UNIVERSITY OF OXFORD

The Alan Turing Institute

MANCHESTER 1824 The University of Manchester

Semantic taxonomies and how to find them

1st Large Language Models in Physics Symposium

Micah Bowles

github.com/mb010/Text2Tag

My Research

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. Q&A

Research interests:

- Big data
- Representations
- Machine learning

My work:

- MIGHTEE-POL (deep radio survey with polarisation measurements)
- Large scale pre-training
- **Semantic language**

Introduction

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. Q&A

Goals:

1. Get the most out of our data,
2. Enable research into high dimensional features of data,
3. Allow users to interact with data intuitively,
4. Reduce barriers to the field (i.e. reduce labelling costs).

"A picture is worth a thousand words."

… what about our data?

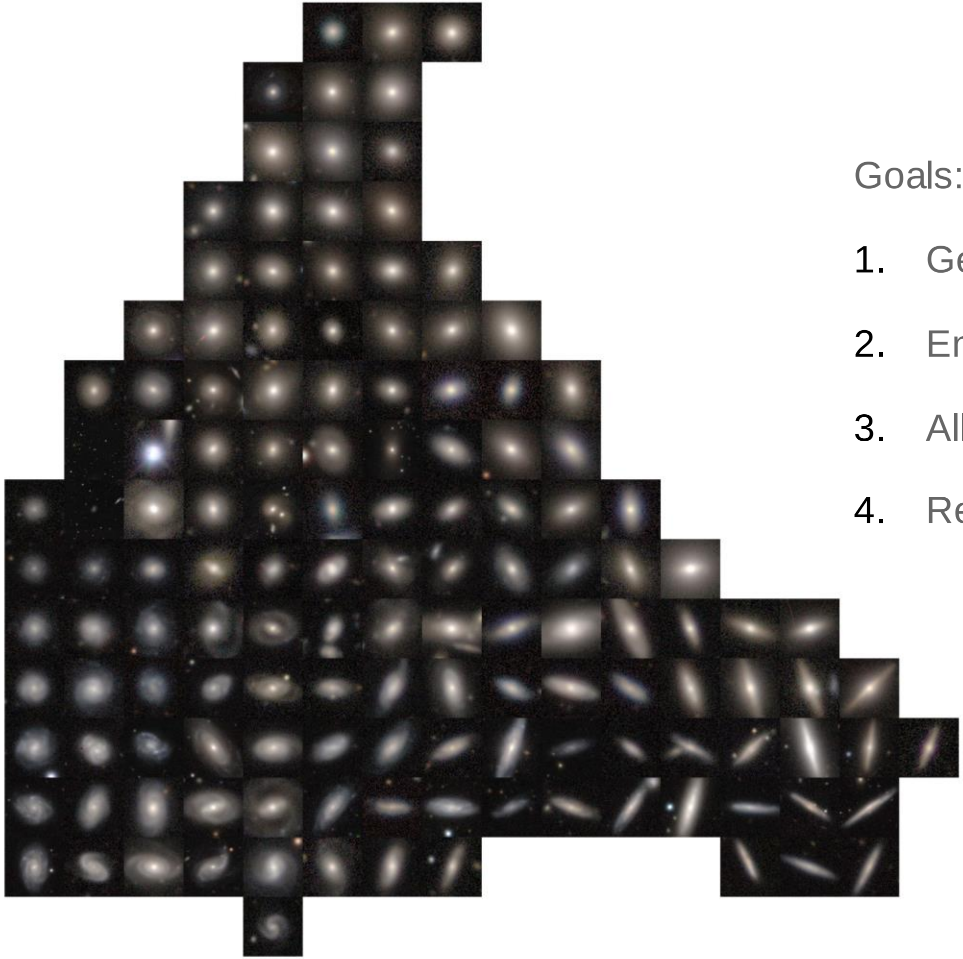


Figure 1. Visualisation of the representation learned by our CNN, showing similar galaxies occupying similar regions of feature space. Created using Incremental PCA and umap to compress the representation to 2D, and then placing galaxy thumbnails at the 2D location of the corresponding galaxy.

Walmsley et al. (2022) https://arxiv.org/abs/2110.12735

Goals:

1. Get the most out of our data,
2. Enable research into high dimensional features of data,
3. Allow users to interact with data intuitively,
4. Reduce barriers to the field (i.e. reduce labelling costs).

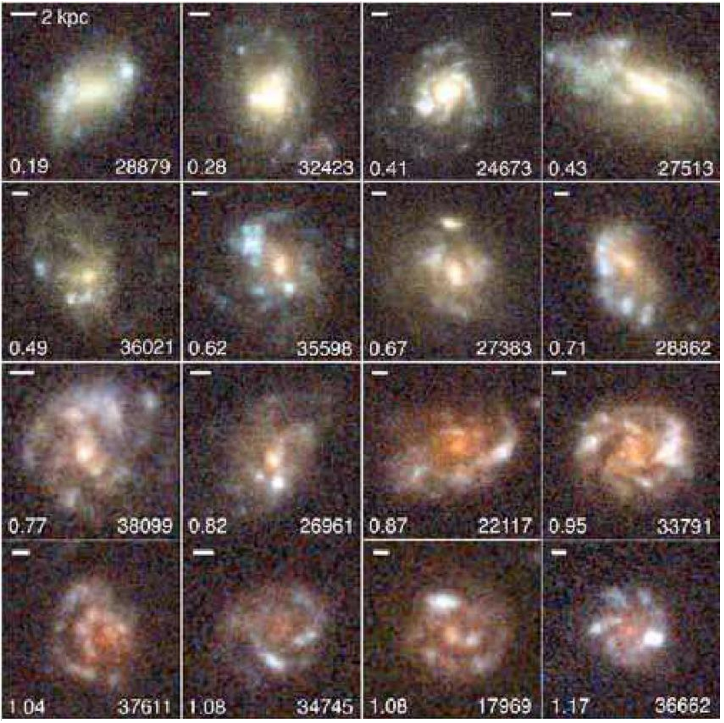
"A picture is worth a thousand words."

… what about our data?

Introduction

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. Q&A

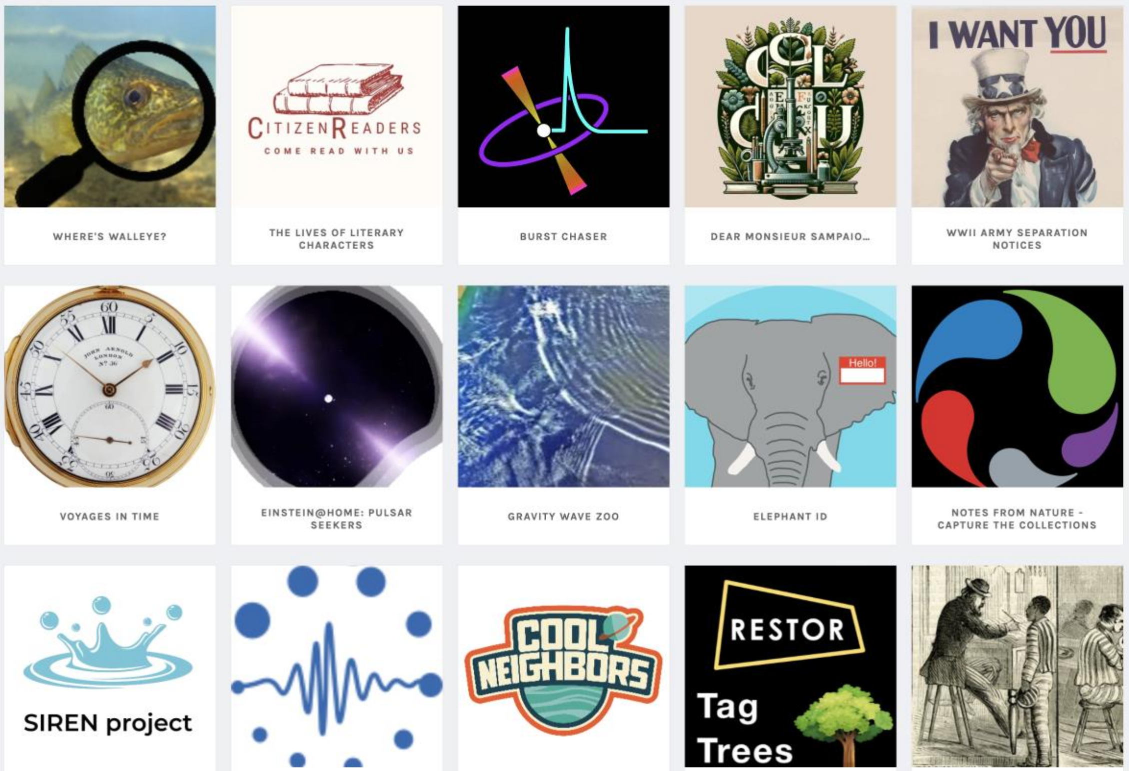
ESA/Hubble NGC 4298



Elmegreen et al. 2009 10.1088/0004-637X/701/1/306.

Introduction

Introduction to Citizen Science.



1. **Introduction**
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. Q&A

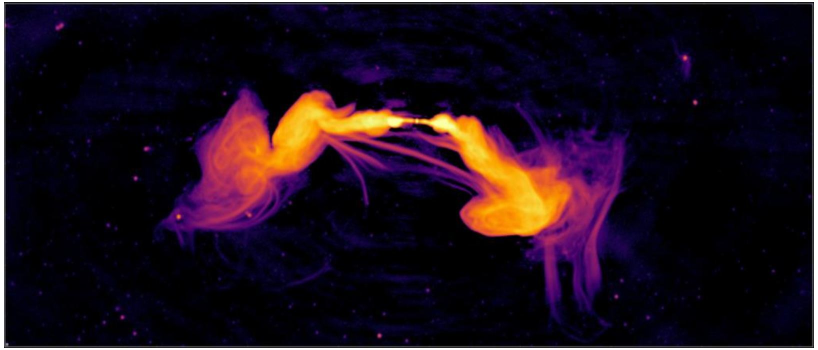
Applications

1. **Introduction**
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. Q&A

Radio galaxy morphology

Initial test bed.

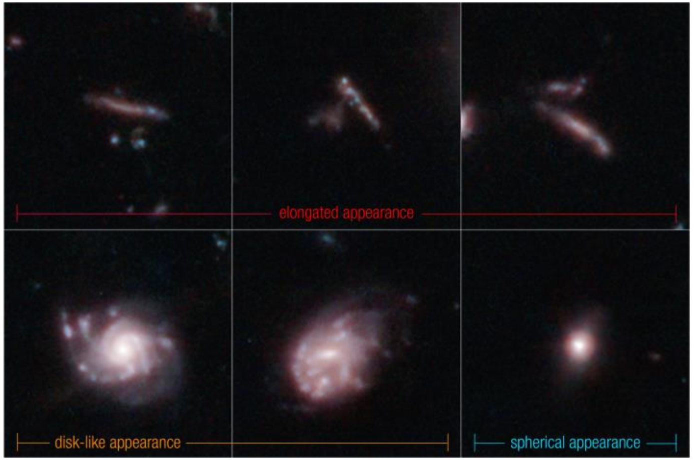
Bowles+22,23



Ramatsoku et al. (2020); South African Radio Astronomy Observatory (SARAO)

JWST galaxy morphology

Hayley Roberts (Minnesota)



Pandya et al. (2024) https://arxiv.org/abs/2310.15232

NASA, ESA, CSA, STScI, Steve Finkelstein (UT Austin), Micaela Bagley (UT Austin), Rebecca Larson (UT Austin)

Applications

1. **Introduction**
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. Q&A

Variable Stars with
Planet Hunters TESS

Discussed with

Chris Lintott (Oxford) &
Nora Eisner (Flatiron)

(Likely to recommence later this year, after I submit my PhD thesis).

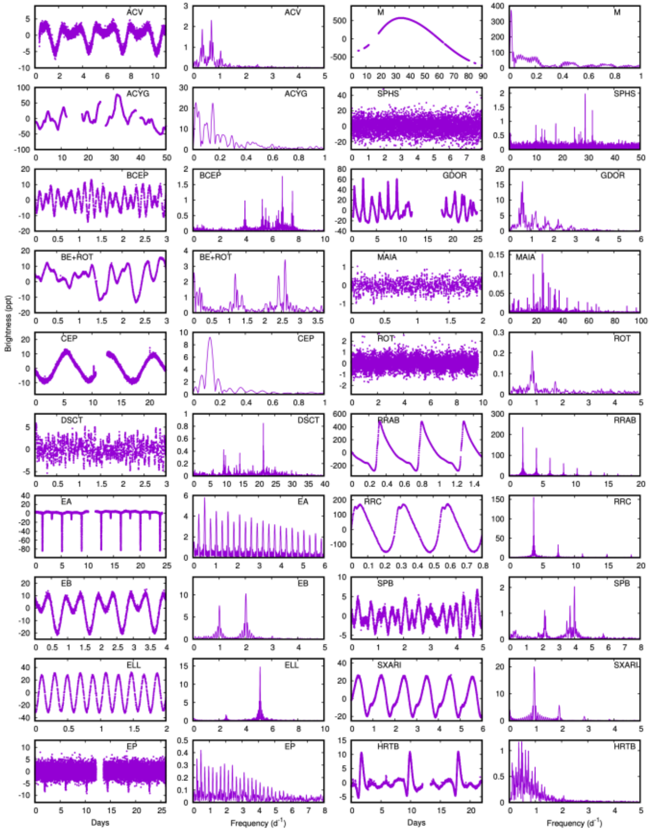


Figure 1. Examples of light curves and periodograms for different classes of variable stars observed by *TESS*. The brightness scale is parts per thousand.

Balona 2023 https://arxiv.org/abs/2212.10776

Pre-Training

1. Introduction
2. **Pre-Training**
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. Q&A

Data volumes to build robust representations for adaptable models:

- ImageNet - 1.4 million samples, 21k labels [1]
- JFT – 4 billion samples, 30k labels [2]

Domain specific data sets **cost** much more to **label**.

"A picture is worth a thousand words."

"Car"

Semantic Class Targets: A Novel Machine Learning Solution

1. Introduction
2. Pre-Training
3. **Semantics**
4. Early approach
5. Application
6. Conclusions
7. Q&A

Problem in technical classifications:

Car

vs

Fanaroff-Riley Type 1 (FR I)

The solution:

"A New Task: Deriving **Semantic Class Targets** for the Physical Sciences"

Bowles et al. (2022) and Bowles et al. (2023).

Ontology ('study of being'): “A set of concepts and categories in a subject area or domain that shows their properties and the relations between them.” (not the ontology of analytic philosophy)

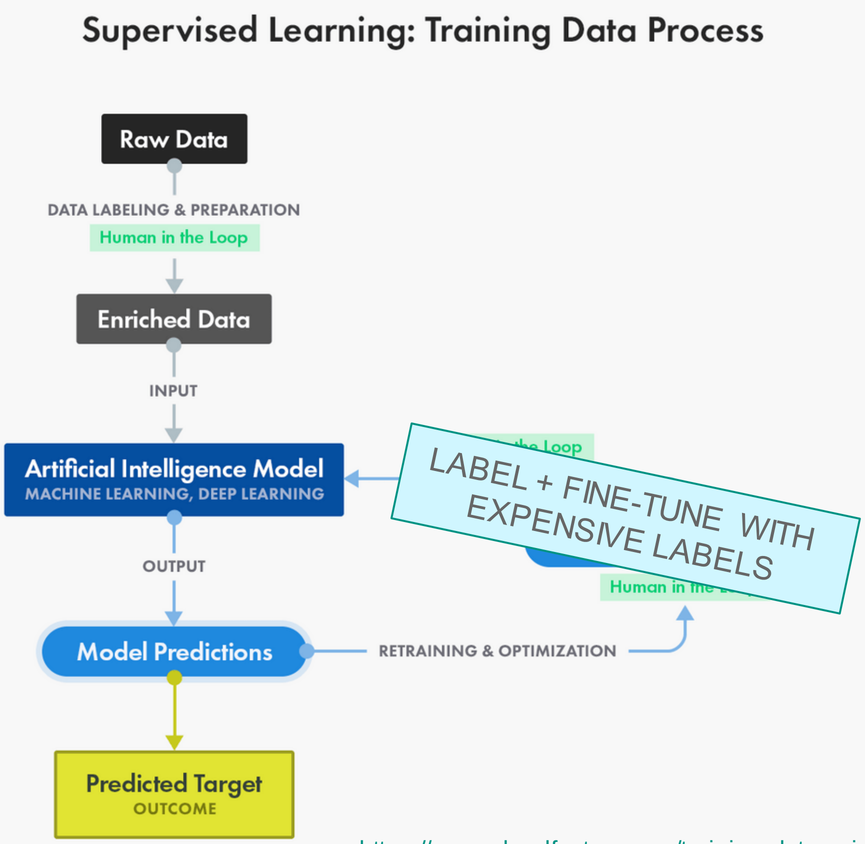
Examples:

- CS: Semantic web (Web Ontology Language) [5]
- BioMed: Basic formal ontology [6], etc.
- AI: WordNet [7] (not quite but pretty close) (ImageNet is built *using* WordNet)

Please note: Semantic taxonomies are not ontologies.

Semantic Classifications for Pre-Training

1. Introduction
2. Pre-Training
3. **Semantics**
4. Early approach
5. Application
6. Conclusions
7. Q&A



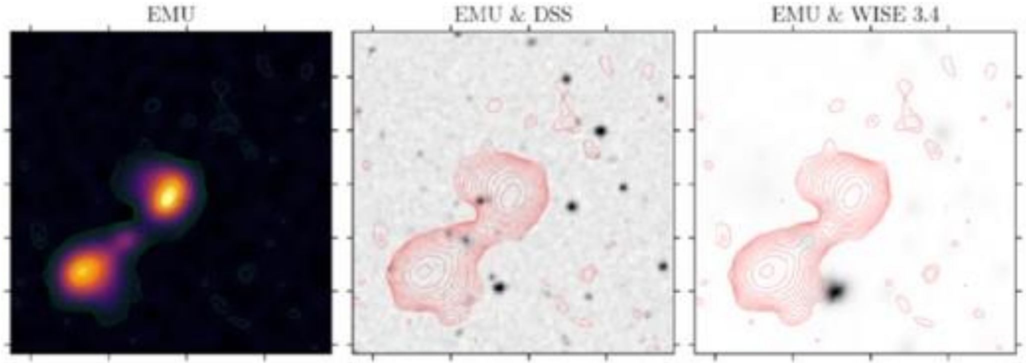
Supervised Learning: Training Data Process
Raw Data
DATA LABELING & PREPARATION
Human in the Loop
Enriched Data
INPUT
Artificial Intelligence Model
MACHINE LEARNING, DEEP LEARNING
OUTPUT
Model Predictions
RETRAINING & OPTIMIZATION
Human in the L
the Loop
Predicted Target
OUTCOME

LABEL + FINE-TUNE WITH EXPENSIVE LABELS

https://www.cloudfactory.com/training-data-guide

The First Approach

1. Introduction
2. Pre-Training
3. Semantics
4. **Early approach**
5. Application
6. Conclusions
7. Q&A

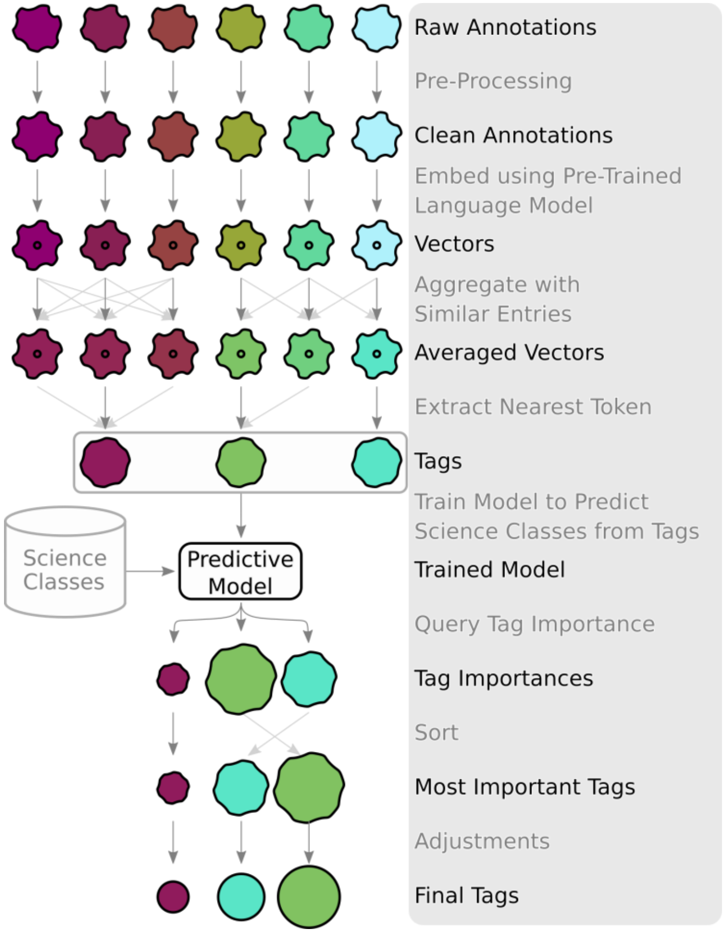


RGZ EMU [8]

The First Approach

1. Introduction
2. Pre-Training
3. Semantics
4. **Early approach**
5. Application
6. Conclusions
7. Q&A

[8]

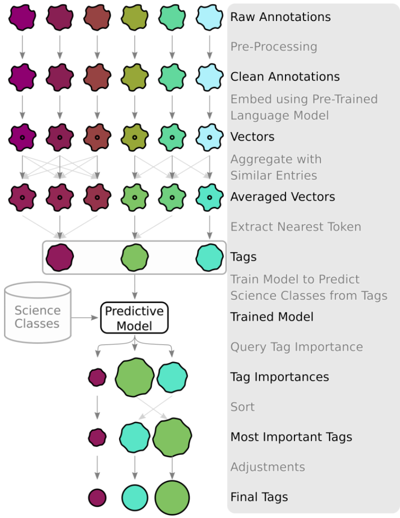


The First Approach

1. Introduction
2. Pre-Training
3. Semantics
4. **Early approach**
5. Application
6. Conclusions
7. Q&A

Goal: From annotations of objects to the most scientifically useful (semantic) plain English labels for that data. [3,8]

1. **Encode** annotations.
2. **Aggregate** with cosine similarity selection.
3. **Extract** nearest token in data space.
4. **Train** random forest to predict science classes.
5. Find most **informative** tokens using Shapley values.
6. **Adjust** grammar of tokens if needed (small data).



Raw Annotations

Pre-Processing

Clean Annotations

Embed using Pre-Trained Language Model

Vectors

Aggregate with Similar Entries

Averaged Vectors

Extract Nearest Token

Tags

Train Model to Predict Science Classes from Tags

Trained Model

Query Tag Importance

Tag Importances

Sort

Most Important Tags

Adjustments

Final Tags

Congratulations! These are now **plain English semantic class targets!**

Semantic Classifications through LLMs

1. Introduction
2. Pre-Training
3. Semantics
4. **Early approach**
5. Application
6. Conclusions
7. Q&A

Ideally done without the special data format.

Unclear how this would work.

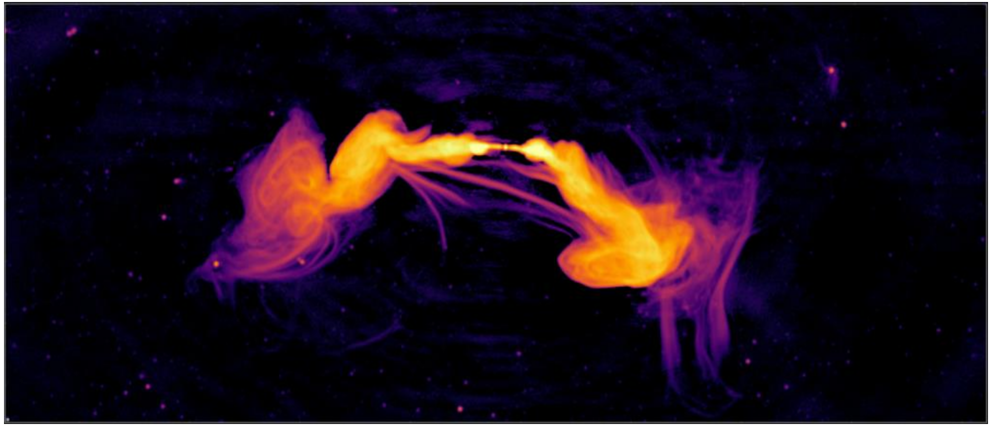
Maybe through concept pruning / debiasing.

Could of course be used as the encoder, however:

- this work was started before GPT3 was released,
- and there was more explainable tooling around to convince my colleagues that this works.

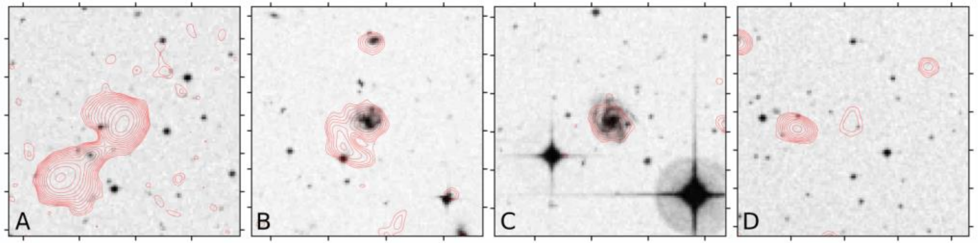
Radio Galaxy Morphology

1. Introduction
2. Pre-Training
3. Semantics
4. **Early approach**
5. Application
6. Conclusions
7. Q&A



[4] South African Radio Astronomy Observatory (SARAO)

Double | Star Forming Galaxy (SFG) | Star Forming Galaxy (SFG) | Single



A: diffuse, double, edge brightened, extended, faint, host, hourglass, jet, lobes, peak

B: bridge, extended, host, merger, traces host galaxy

C: amorphous, compact, extended, host, traces host galaxy

D: amorphous, compact, core, faint, host, small

[3]

Classes vs. Semantic Taxonomy

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. **Application**
6. Conclusions
7. Q&A

Normal domain specific classes:

Single, Double, Classical double, Triple, Narrow-angle tail (NAT), Wide-angle tail (WAT), Bent tail, Fanaroff & Riley Class 1 (FR I), Fanaroff & Riley Class 2 (FR II), Fanaroff & Riley Class 0 (FR 0), Hybrid, X-shaped, S-shaped, C-shaped, Diffuse, Double-double (DDRG), Core-dominant, Core-jet, Compact Symmetric Object (CSO), 1-sided, Odd Radio Circle (ORC), Star-Forming Galaxy (SFG)

Our final semantic classes:

amorphous, asymmetric brightness, asymmetric structure, bent, bridge, compact, core, diffuse, double, edge brightened, extended, faint, host, hourglass, jet, lobe, merger, peak, plume, small, tail, traces host galaxy

[8]

Labelling in Practice

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. **Application**
6. Conclusions
7. Q&A

Hard to compute – easy for non-expert labelers:

amorphous, bent, bridge, core, hourglass, jet, lobe, merger, plume, tail

Computable – given our expected data products:

asymmetric brightness, asymmetric structure, compact, diffuse, double, edge brightened, extended, faint, host, peak, small, traces host galaxy

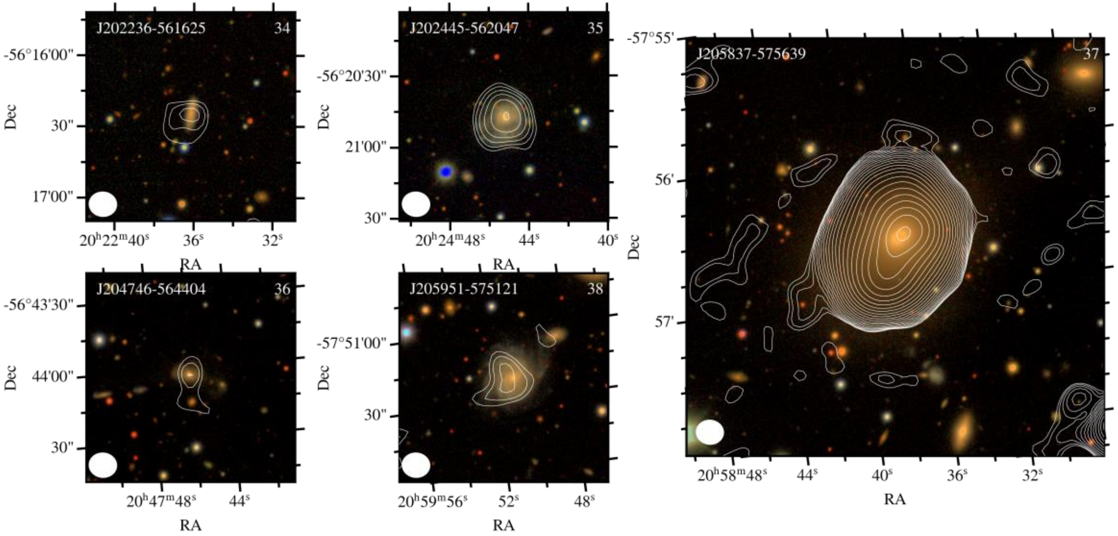
[8]

Science Case 1: Existing populations

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. **Application**
6. Conclusions
7. Q&A

Traces host galaxy

recovered 37/45 of the confidently classified "Star Forming Galaxies" in the sample. [8]



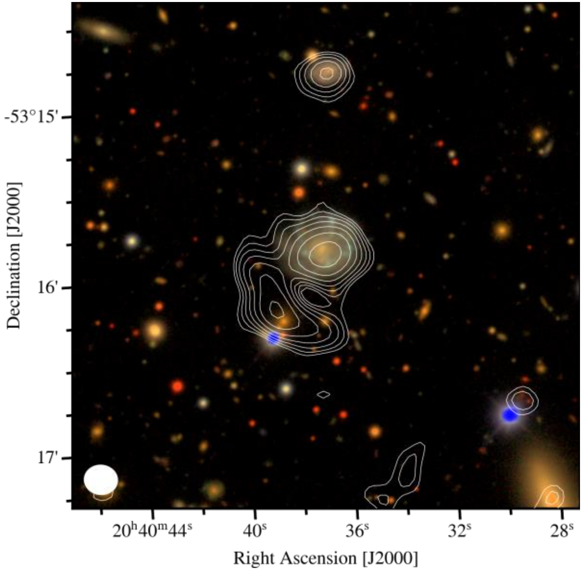
Science Case 2: Outlier detection for all!

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. **Application**
6. Conclusions
7. Q&A

Hourglass

but not tagged as

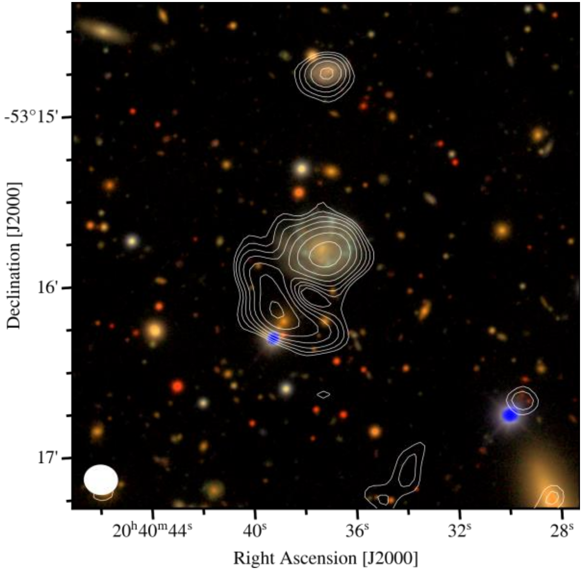
amorphous, **traces host galaxy** or **bent**. [8]



Science Case 3: New populations, New probes?

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. **Application**
6. Conclusions
7. Q&A

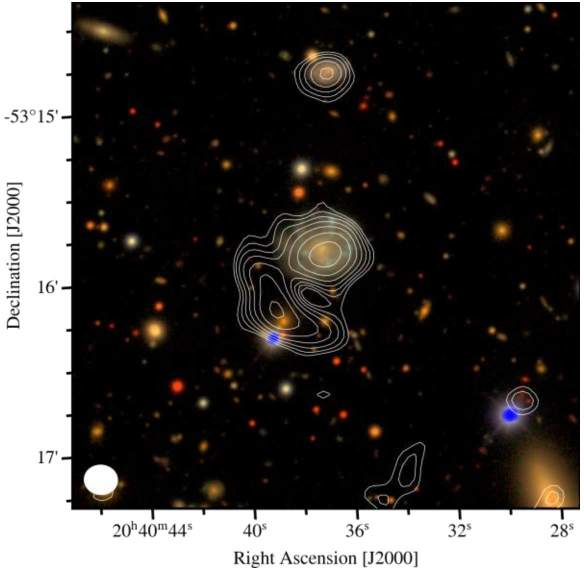
Hourglass
but not tagged as
amorphous, **traces host galaxy** or **bent**. [8]



Science Case 3: New populations, New probes?

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. **Application**
6. Conclusions
7. Q&A

Hourglass
but not tagged as
amorphous, **traces host galaxy** or **bent**. [8]



Conclusion

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. **Conclusions**
7. Q&A

- Reduce data labelling costs with cheaper non-expert labelling.
- Reduce barriers to participation and interdisciplinarity.
- Mitigates against learned biases from historic labelling schemes, and allows for new relationships (and potentially new physics) to be identified.
- Must be mindful of the anglocentric nature of our current approach and the potential biases that may introduce.
- Fine tune models trained on semantic targets to your specific science case with expert labels.

Q&A

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. **Q&A**

- Reduce data labelling costs with cheaper non-expert labelling.
- Reduce annotator education costs by using plain English.
- Broader impact on collaboration, inclusivity, language barriers, barriers to participation and interdisciplinarity.
- Moving away from historical labelling schemes mitigates against learned biases and allows for new relationships (and potentially new physics) to be identified.
- Must be mindful of the anglocentric nature of our current experiment and the potential biases that may introduce.
- Fine tune your semantically trained models to your specific science case with expert labels.

micah.bowles@postgrad.manchester.ac.uk
mb010.github.io
github.com/mb010/Text2Tag

References

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. **Q&A**

[1] Deng, J. et al., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. https://www.image-net.org/

[2] Dehghani, M. et al. 2023. Scaling Vision Transformers to 22 Billion Parameters. arXiv e-prints. https://arxiv.org/abs/2302.05442

[3] Bowles, M. et al. 2022. A New Task: Deriving Semantic Class Targets for the Physical Sciences. NeurIPS 2022: Machine Learning and the Physical Sciences Workshop. https://arxiv.org/abs/2210.14760

[4] Ramatsoku, M. et al. 2020. Collimated synchrotron threads linking the radio lobes of ESO 137-006. A&A. https://doi.org/10.1051/0004-6361/202037800

[5] Antoniou, G. and van Harmelen, F. 2004. Web Ontology Language: OWL, Springer Berlin Heidelberg, 67-92, https://doi.org/10.1007/978-3-540-24750-0_4

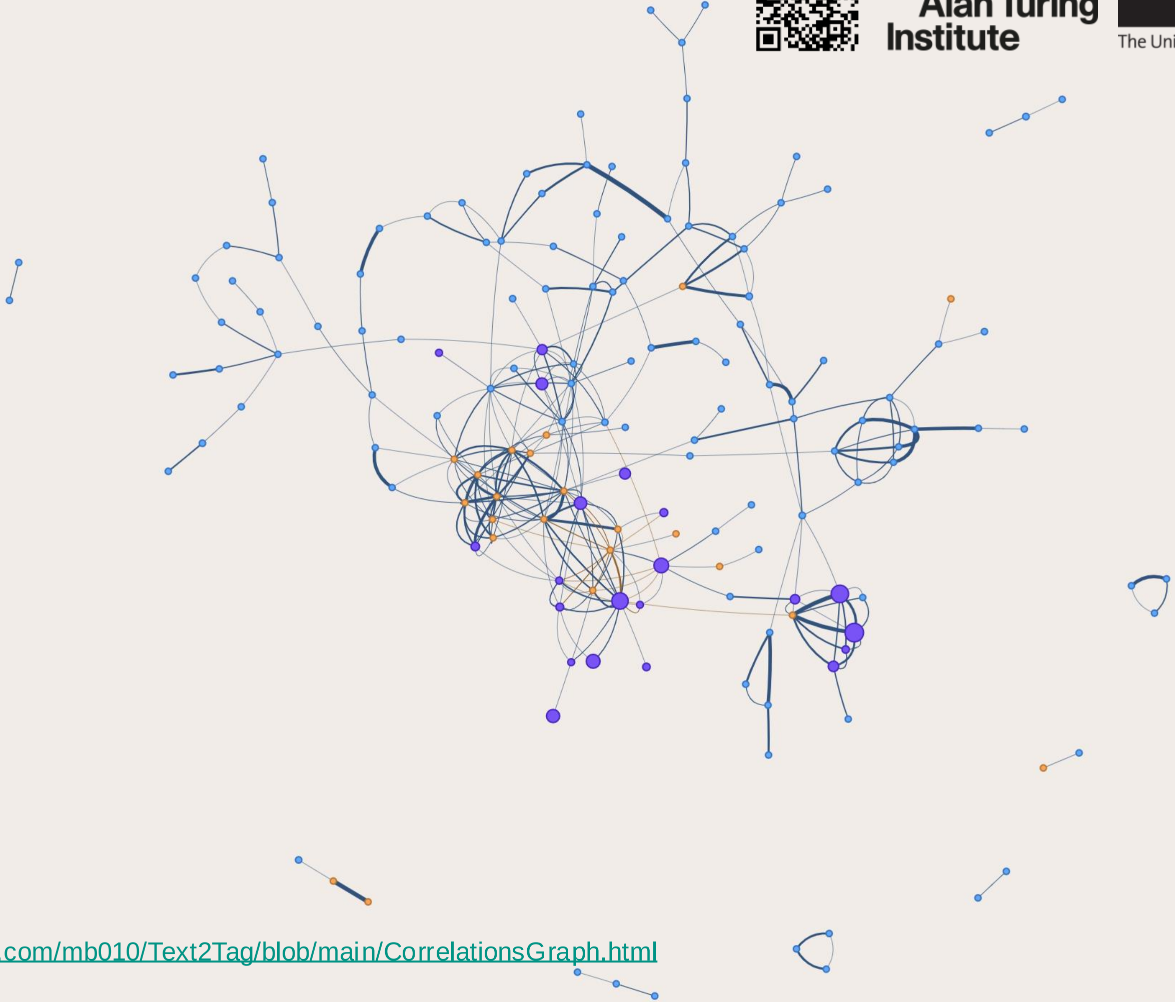
[6] Arp, R. et al. 2015. Building Ontologies with Basic Formal Ontology. The MIT Press. https://www.jstor.org/stable/j.ctt17kk7ww

[7] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41. https://wordnet.princeton.edu/

[8] Bowles, M. et al., 2023. Radio galaxy zoo EMU: towards a semantic radio galaxy morphology taxonomy. MNRAS. https://doi.org/10.1093/mnras/stad1021

WordNet

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. **Q&A**



github.com/mb010/Text2Tag/blob/main/CorrelationsGraph.html

WordNet

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. **Q&A**

“WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations”

WordNet
https://wordnet.princeton.edu/

“WordNet is sometimes called an ontology, a persistent claim that its creators do not make.”

Wikipedia
https://en.wikipedia.org/wiki/WordNet#As_a_lexical_ontology

Challenging Language

1. Introduction
2. Pre-Training
3. Semantics
4. Early approach
5. Application
6. Conclusions
7. **Q&A**

The Alan Turing Institute

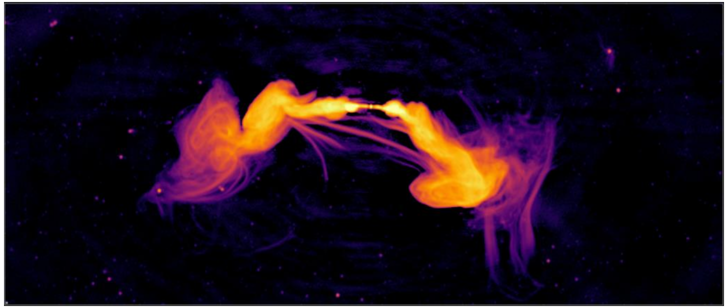
MANCHESTER 1824 The University of Manchester

Case study: Radio Astronomy

Specifically: Radio Galaxy Morphology

Current Examples:

- FRI / FRII
- Hybrid
- WAT
- NAT
- X-Shaped
- ORC



Ramatsoku et al. 2020 || South African Radio Astronomy Observatory (SARAO)

Limitations:

1. Doesn’t span the full space of possible morphologies
2. No full information (unlike "neutron").
3. Additional education required.
4. Occasionally no consensus as the bounds of classes are neither linear nor clear!