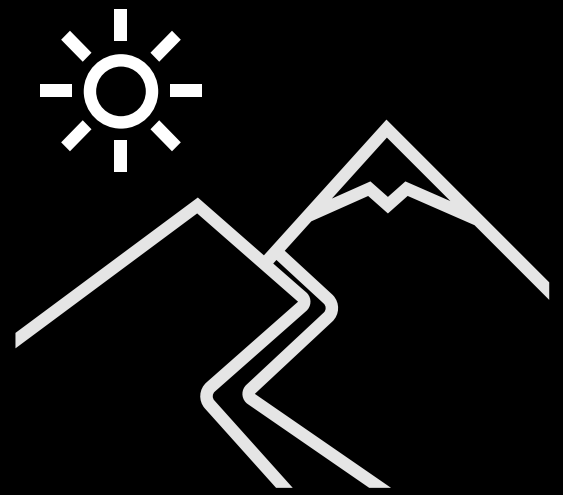
Astrotag: #Tagging Radio Galaxies using Machine Learning

Vinay Kerai | Supervisors: Dr. O Ivy Wong, Dr Cheng Soon Ong, Dr Dawei Chen, Dr Zeyi Wen

Roadmap

- RadioTalk Dataset and Motivations
- Research Aims
- Methods
- Results
- Conclusion and Future Work



What is a Radio Galaxy?

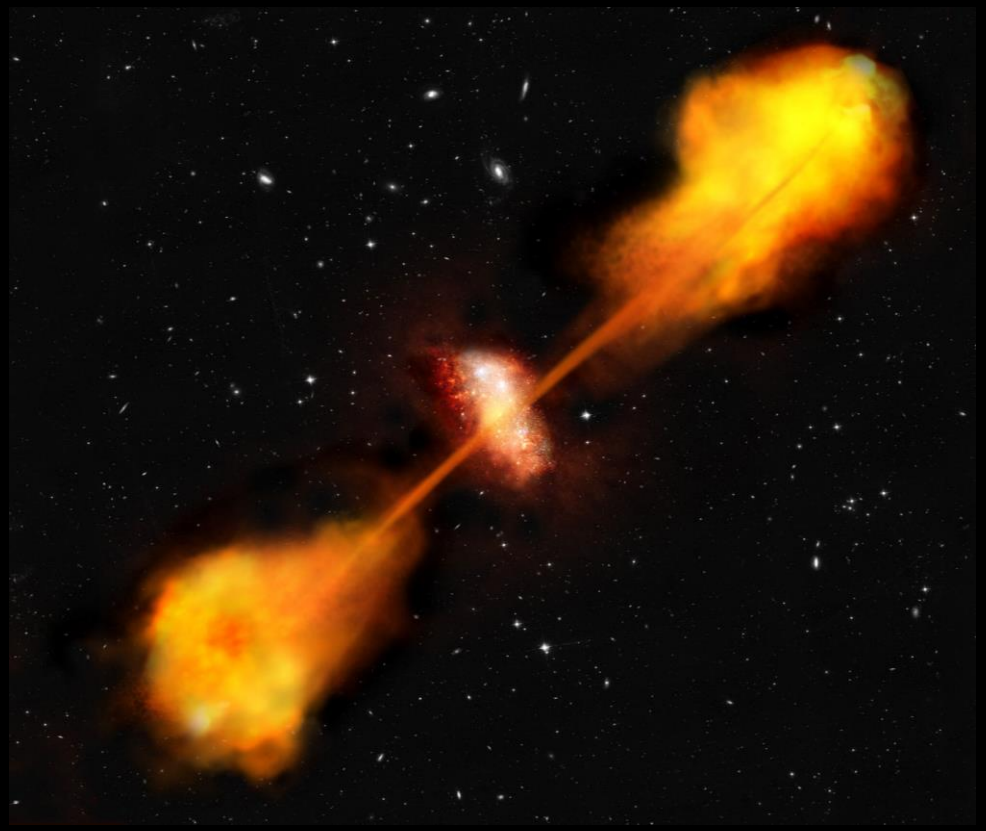


Figure 1: A Radio Galaxy

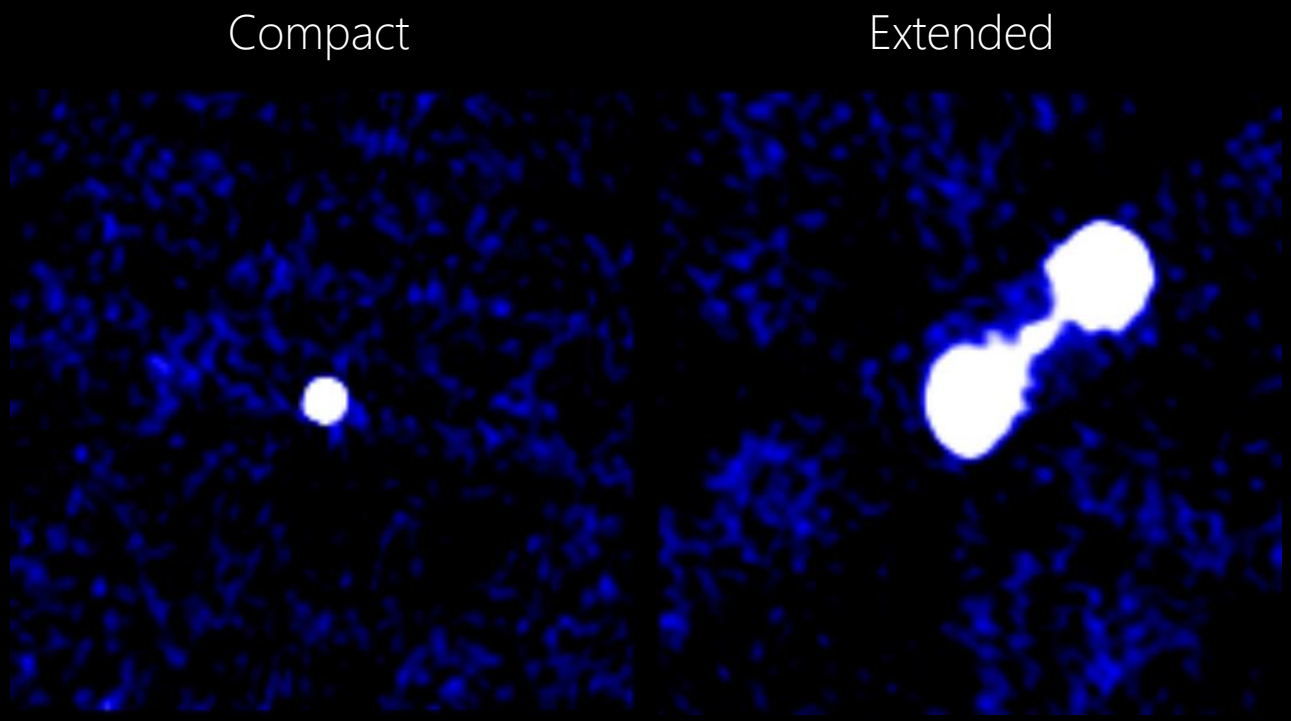


Figure 2: Compact vs Extended Sources

Labelling Extended Radio Sources

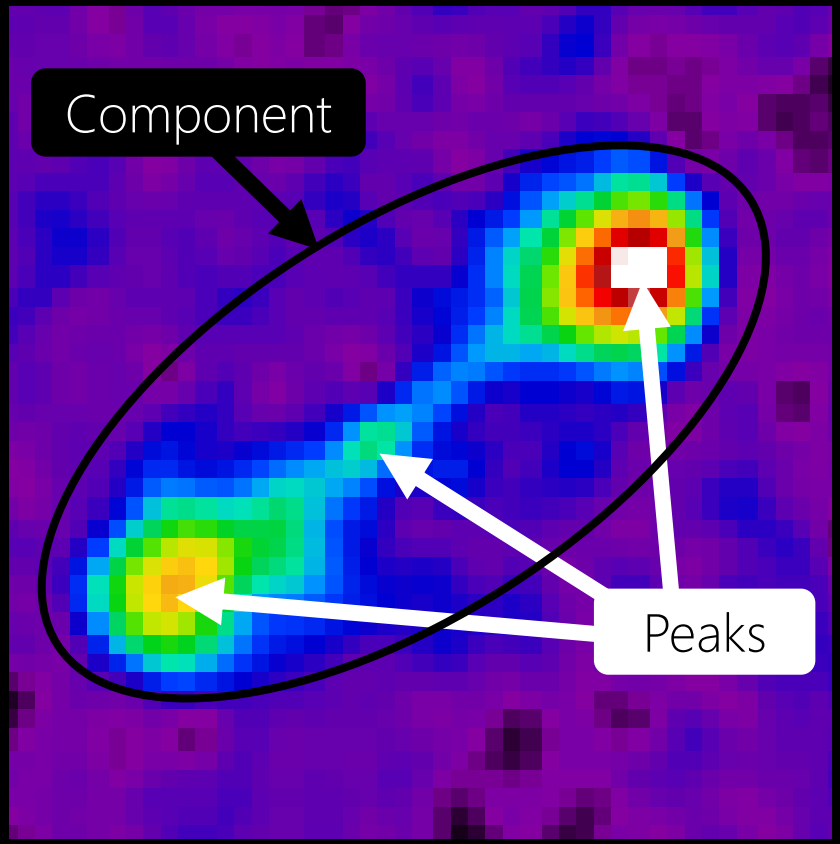


Figure 3: Labelling a Radio Source as 1C_3P

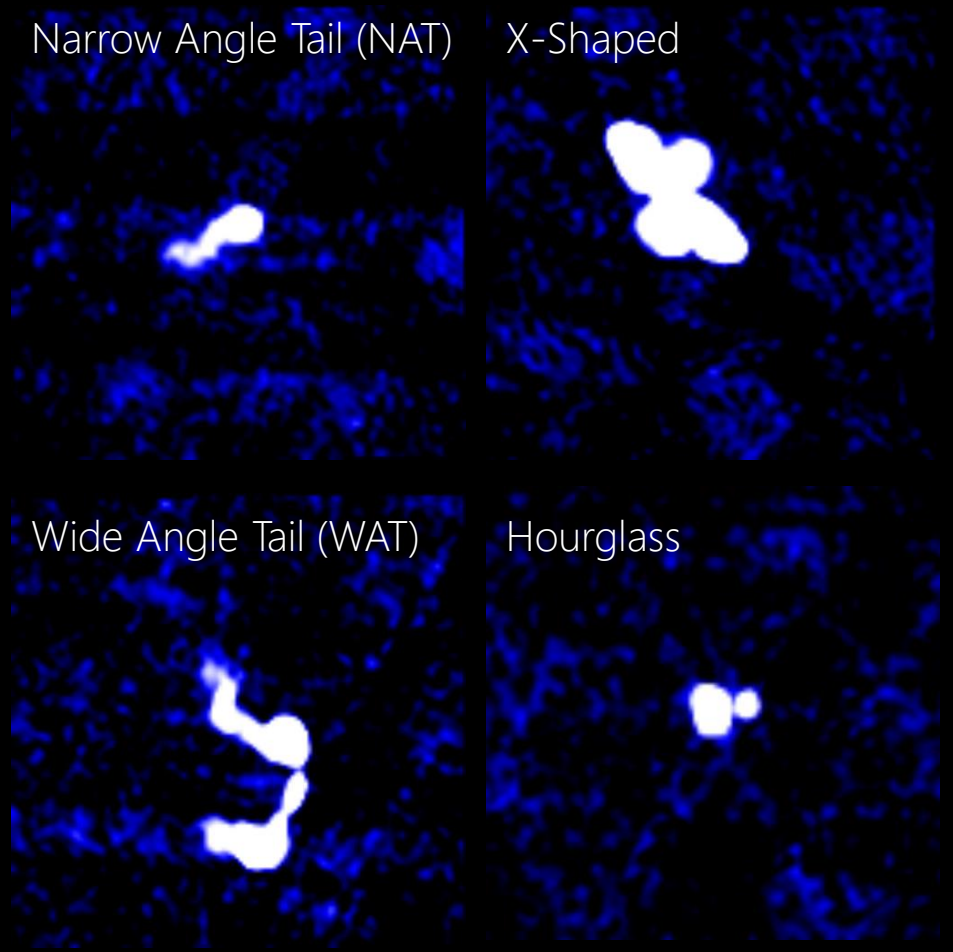


Figure 4: Sample of Unique Morphologies

RadioTalk Dataset

- Radio Galaxy Zoo (RGZ) utilised citizen science to build a catalogue of over 170,000 radio sources.
- Volunteers were asked to locate and label any radio sources present in a Radio Subject.
- RadioTalk is a platform for keen volunteers to provide more descriptive labels in the form of tags and comments.
- Additional descriptions for >30,000 Radio Subjects are available via RadioTalk and currently not present in the catalogue.

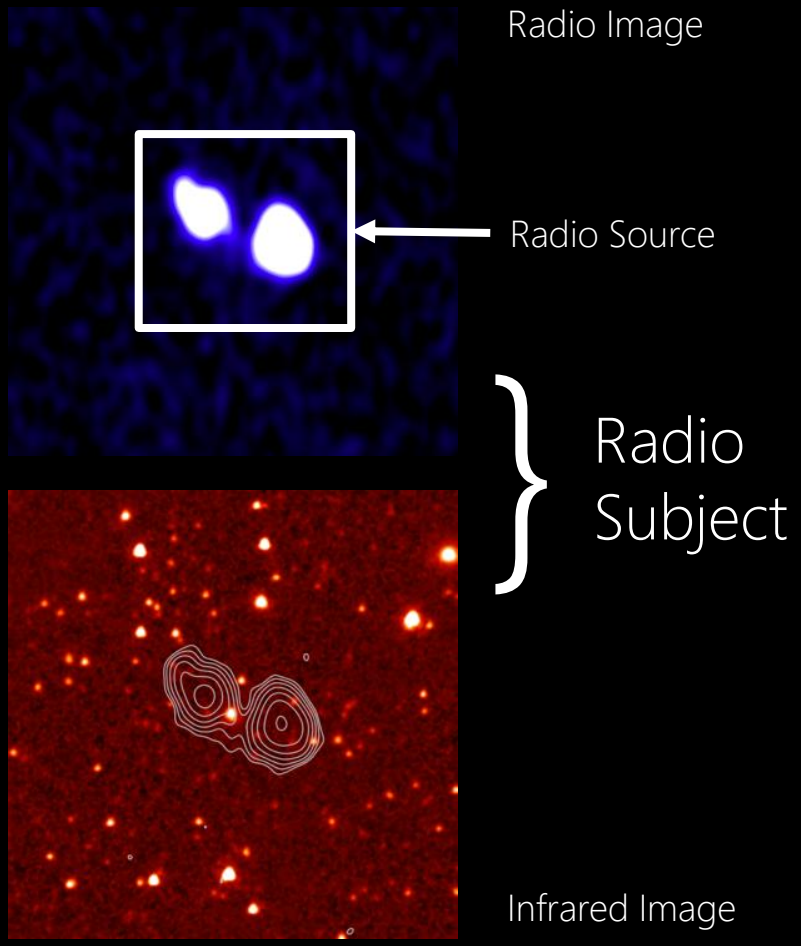


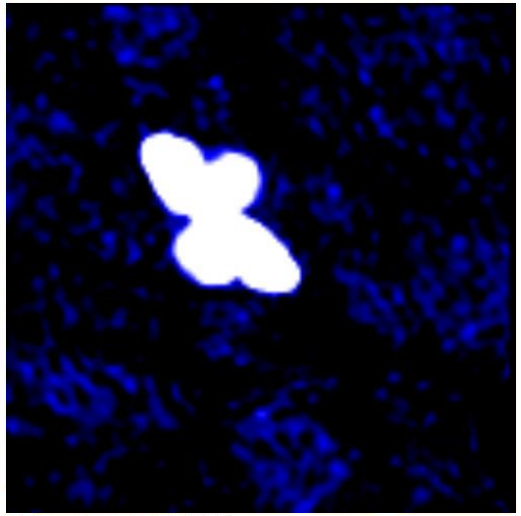
Figure 5: Breakdown of Radio Subject

Radio Subject Threads

Radio Galaxy Zoo Talk

RECENT DISCUSSION BOARDS

Subject: ARG0002ato



Full subject data (JSON)

TAGS

butterfly, 4lobes, nice, largesource, symmetric, x-shaped, quad, goldstandard, reflection, messy, xshaped, closerlook, beautiful, large, cloverleaf

COMMENTS

by antikodon

fat hourglass?

Posted December 26, 2013 11:12 AM

BOARD DISCUSSIONS

CHAT

SCIENCE

COMMENTS

by antikodon

fat hourglass?

Posted December 26, 2013 11:12 AM

by Dolorous_Edd

#butterfly?

Posted December 26, 2013 11:16 AM

by antikodon

or bent doublelobe?

Posted December 26, 2013 11:17 AM

by ivywong SCIENTIST, ADMIN

Wow! This is an #Xshaped source! This could be a double black hole system with 2 separate jets! Most awesome find 😃

Posted December 26, 2013 11:39 AM

by Dolorous_Edd

here is another one

http://radiotalk.galaxyzoo.org/#/subjects/ARG0002ztd

Posted December 26, 2013 11:41 AM

by bretarn

just WOW.😃

Posted December 31, 2013 5:25 PM

Figure 6: An example Radio Subject thread from RadioTalk. *Radio Image shown in-place of Infrared Image*

Research Problems

Problem: Volunteers could generate new tags freely → lack of tag coherence

Question: Can we identify the relationships between tags by learning embeddings?

Problem: Assigning tags was optional → tagging could be incomplete

Question: Can we use the subject comments to perform tag recommendation?

Overall Goal: Maximise Science Output from Radio Galaxy Zoo Project

Discovery of Missing Subjects

- Discovered **10,810** new subjects that were present in forum but not in the catalogue
- It would take an astronomer **8.7yrs** at **40hrs/week** to label this many subjects!
- Overall these subjects were complex, extended and difficult to classify

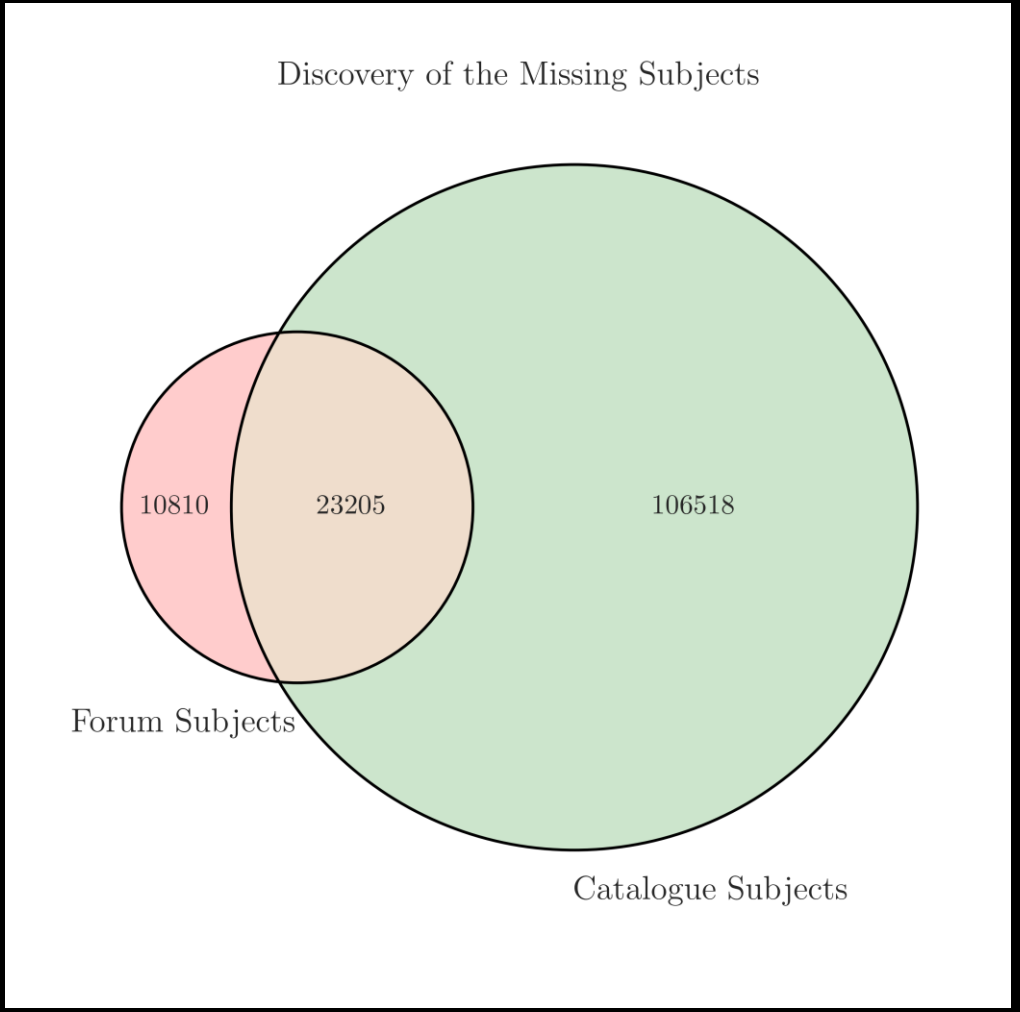
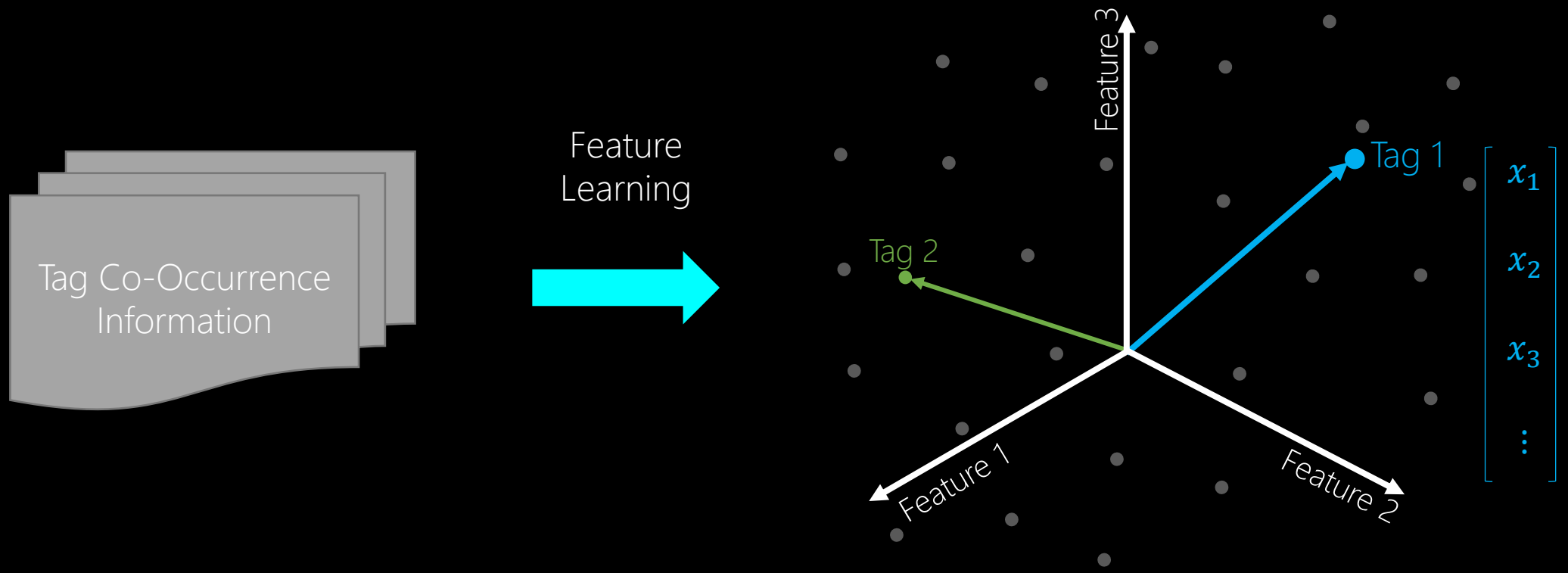


Figure 7: Discovery of Missing Subjects

Research Aim – Learn Tag Embeddings

Can we learn Tag Embeddings from Co-Occurrence Information?

- Are these embeddings interpretable?
- Do these embeddings uncover relationships or form clusters?



Method – Learn Tag Embeddings

- **Construct Subject-Tag Co-Occurrence Matrix X**
  - $X_{i,j}$ denotes how many times the tag j appeared in the comments for subject i.
- **Perform Non-Negative Matrix Factorisation (NMF)**
  - Technique that has demonstrated success in generating interpretable embeddings
- **Probe at Embeddings**
  - Explore Feature Axes
  - Use t-SNE to visualise relationships
  - Use k-means to cluster tags

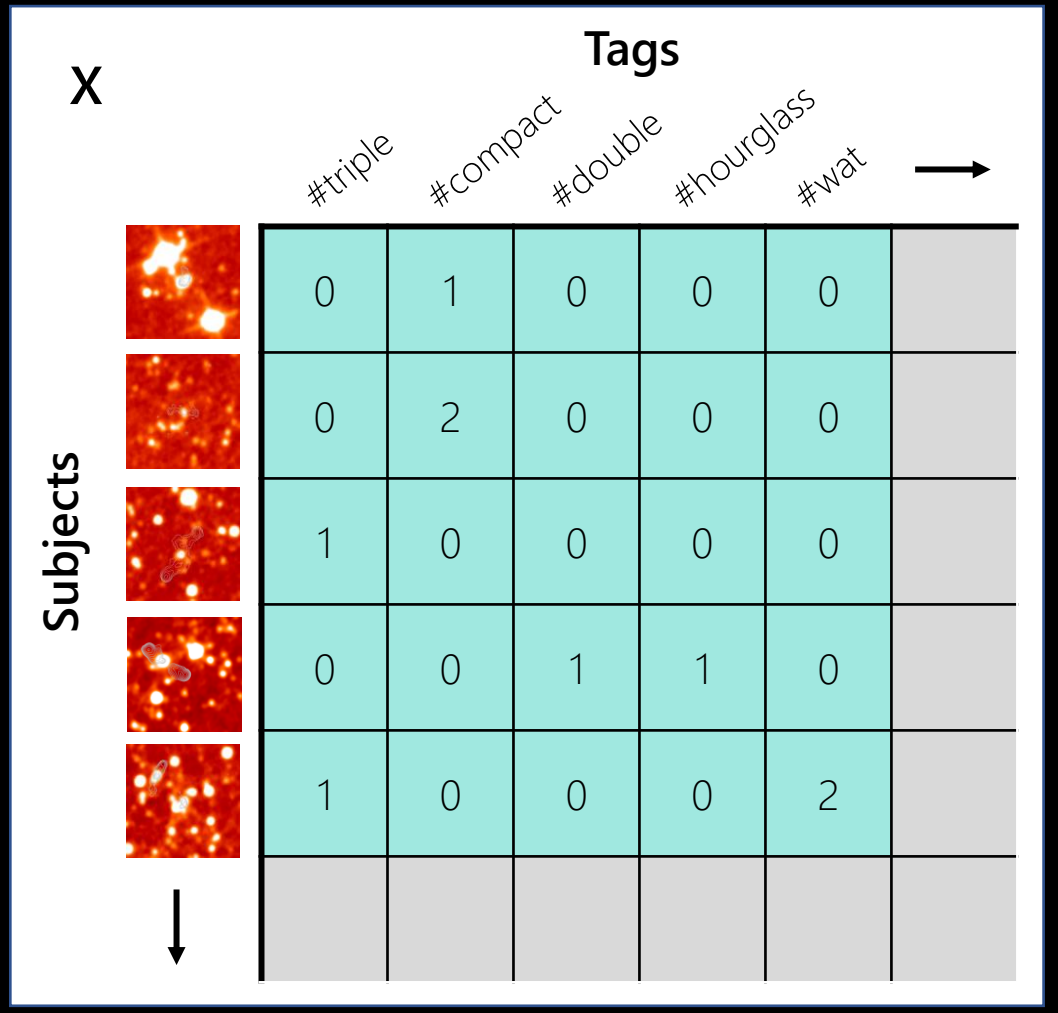


Figure 8: Co-Occurrence Matrix **X**

Non-Negative Matrix Factorisation (NMF)

- Decomposes a Non-Negative Matrix into the product of two non-negative matrices **W** and **H**
- Non-Negative constraint helps interpretability
- Rows of **W** are subject embeddings
- Rows of **H** are tag embeddings
- Chose k=53 after accepting a reconstruction error of 20%

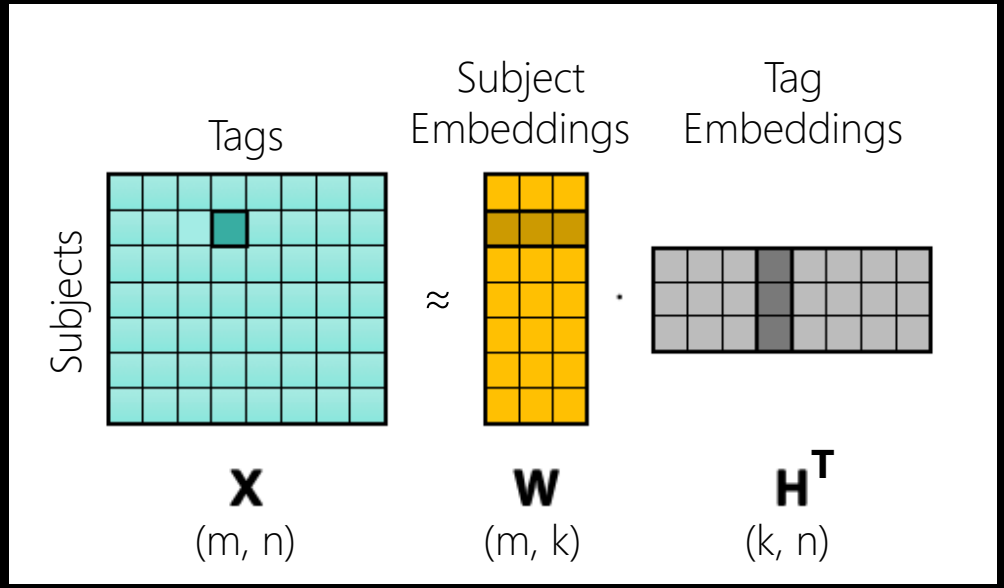


Figure 9: Non-Negative Matrix Factorisation

Results – Tag Embeddings

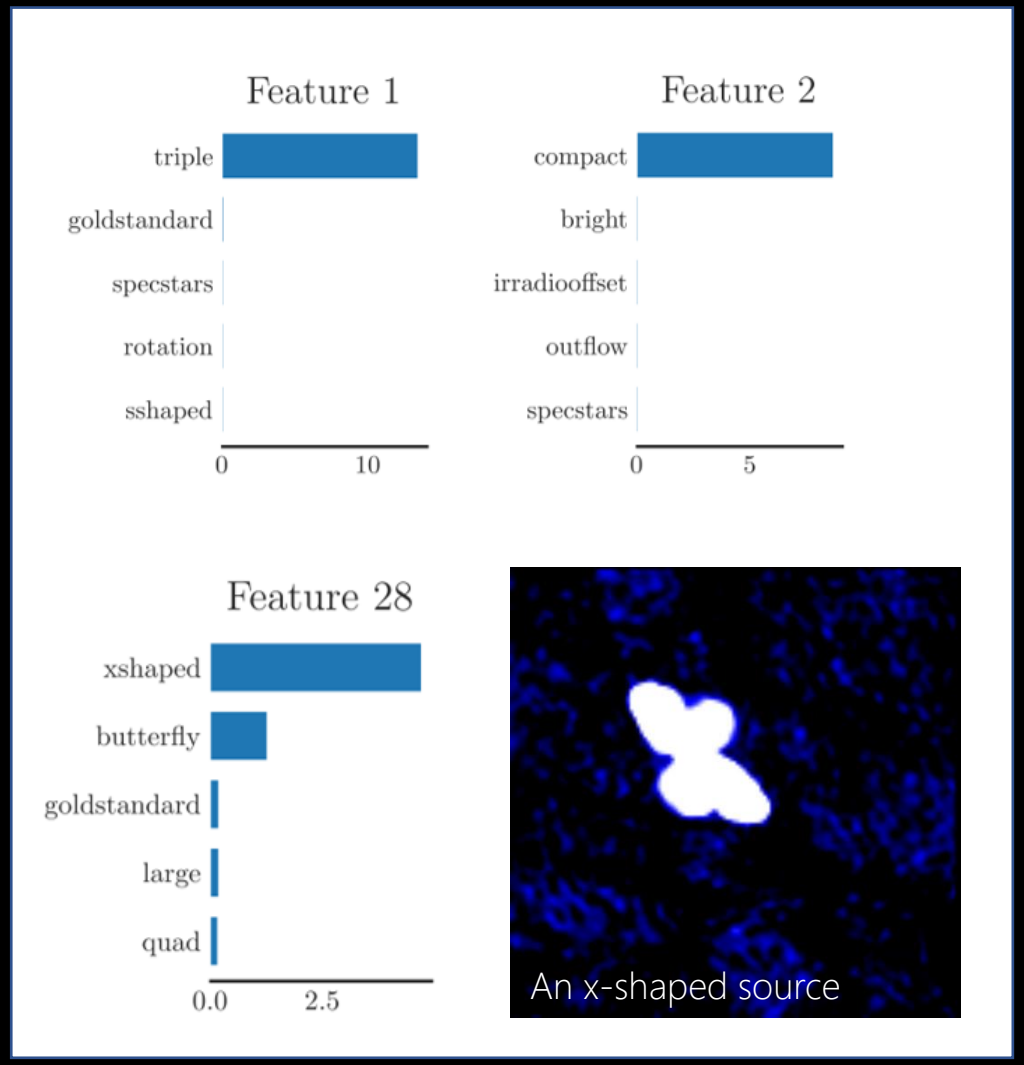


Figure 10: Sample of Top Tags along Feature Axes

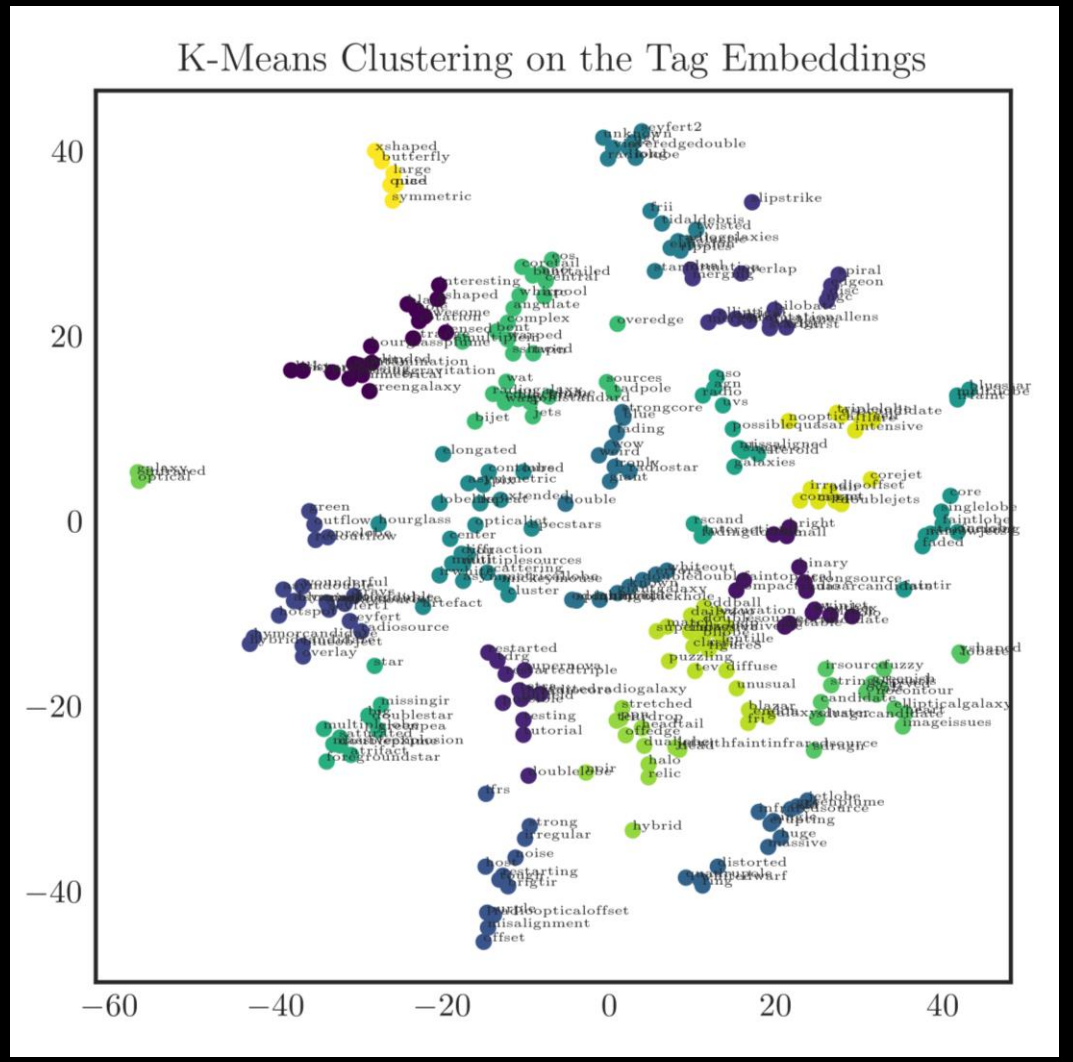
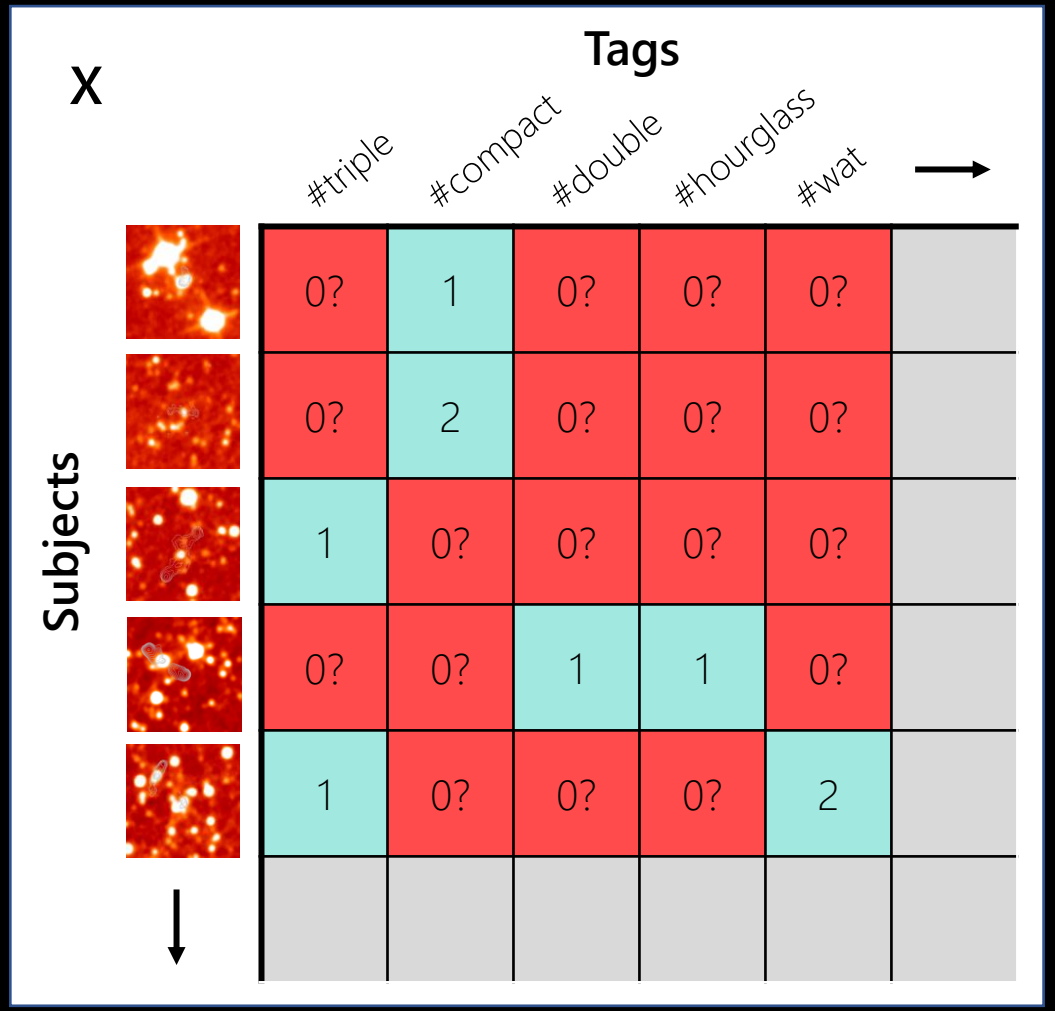


Figure 11: K-Means and t-SNE visualisation of tag embeddings

Research Challenges

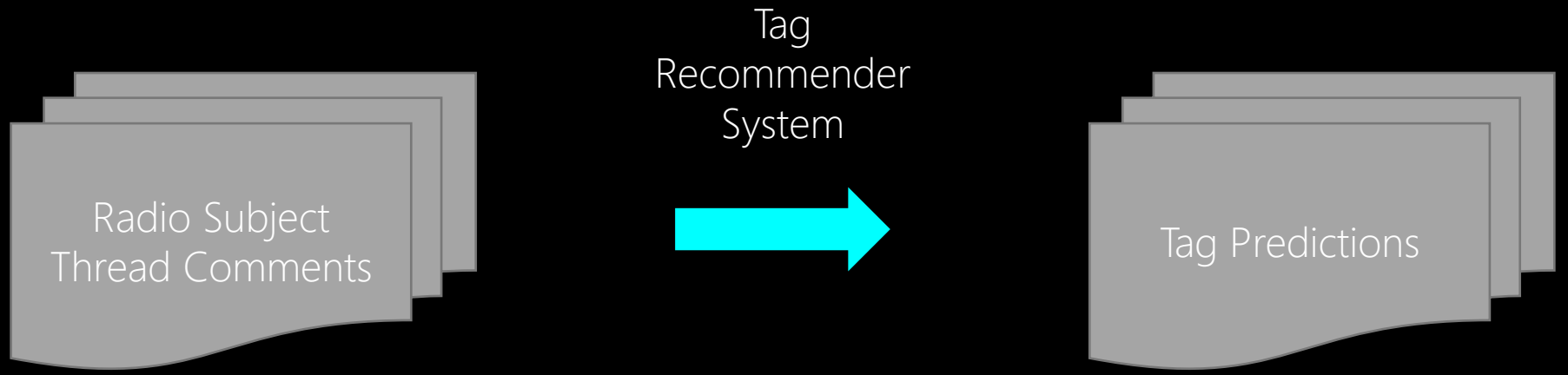
- **Positive-Unlabelled Data (PU-Data)**
  - The dataset contains only positive labels. The absence of a tag in the RadioTalk forum does not imply it is not applicable.
- **Lack of Data**
  - The co-occurrence matrix is highly sparse. ~99.4% of entries are zero.



Research Aim – Tag Recommendation

How well can we predict the tags of a radio subject given user text descriptions?

- Can we leverage a pre-trained language model?
- Can we recommend tags to subjects that previously had none?



Methods – Tag Recommendation Pipeline

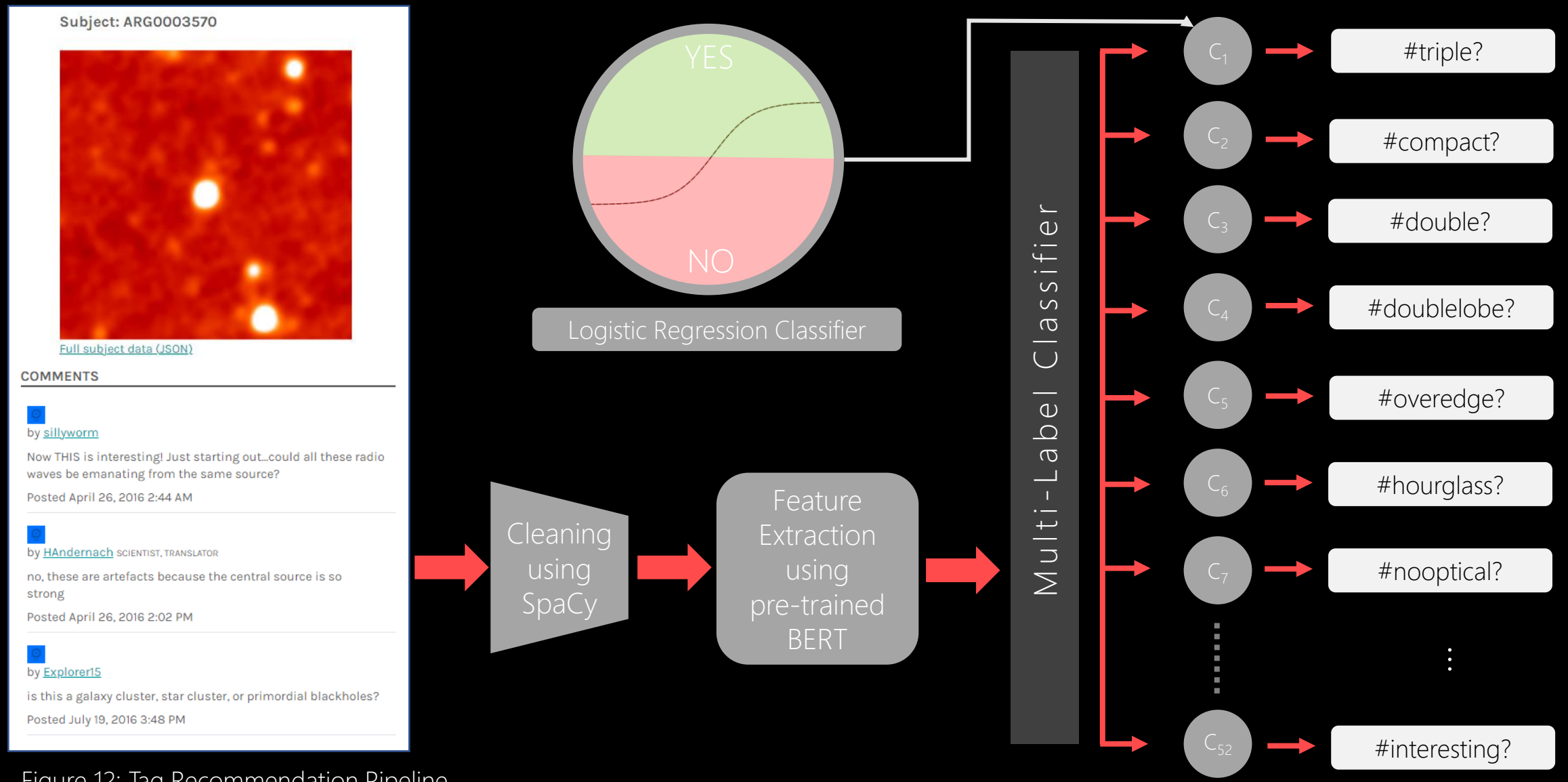


Figure 12: Tag Recommendation Pipeline

BERT (Bidirectional Encoder Representations from Transformers)

- State-of-the-art language model based on bi-directional transformer architecture.
- Many variants, we used BERT-base-uncased (110M parameters)
- Model is pre-trained on over 3 billion words from the English Wikipedia and BooksCorpus datasets.

Methods – Feature Extraction using BERT

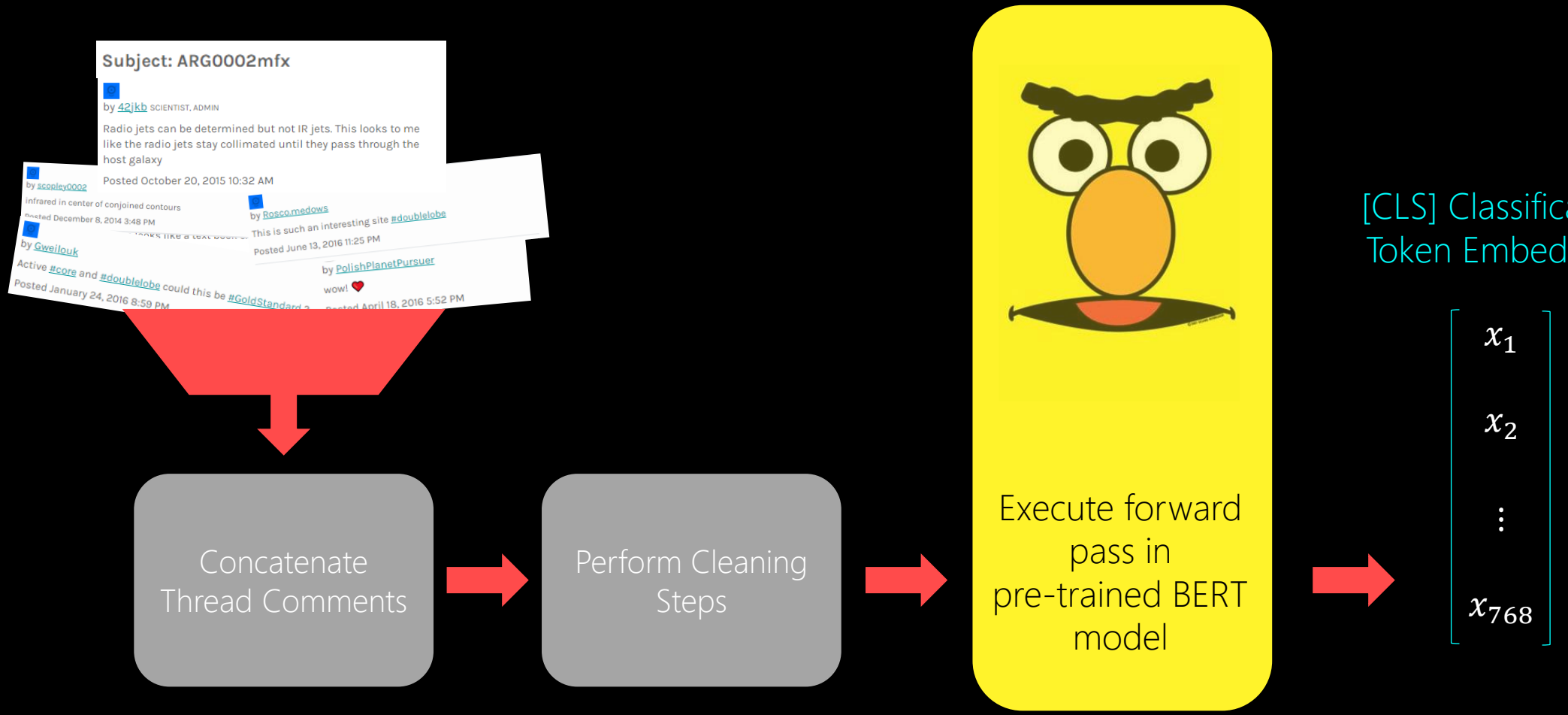
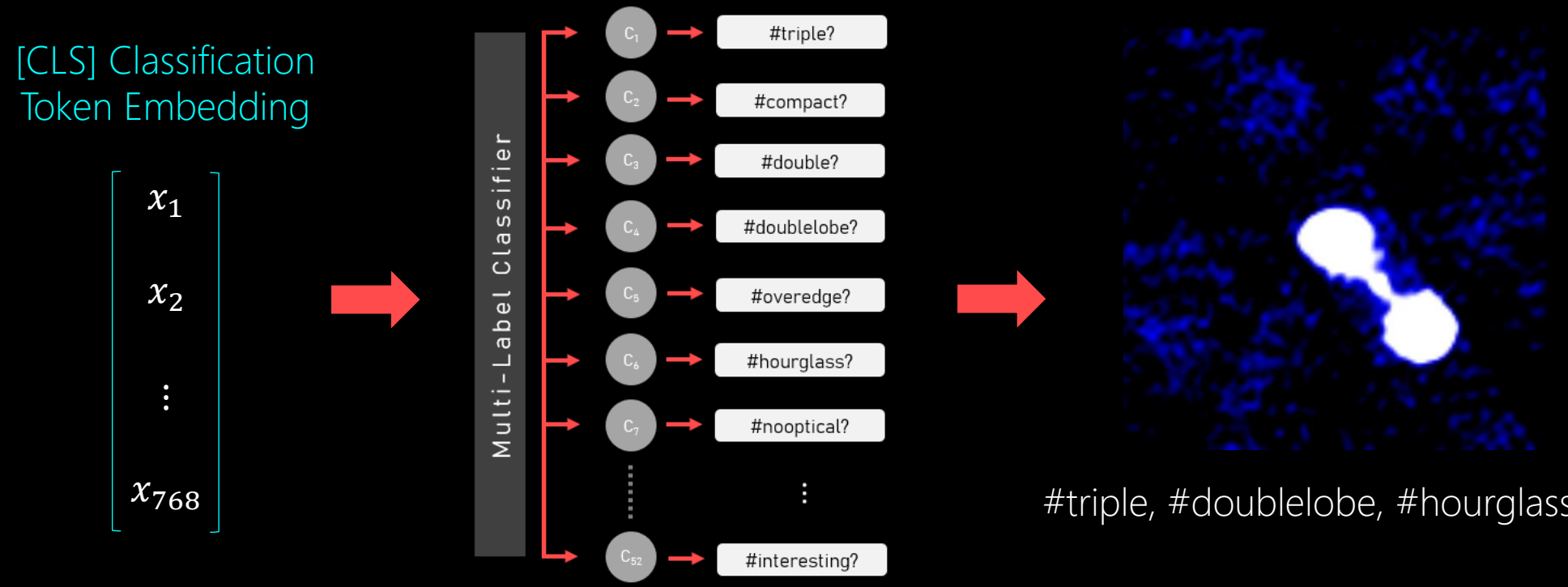


Figure 13: Feature Extraction Pipeline

Multi-Label Classification

- Multi-Label Classification implies more than one label can be assigned to each instance
- **Binary Relevance**
  - Train an independent Binary Classifier for each label to be predicted
  - Common approach to multi-label classification
  - Cannot capture dependencies between labels



Results – Multi-Label Classification

Table 1: 5-Fold Nested Cross-Validation (Macro-Averaged)

Precision	Recall	F1-Score
0.64 ± 0.02	0.66 ± 0.01	0.65 ± 0.01

Table 2: Best and Worst Performing Tags

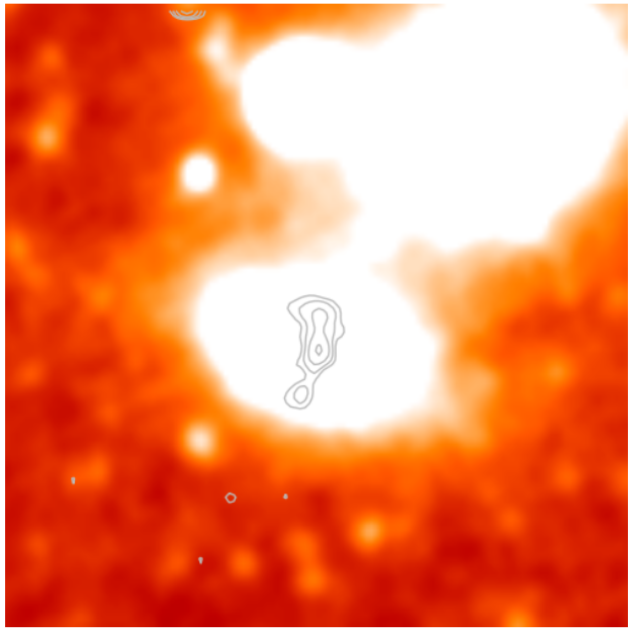
Tag	Precision	Recall	F1-Score
nooptical	0.92 ± 0.01	0.95 ± 0.01	0.93 ± 0.01
hourglass	0.88 ± 0.02	0.92 ± 0.01	0.90 ± 0.01
compact	0.88 ± 0.02	0.92 ± 0.02	0.90 ± 0.00
…	…	…	…
star	0.38 ± 0.06	0.47 ± 0.09	0.41 ± 0.05
spiral	0.40 ± 0.08	0.37 ± 0.10	0.37 ± 0.05
qso	0.33 ± 0.06	0.34 ± 0.05	0.33 ± 0.04

Astrotag Explorer

Astrotag Explorer

Interactively view the tag recommmendation results.

Subject Image



Thread Text

tiny #hourglass more in NVSS
NGC 3653 = 2MASX J11223008+2416457 = SDSS J112230.06+241645.3
FIRST emission perpendicular to central dust lane
NVSS-size 5.2' or 175 kpc

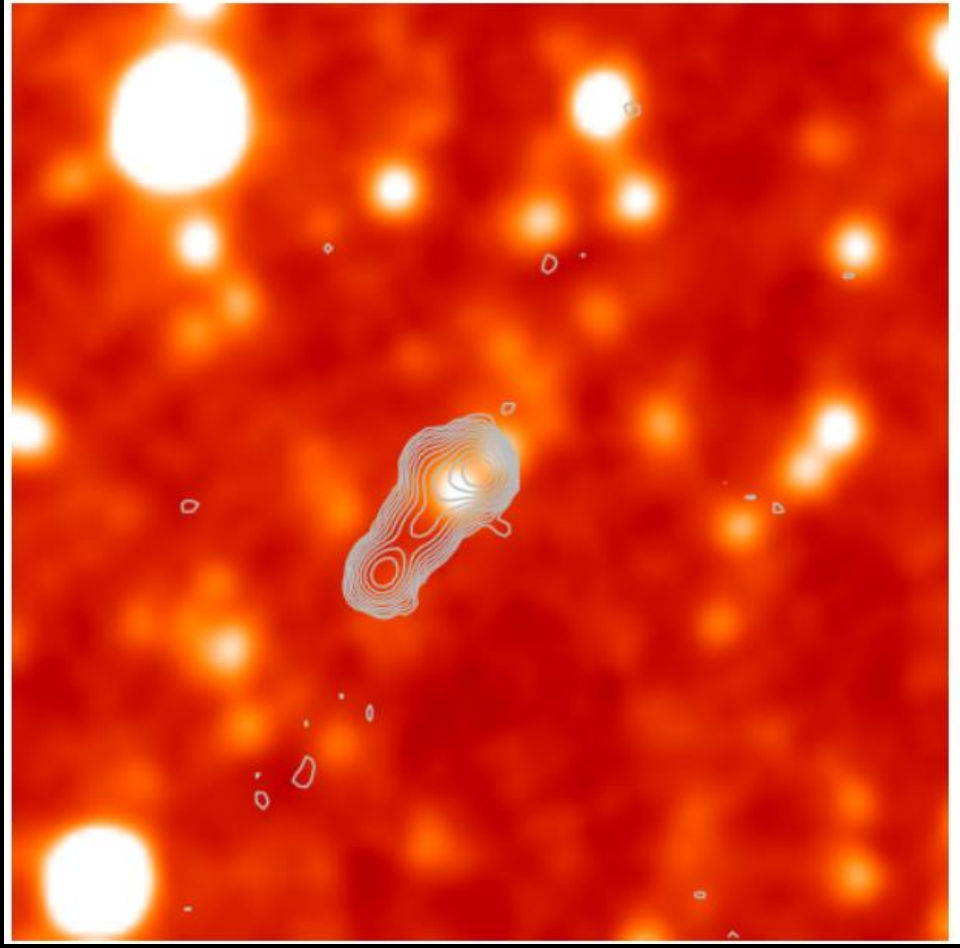
Search..

SubjectID	Tags	BERT_Predicted_Tags
ARG0003lw6	bent	bent (0.9966)
ARG0001rrt	double	double (0.9996)
ARG0001oaq	double	double (0.9996)
ARG0001i5y	doublelobe,compact,infrared,galaxy	infrared (0.9985), compact (0.9558), galaxy (0.8366), doublelobe ...
ARG0000dg0	bent,doublelobe,hybrid	doublelobe (0.9872), green (0.9719), bent (0.8221)
ARG00022x5	hourglass	hourglass (0.8687)
ARG0001nj2	compact	compact (0.9999)
ARG0000t2l	hourglass	hourglass (1.0000)
ARG00033sq	hourglass,star	hourglass (0.9989), star (0.8842)
ARG0003oyn	doublelobe	doublelobe (1.0000)
ARG0000j0k	compact	compact (0.9987)
ARG0001fc2	relic	diffuse (0.9698), relic (0.9461)
ARG0002wlw	triple,overedge	
ARG0001fat	hourglass,green	hourglass (0.9995)
ARG0003jfx	overedge,corejet,doublelobe	overedge (0.8912), corejet (0.7443)
ARG0001op1	compact	compact (0.8995)
ARG000205q	bent,asymmetric	asymmetric (0.8614)
ARG0000xrt	double	double (0.7426)
ARG0000dg6	hourglass	hourglass (0.9995), nat (0.4312)
ARG0003jze	noir,overedge,faintir,ifrs,compact,hourglass,nooptical	noir (1.0000), nooptical (1.0000), compact (0.9998), ifrs (0.8644)
ARG0002yxp	twin	twin (0.9995)
ARG00038ac	overedge,double	compact (0.9553), triple (0.7207), double (0.4946)
ARG00017cd	compact	compact (0.9986)
ARG0000vy8	double	double (0.9505)
ARG0001spg	green,double	green (0.9977)
ARG0000av8	doublelobe	doublelobe (0.9963)
ARG0001l4f	ifrs,compact,nooptical	compact (0.9885), ifrs (0.9867), nooptical (0.9326)
ARG0002egj	compact	
ARG00004i9	triple,faintir,compact	faintir (0.9345), compact (0.9194), triple (0.6544)
ARG0000gxk	triple,radio	radio (0.8524)
ARG00006jl	doublelobe,bhgroup,compact	bhgroup (1.0000), doublelobe (0.8151), double (0.5020)
ARG0003ceq	wat,hourglass	hourglass (0.7674), overedge (0.6547)

Figure 14: Astrotag Explorer Web Tool

Case 1: Tag Appears Verbatim

Subject Image



Thread Text

highly asym #triple host merger zph 0.60 J145550.10+350947.0
1237662306192982490

Volunteer Assigned Tags:

#triple

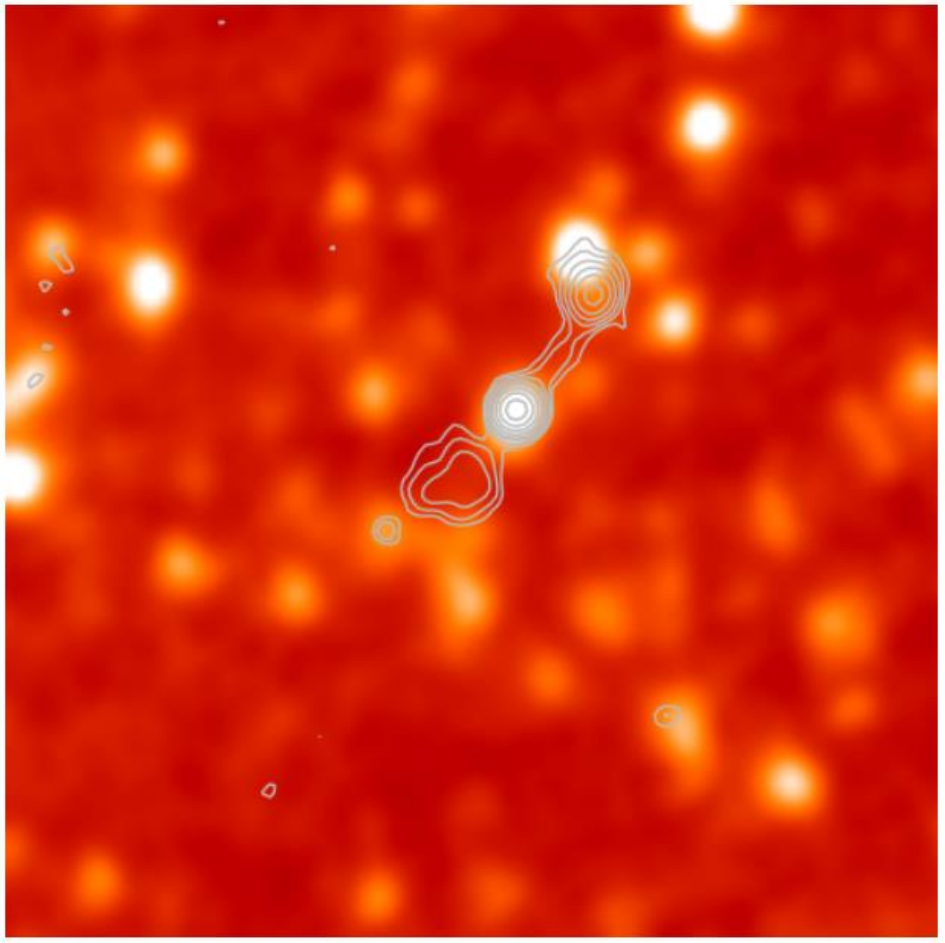
BERT Classifier Predicted Tags:

#asymmetric (0.9724)
#triple (0.9017)
#merger (0.6133)

Figure 15: Example of good recommendation

Case 1: Tag Appears Verbatim

Subject Image



Thread Text

Hybrid ? Host SDSS J154813.32+241149.2
Same as ARG000233k Don't think it is a hybrid, though More like assym triple

Volunteer Assigned Tags:

None

BERT Classifier Predicted Tags:

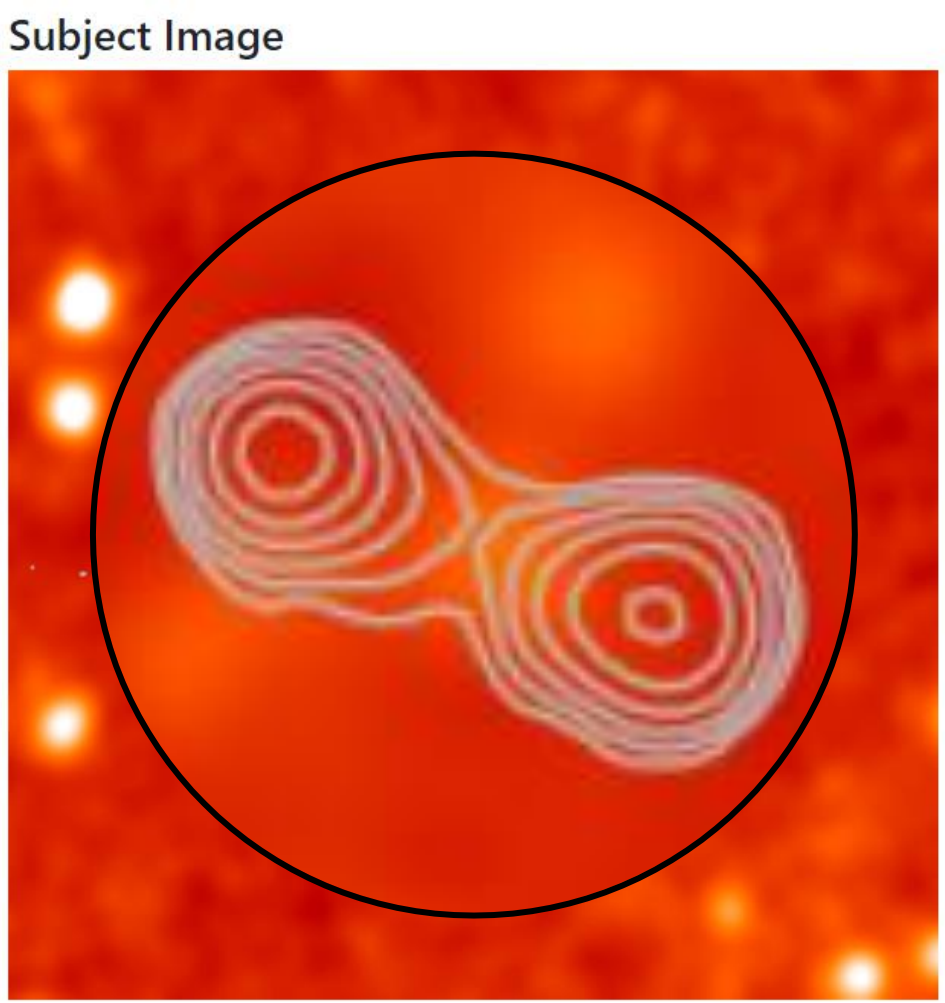
#hybrid (1.0000)
#hymor (0.7483)

Missed Tags:

#asymmetrical, #triple

Figure 16: Example of poor recommendation

Case 2: Descriptions Only



Thread Text

Blank pixels in FIRST Mobius-loop shape...

Volunteer Assigned Tags:

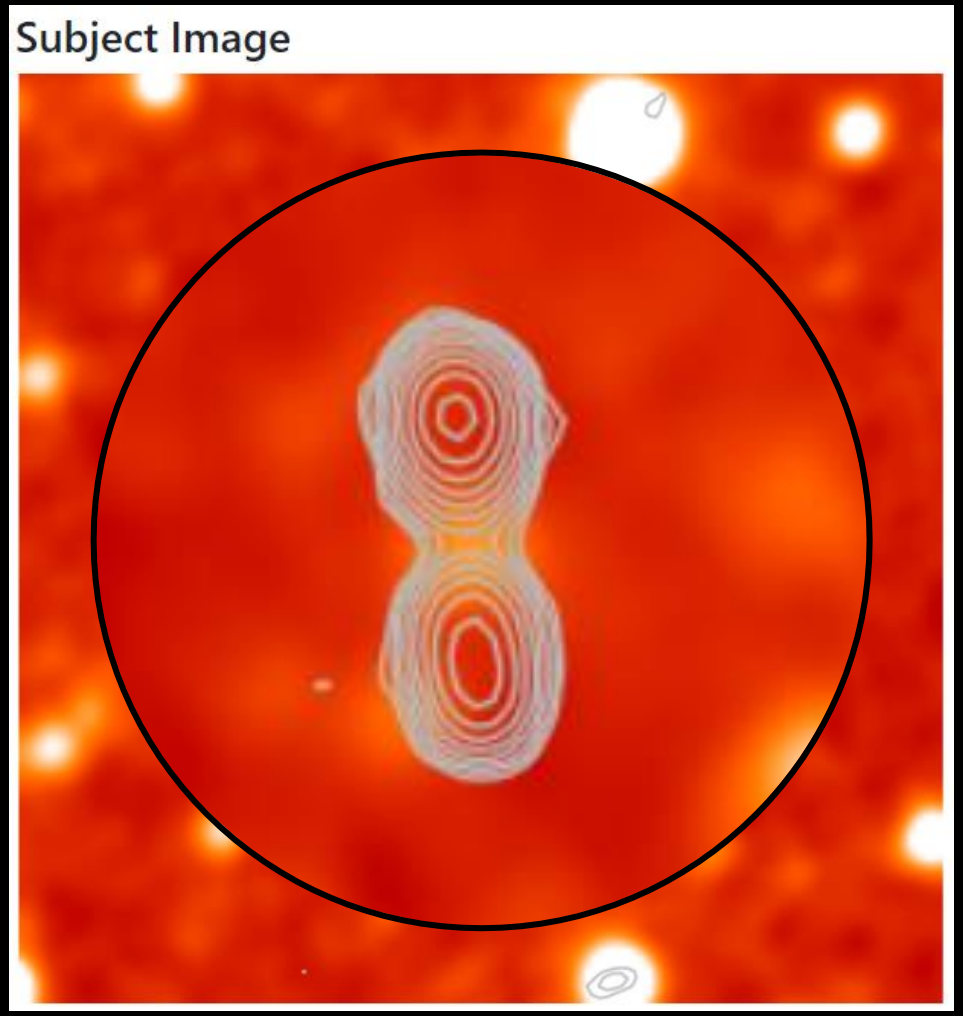
None

BERT Classifier Predicted Tags:

#bijet (0.8363)
#hourglass (0.5581)

Figure 17: Example of good recommendation

Case 2: Descriptions Only



Thread Text

Compact radio source, strong infrared signal. And a typical jet ejected from both sides, faint infrared sources.

Volunteer Assigned Tags:

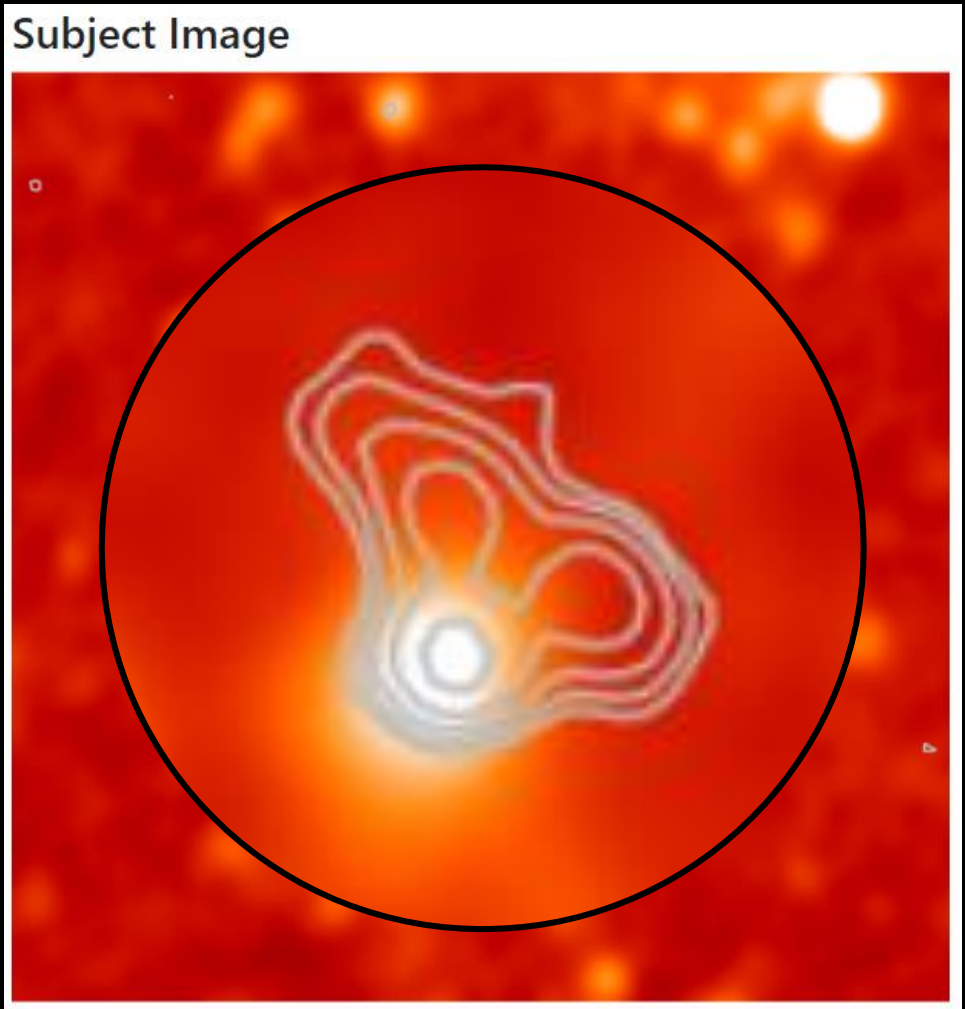
None

BERT Classifier Predicted Tags:

#hourglass (0.7725)

Figure 18: Example of good recommendation

Case 2: Descriptions Only



Thread Text

A little #mouse looks at us from the sky
This one is great! I have no idea what the radio source is, looks very odd.
The inner contour at the center looks like mickey mouse :-)

Volunteer Assigned Tags:

#mouse (Not included in our recommendation system)

BERT Classifier Predicted Tags:

None

Figure 19: Example of poor recommendation

Classifier Improvements

Before Astrotag : 4892 subjects without tag
After Astrotag : 2816 subjects without tag

Problem:
Text descriptions are insufficient

- URLs
- Astronomical Co-Ordinates
- Non-English Text
- Not descriptive enough
- No text at all (Cold Start Problem)

Improvement:
Include additional features from the image and/or catalogue data

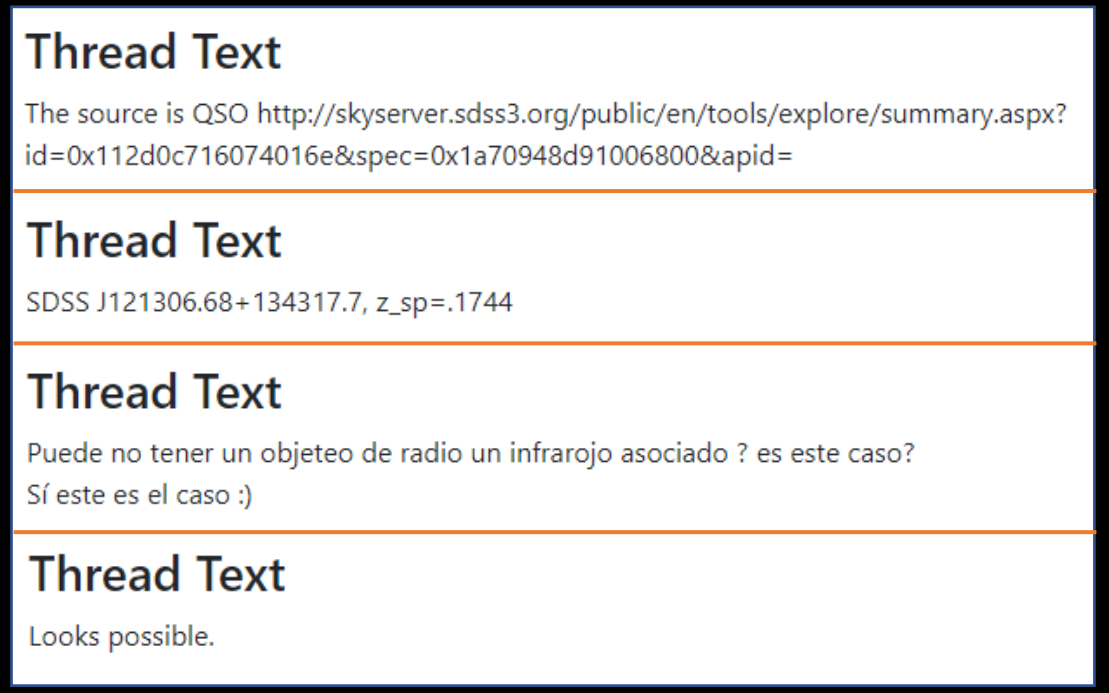


Figure 20: Sample of poor input

Research Conclusions

Problem: Volunteers could generate new tags freely → Lack of tag coherence

Question: Can we identify the relationships between tags by learning embeddings?

Our Findings: There is a lack of co-occurrence data to learn from. However our results may assist a manual approach to increase tag coherence.

Problem: Assigning tags was optional → tagging could be incomplete

Question: Can we use the subject comments to perform tag recommendation?

Our Findings: Our classifier demonstrates the potential to utilise the text data but a hybrid approach utilising additional features may perform better.

Any Questions?