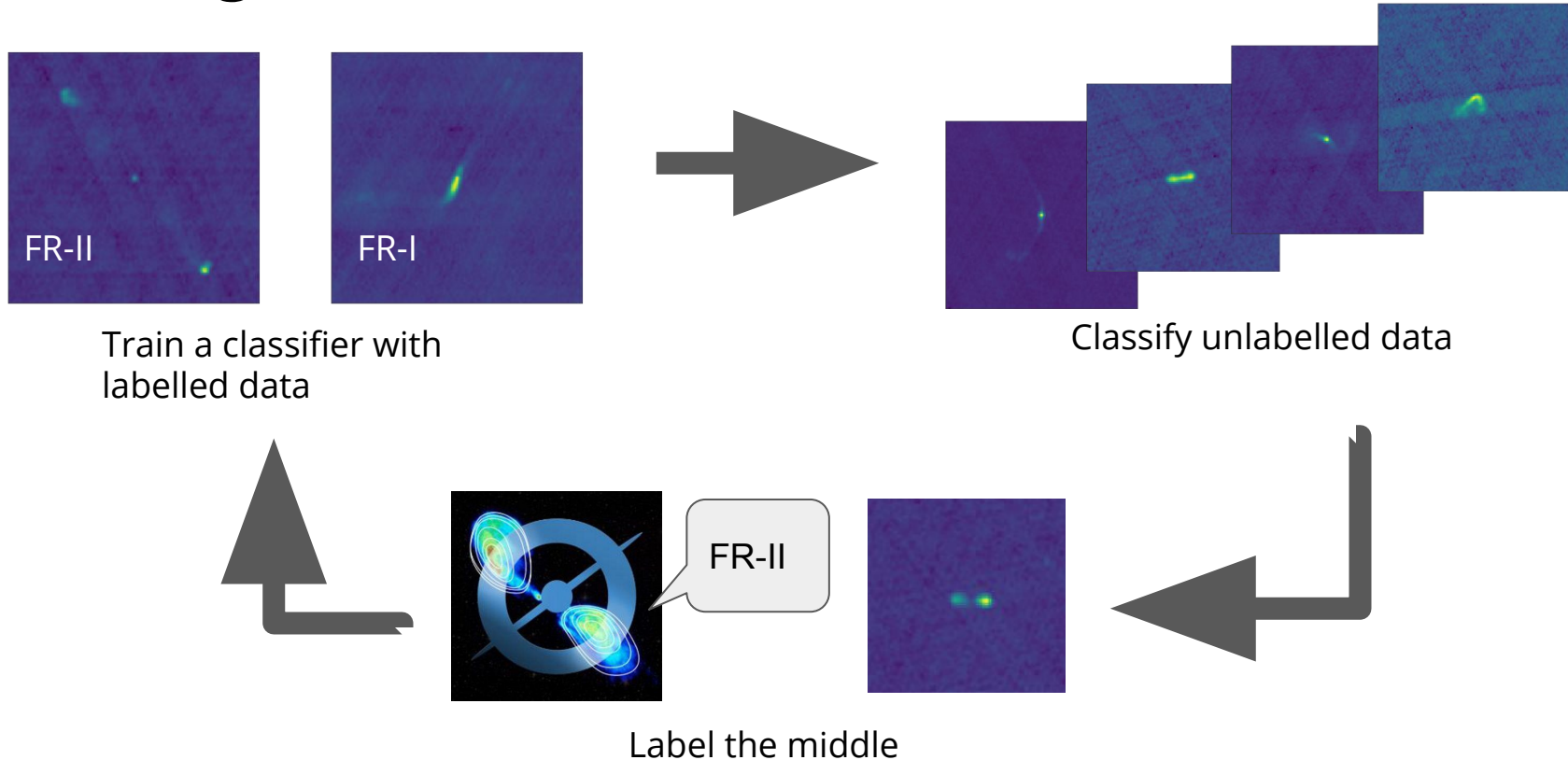
Label the middle

Building a better classifier



Label the middle

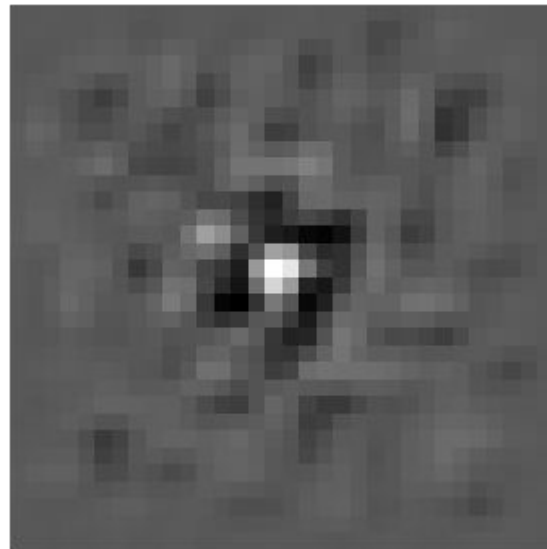
# Getting better results from citizen science





# Analysing the classifier

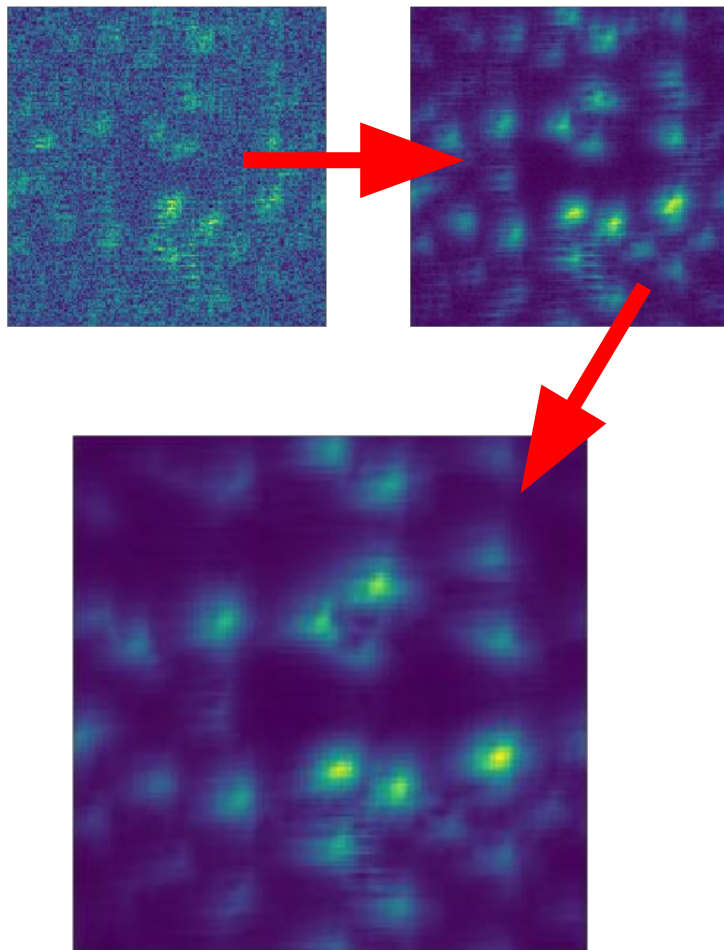
- How do we know what our black box is doing?
- If the classifier is differentiable, differentiate it



Gradients indicate how much each pixel contributes to the “FR-II-ness” of an image.

# Maximising $f(\mathbf{x})$

- If your classifier is differentiable, you can differentiate it
- Use the gradient to make your inputs more like the target class
- $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \varepsilon \nabla f_{\mathbf{x}}$
- ...But a classifier's idea of the target class might be different to yours



# What can you do with a classifier?

- Classify objects
- Sort a list of objects
  - Top and bottom of list tell you classes you care about
  - Middle of list provides “interesting” cases
  - “Interesting cases” useful for learning and science
- Analyse your classifier
  - Classifier may not be looking for what you expect