Learning to Cross-identify Wide-area Radio Surveys with Radio Galaxy Zoo: Data challenges in citizen science

Matthew Alger (ANU/Data61)
Julie Banfield (ANU)
Cheng Soon Ong (Data61/ANU)

Slides: http://www.mso.anu.edu.au/~alger/c3dis

Australian National University

What is citizen science?

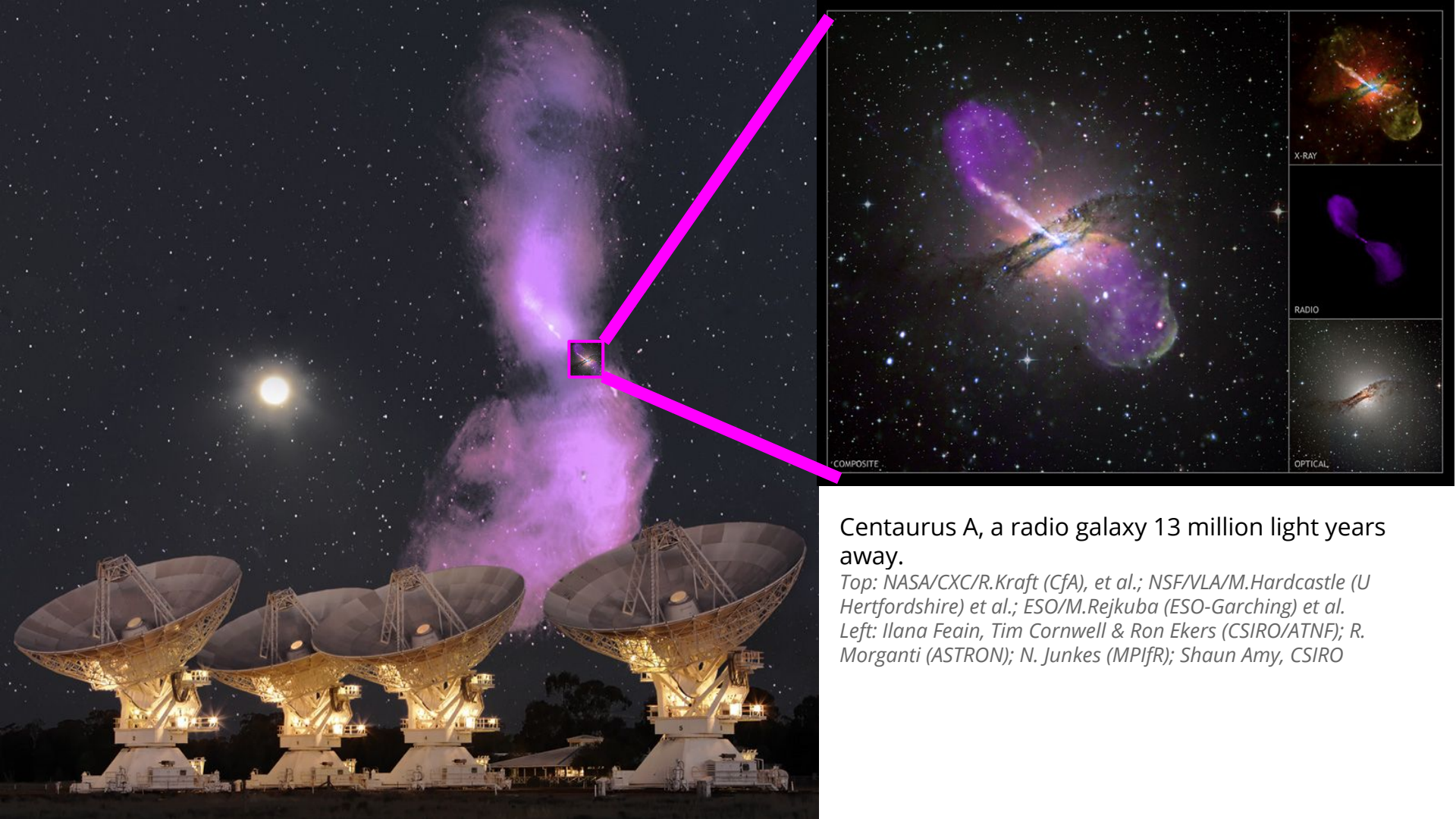
“Citizen science projects involve non-professionals taking part in crowdsourcing, data analysis, and data collection. The idea is to break down big tasks into understandable components that anyone can perform.”

— Robert Simpson (Zooniverse), The Conversation, 15/08/13

Validating machine learning models

- Assume we have a performance measure (hard!)
  - Call this “accuracy”
- Source validation data from citizen science
- Compute accuracy by comparing predictions to validation data

How does citizen science affect our validation?

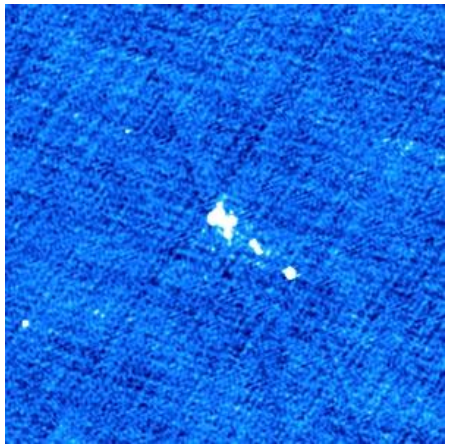


Centaurus A, a radio galaxy 13 million light years away.

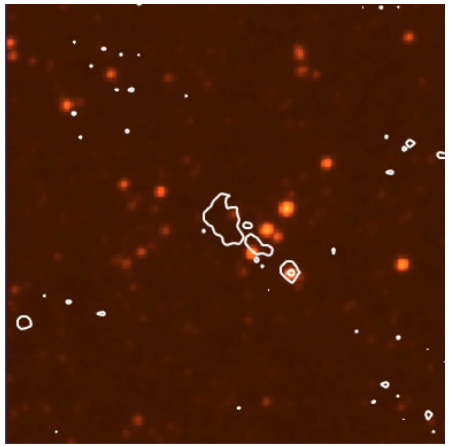
Top: NASA/CXC/R.Kraft (CfA), et al.; NSF/VLA/M.Hardcastle (U Hertfordshire) et al.; ESO/M.Rejkuba (ESO-Garching) et al.
Left: Ilana Feain, Tim Cornwell & Ron Ekers (CSIRO/ATNF); R. Morganti (ASTRON); N. Junkes (MPIfR); Shaun Amy, CSIRO

Case study: radio/infrared cross-identification

- Problem:
  - Match radio emission to the corresponding galaxy in infrared
  - Important for understanding galaxies throughout cosmic time
- Hard:
  - Radio emission can be very extended across the sky
  - Often no clear relationship between radio emission and the emitting galaxy



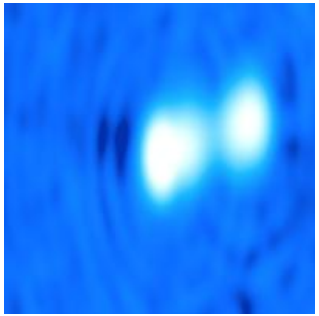
A radio galaxy imaged with the VLA, a radio telescope.
Image: FIRST



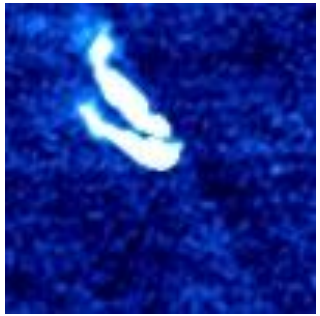
The same patch of sky imaged with *WISE*, an infrared telescope. White contours show the radio image on the left.
Image: WISE

Radio Galaxy Zoo (http://radio.galaxyzoo.org/)

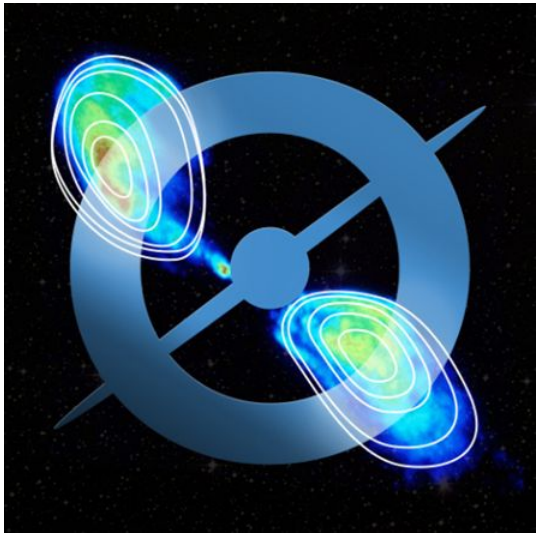
- Crowdsourced, citizen science project
- ~75000 labelled radio objects in 3 years
- Volunteers identify infrared galaxy counterparts to radio emission

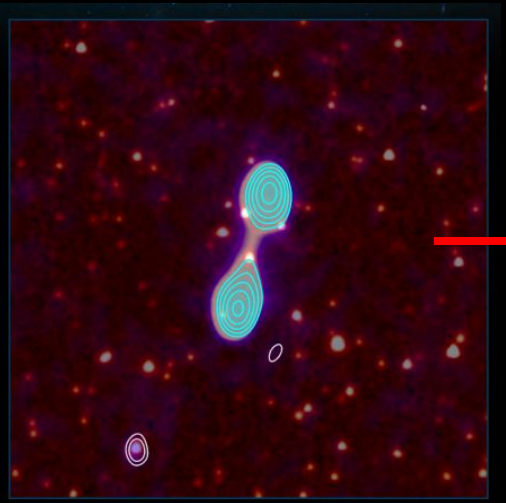


An image from the Australia Telescope Compact Array.



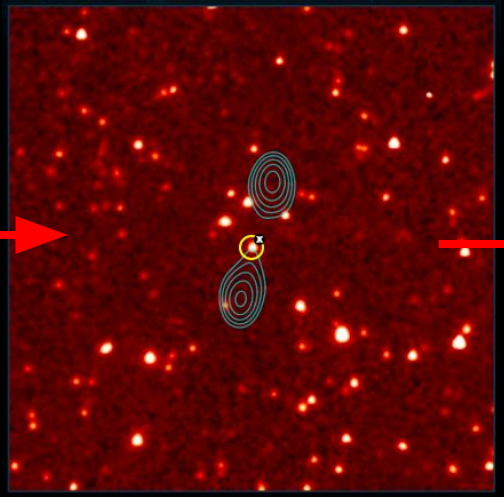
An image from the Very Large Array.





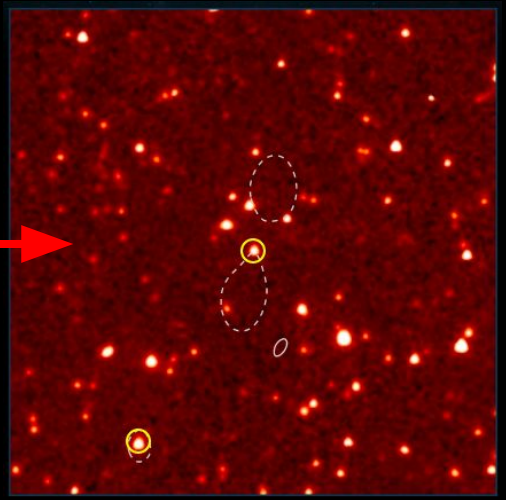
Radio IR

Click on any radio contour or pair of jets



Radio IR

Click the associated infrared source(s)

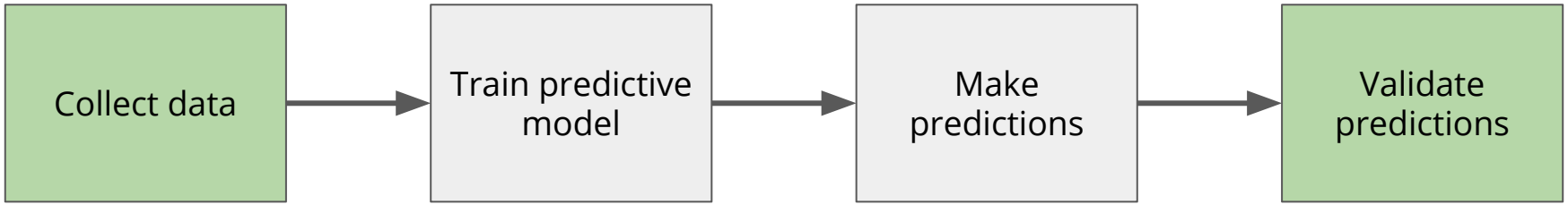


Radio IR

Are there any more sources?

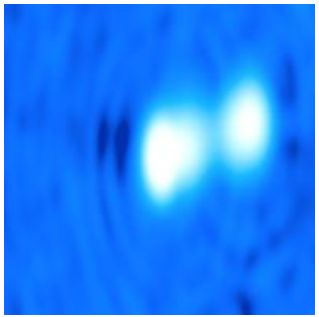
Data quality for machine learning

- Data should be
  - Accurate
  - Complete: we want to sample the full data space
  - Close to its usage
  - Understandable: we want to know when it’s wrong
- Issues in any of these affect training and evaluation

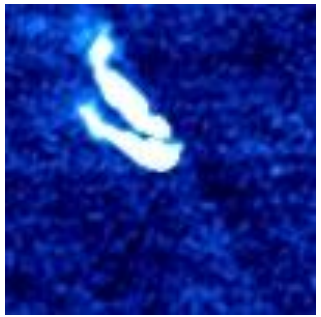


Citizen scientists are noisy

- Validating citizen scientist labels is already a hard task
- We don’t know how!
  - Are some things intrinsically hard to label?
  - Are volunteers better at certain tasks?
  - Do volunteers “overfit” to some datasets?



Accuracy: 30%

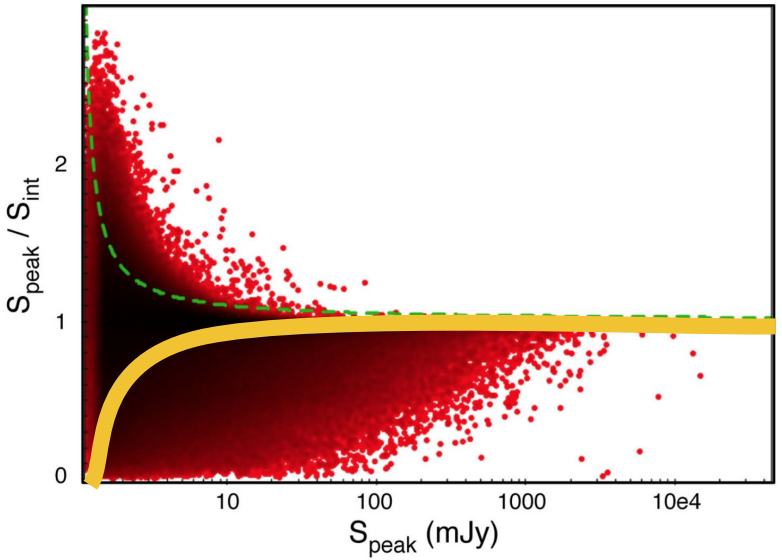


Accuracy: 80%

Accuracy of Radio Galaxy Zoo varies between ATCA (left) and VLA (right) observations. But why?

Citizen science only examines interesting objects

- We only show interesting objects
- At odds with good validation
  - Citizen science projects by design have different distributions to target distribution
  - Validation will favour particular parts of the feature space
  - Unexpected results: ML methods may fail on “simple” cases!

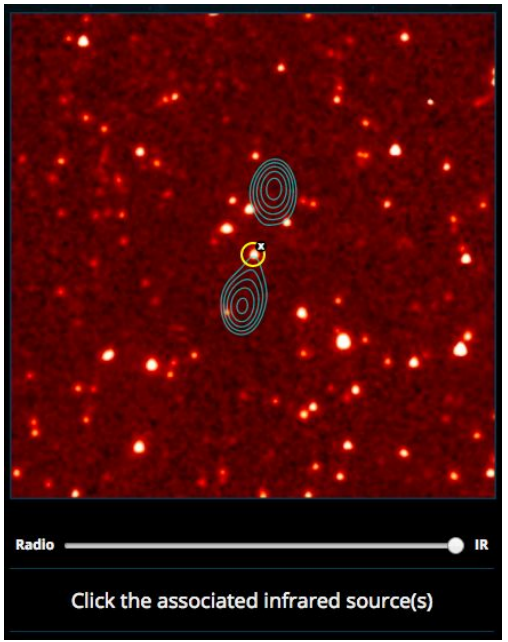


Radio Galaxy Zoo only shows objects below the gold line to volunteers.

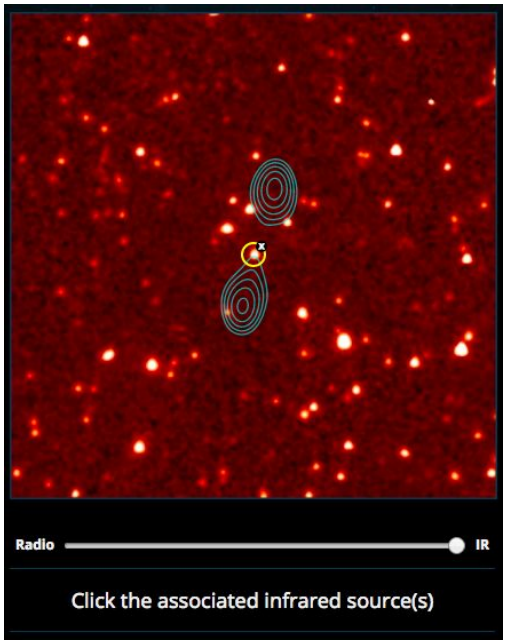
Image: Adapted from Banfield+15

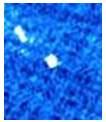
We don’t get the labels we need

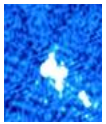
- People are inherently better at some tasks than others
- We present non-experts with suitable tasks that don’t require domain knowledge
- Three distinct contexts:
  - Tasks volunteers are good at
  - Tasks ML is good at
  - Physics questions we want to answer
- Are we validating the method, or validating our data transformation?



Different tasks for volunteers and machine learning



( , ) → 1

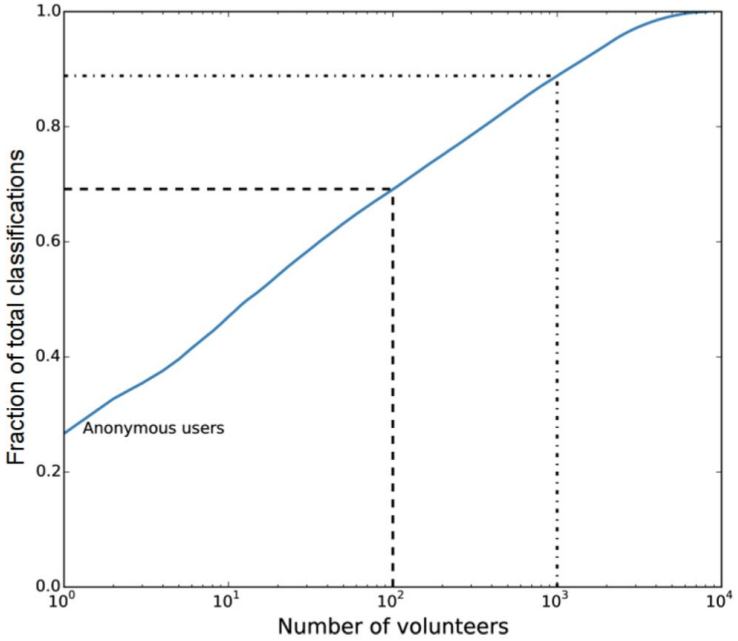
( , ) → 0

Representation of galaxy

Whether galaxy has an AGN

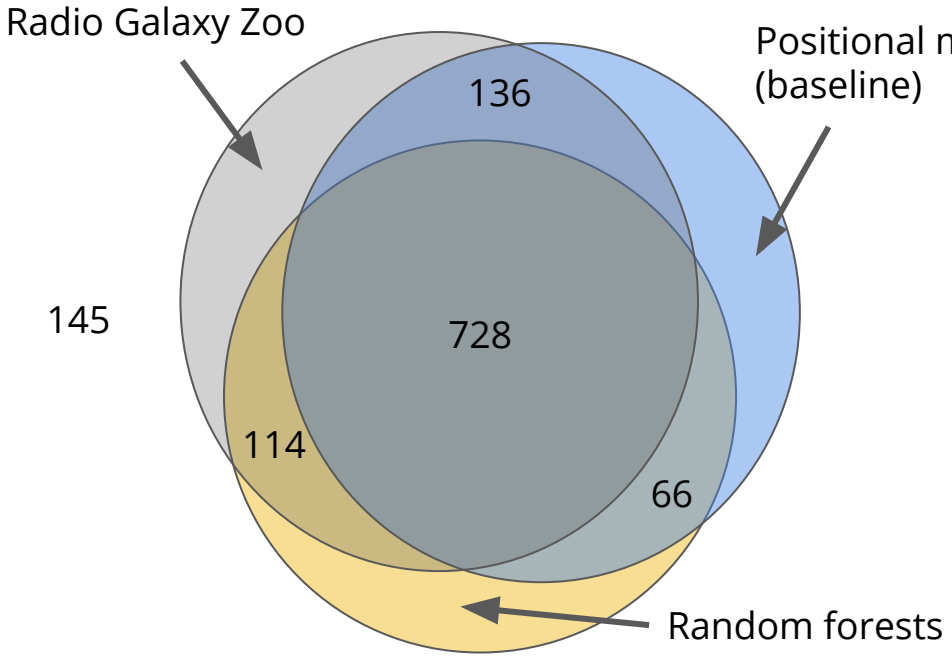
Noise is hard to characterise

- Different tasks for citizen scientists and machine learning
  - Even when citizen science noise is quantified this doesn't map clearly to label noise or uncertainty
- 10,000s of annotators make standard label noise estimation hard
  - Joint groundtruth-model estimators fail

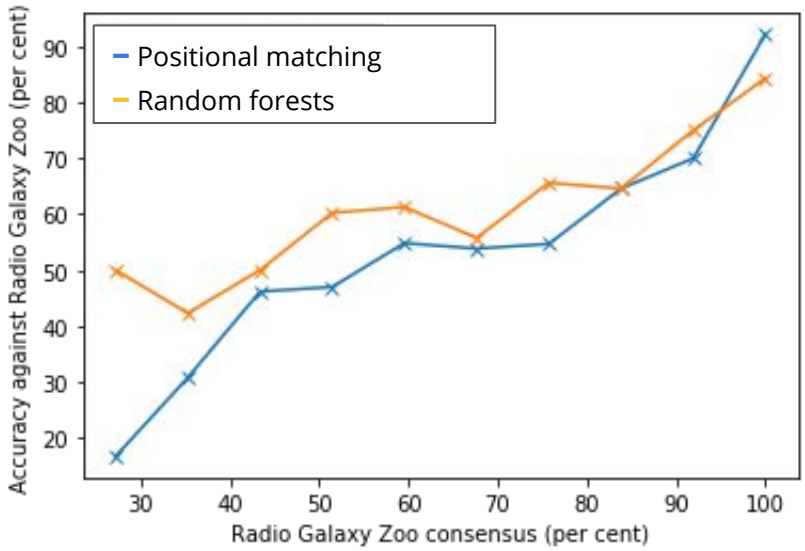


Cumulative distribution of classifications.
Image: Banfield+15

Evaluating our model on Very Large Array data



Agreement of different cross-identifiers.



Accuracy against volunteer agreement.

Summary

- Using citizen science data to validate our models introduces unique challenges and noise
- Good citizen science tasks are not good machine learning tasks
  - Bias toward interesting objects
  - Different kinds of labels

References

- Banfield et al. (2015) DOI: 10.1093/mnras/stv1688
- Alger et al. (in press) DOI: 10.1093/mnras/sty1308
- Wong et al. (in prep)

Related talks

- Minh Huynh: *PINK, a Self-Organised Map for Radio Astronomy*
  Room 105/106, 13:50
- Ray Norris: *Discovering the Unexpected in Radio-astronomy data*
  Room 104, 14:10
- Ray Norris: *Machine Learning in Astronomy*
  Room 105/106, 16:30