ZOØNIVERSE

Label-Efficient Learning at Galaxy Zoo

Mike Walmsley, with Micah Bowles, Inigo Slijepcevic, Anna Scaife

University of Toronto

UNIVERSITY OF TORONTO

MANCHESTER 1824

The Alan Turing Institute

zooniverse.org

@mike_walmsley_

ZOØNIVERSE

Live Demo

bit.ly/decals_viz

Try it yourself: **bit.ly/decals_viz**

ZOONIVERSE

Has spiral arms?

Answer

Yes

Posterior Mean

0.73 1.00

0.00 1.00

Spiral arm count?

Answer

1

Posterior Mean

0.00 1.00

0.00 1.00

Spiral winding?

Answer

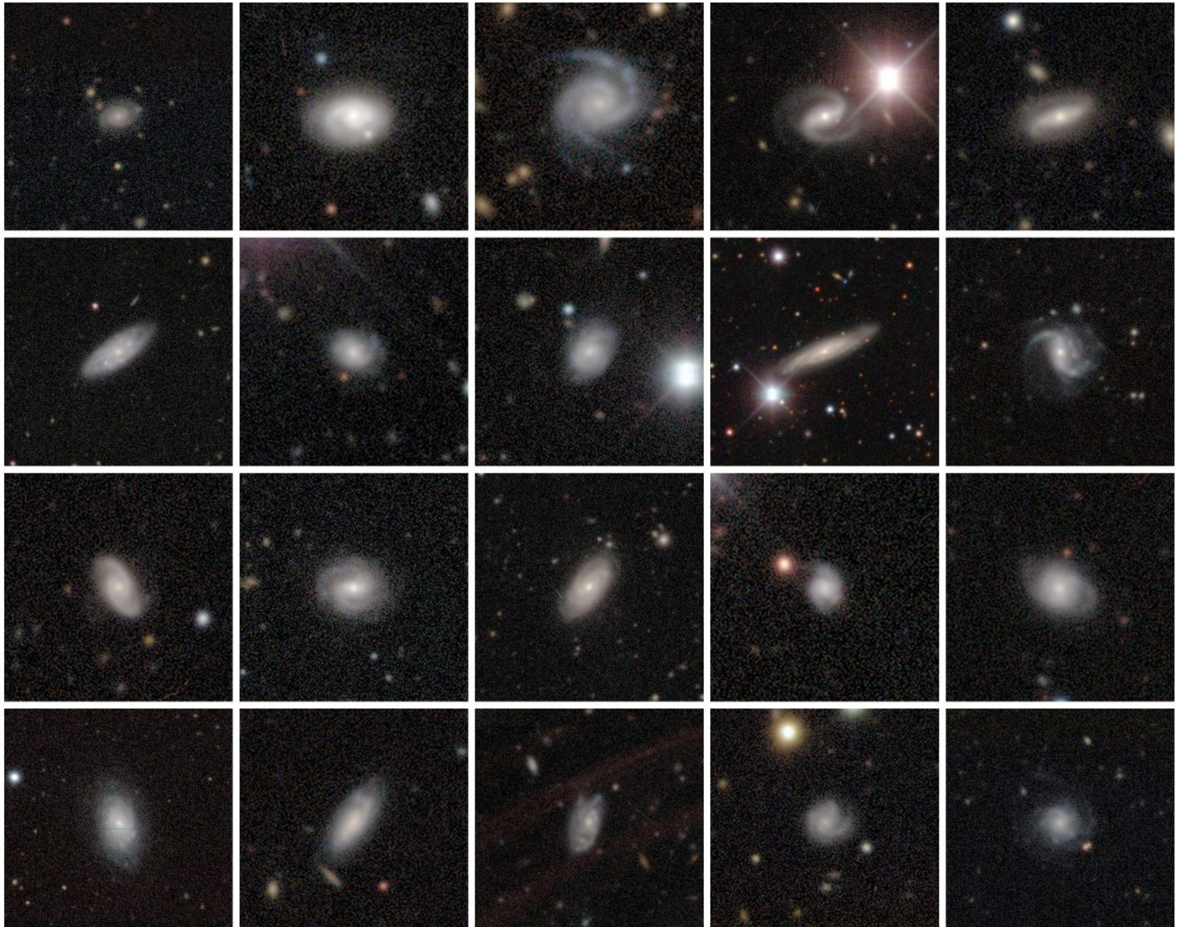
Tight

Posterior Mean

0.00 1.00

0.00 1.00

Merging?



Try it yourself: **bit.ly/decals_viz**

ZOØNIVERSE

Has spiral arms?

Answer

Yes

Posterior Mean

0.73 1.00

0.00 1.00

Spiral arm count?

Answer

2

Posterior Mean

0.74 1.00

0.00 1.00

Spiral winding?

Answer

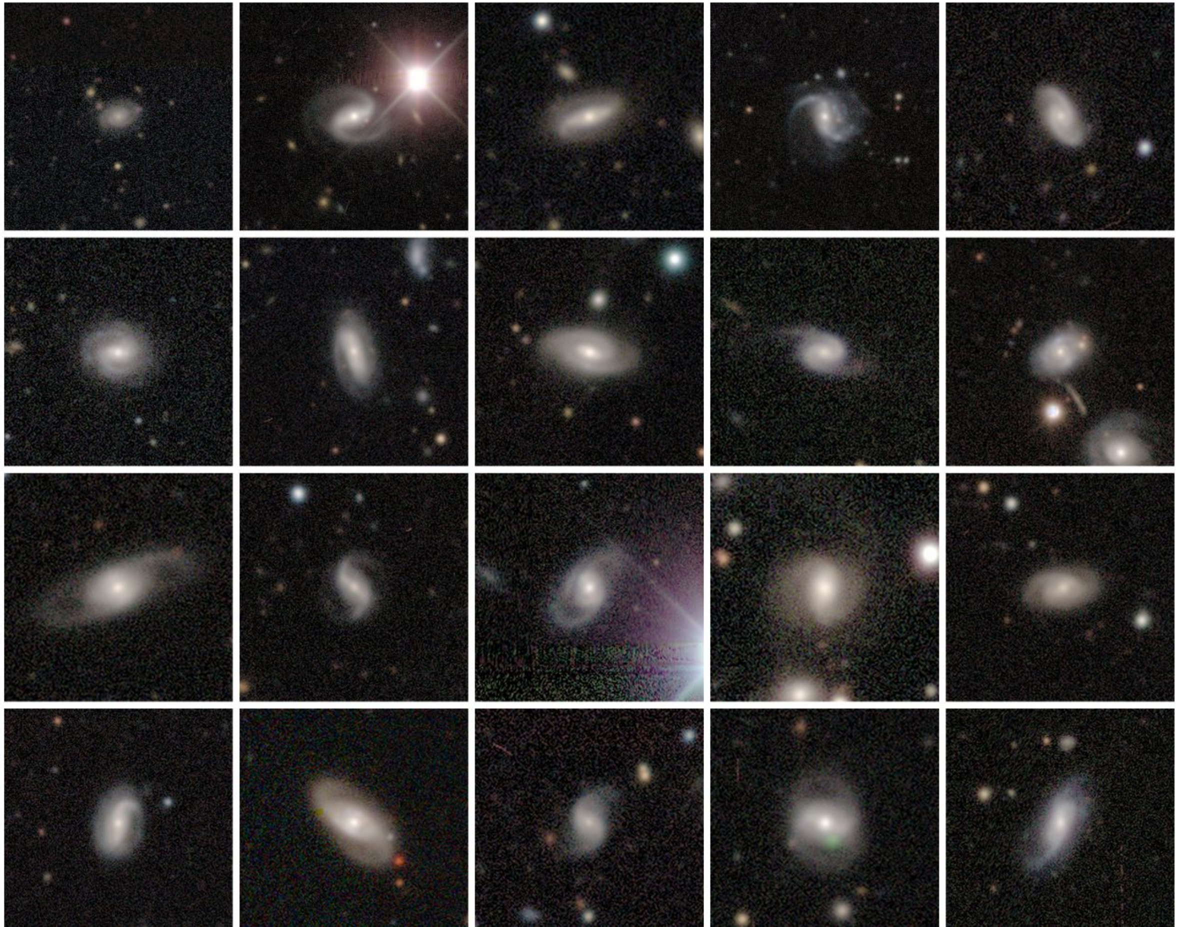
Tight

Posterior Mean

0.00 1.00

0.00 1.00

Merging?



Try it yourself: **bit.ly/decals_viz**

ZOONIVERSE

Has spiral arms?

Answer

Yes

Posterior Mean

0.74 1.00

0.00 1.00

Spiral arm count?

Answer

2

Posterior Mean

0.74 1.00

0.00 1.00

Spiral winding?

Answer

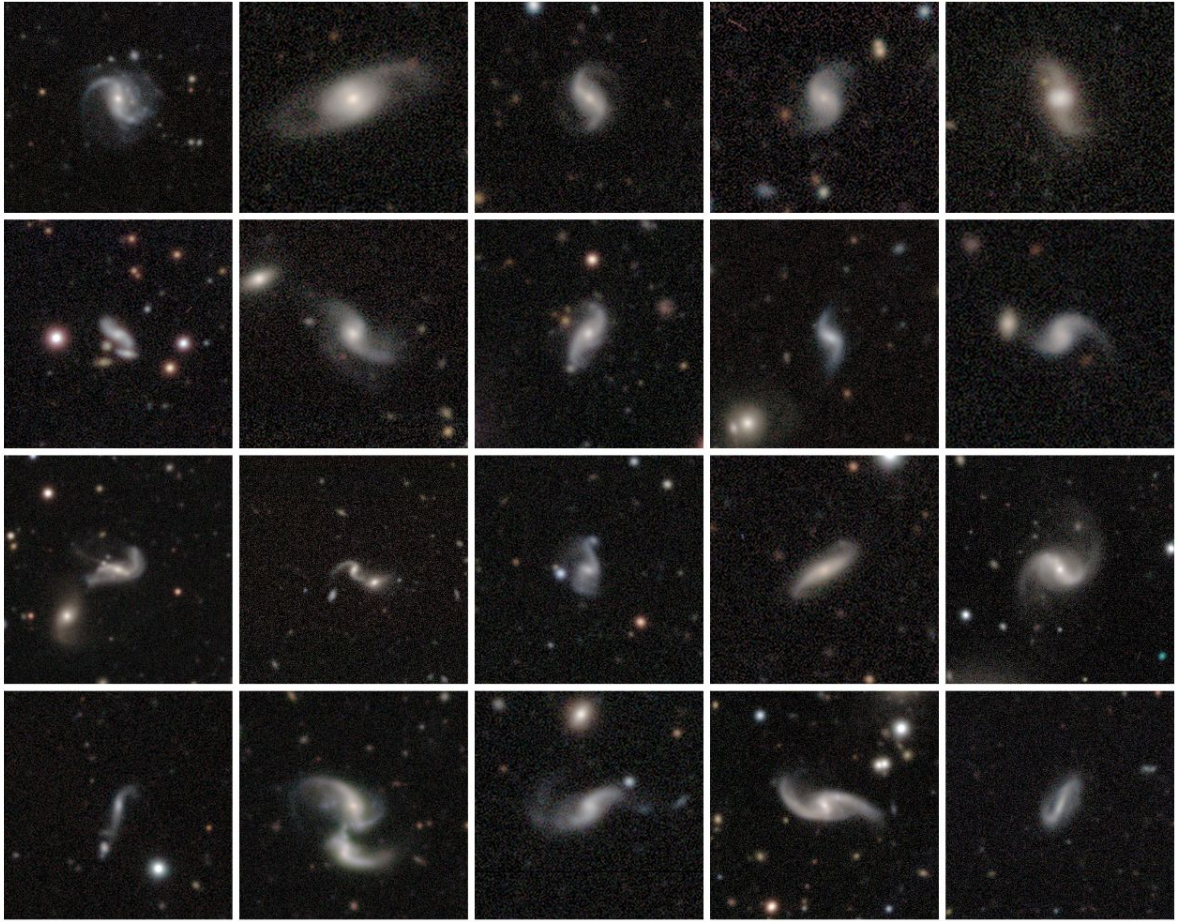
Loose

Posterior Mean

0.71 1.00

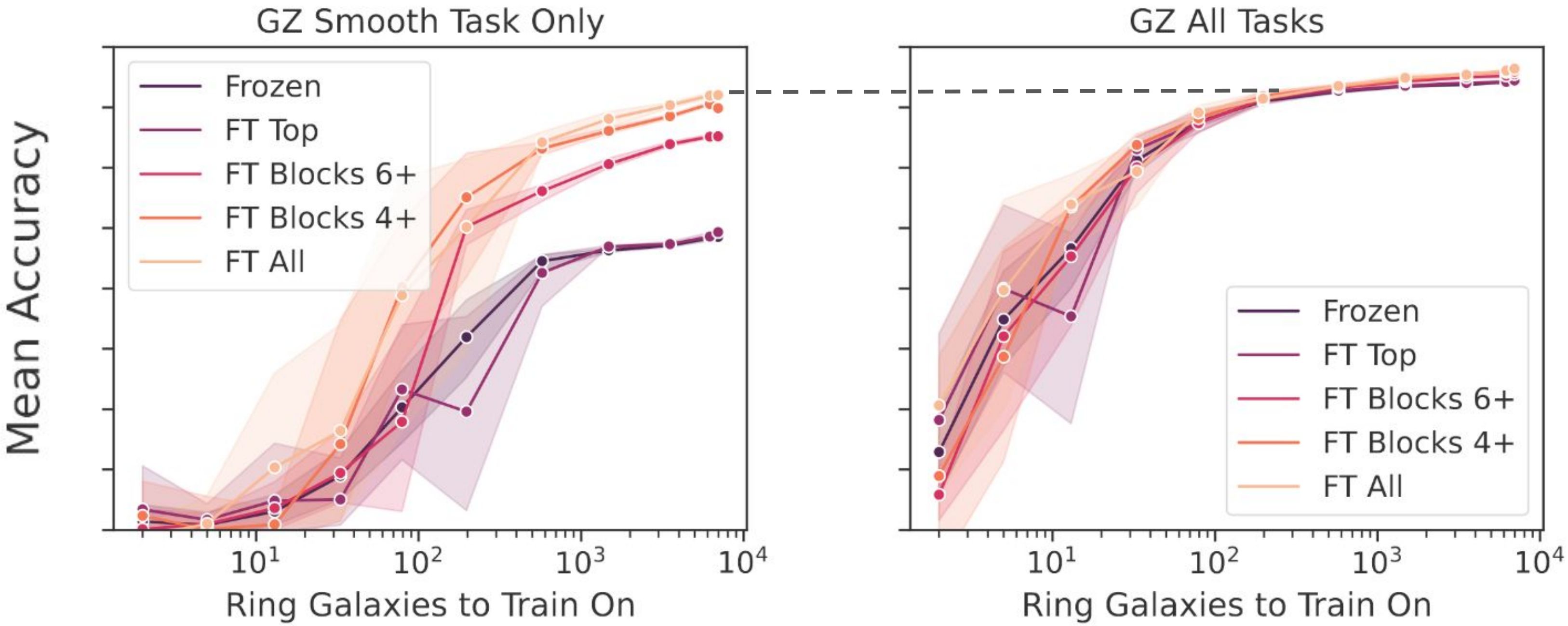
0.00 1.00

Merging?



1. Multi-task — *all GZ questions, all surveys*
2. Pretraining — *contrastive + supervised*
3. Segmentation — *pixelwise labels & prediction*

Many Questions = More General Representation



Mike Walmsley et al

Live Demo #2

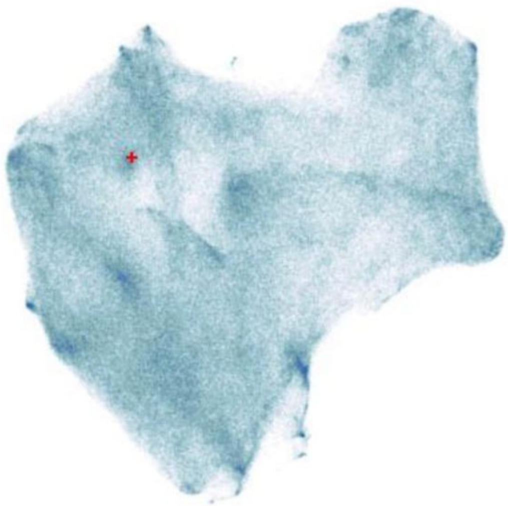
bit.ly/gz-explorer

Move Around

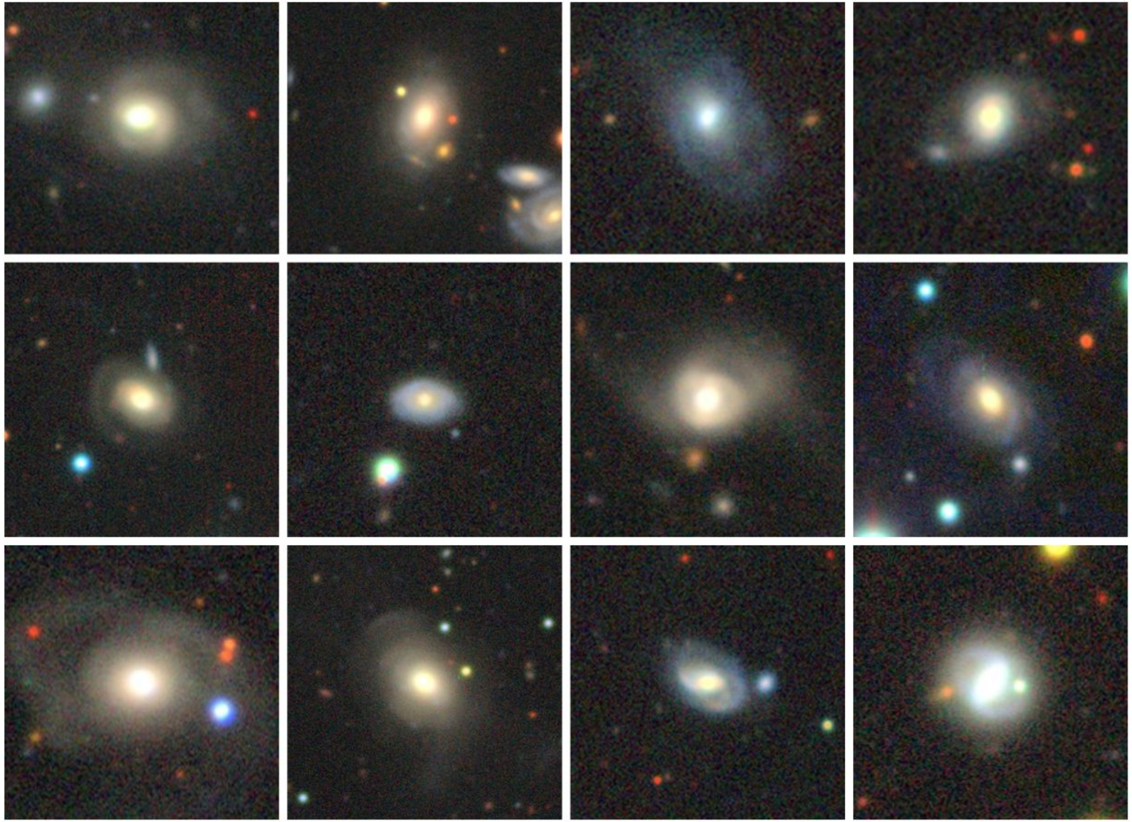
Click anywhere on the latent space to view galaxies.

Select Reduction

Featured v2



Location: (-0.109, 6.410)



Download CSV of the 1000 galaxies closest to your search

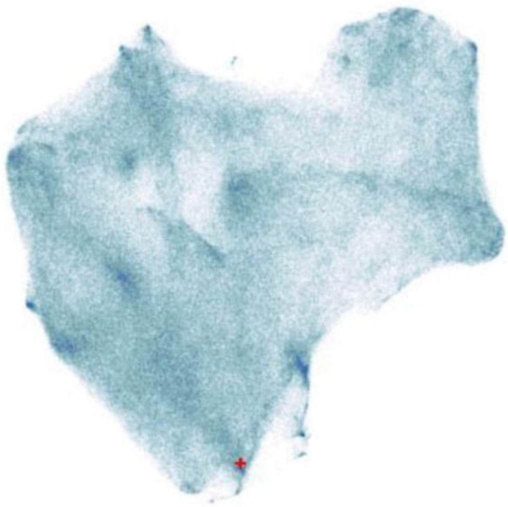
ZOONIVERSE

Move Around

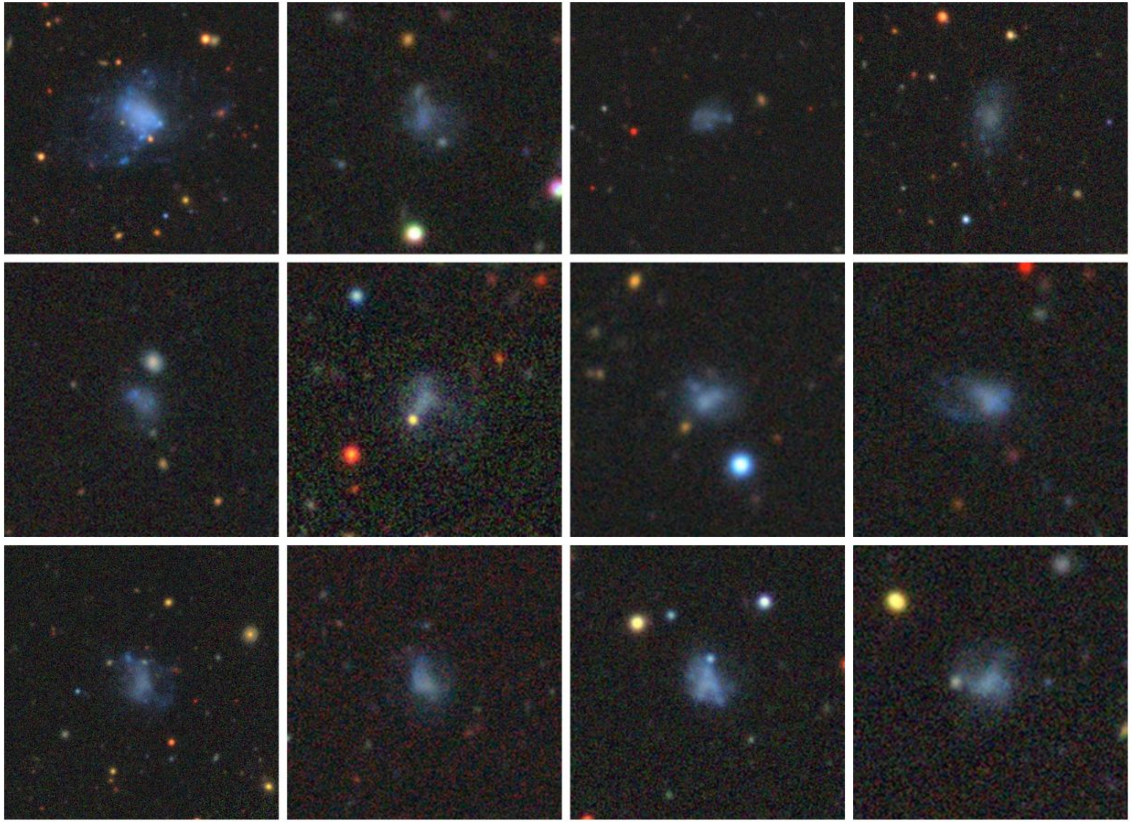
Click anywhere on the latent space to view galaxies.

Select Reduction

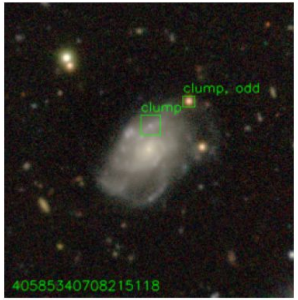
Featured v2



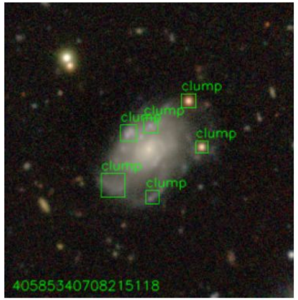
Location: (2.809, 0.812)



Download CSV of the 1000 galaxies closest to your search



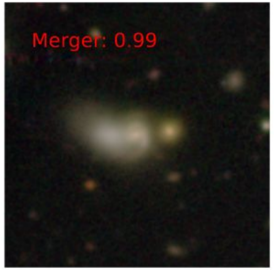
Scratch



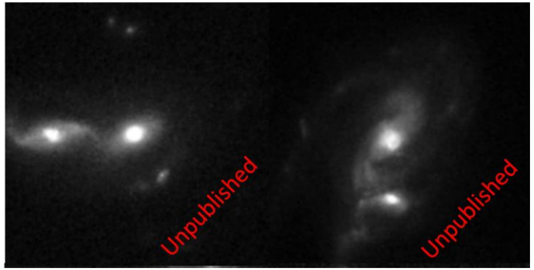
Zoobot

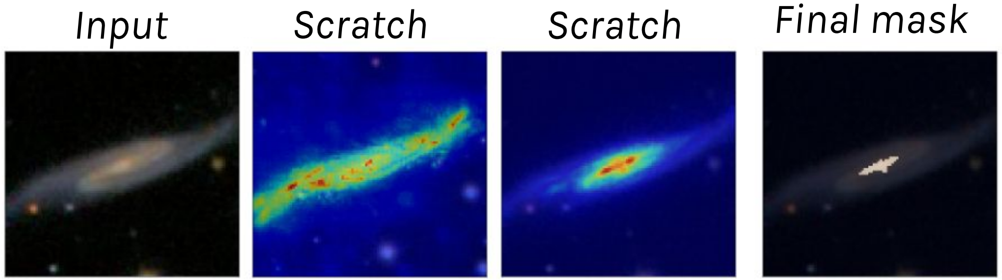
Jürgen Popp (prep.)

O'Ryan+23



Omori+23





Bhambra+ 22

```python
import pandas as pd
from galaxy_datasets.pytorch.galaxy_datamodule import GalaxyDataModule
from zoobot.pytorch.training import finetune

# csv with 'ring' column (0 or 1) and 'file_loc' column (path to image)
labelled_df = pd.read_csv('/your/path/some_labelled_galaxies.csv')

datamodule = GalaxyDataModule(
  label_cols=['ring'],
  catalog=labelled_df,
  batch_size=32
)

# load trained Zoobot model
model = finetune.FinetuneableZoobotClassifier(checkpoint_loc, num_classes=2)

# retrain to find rings
trainer = finetune.get_trainer(save_dir)
trainer.fit(model, datamodule)
```

Quickstart example from ***github.com/mwalmsley/zoobot***

1. Multi-task *all GZ questions, all surveys*

2. Pretraining *contrastive + supervised*

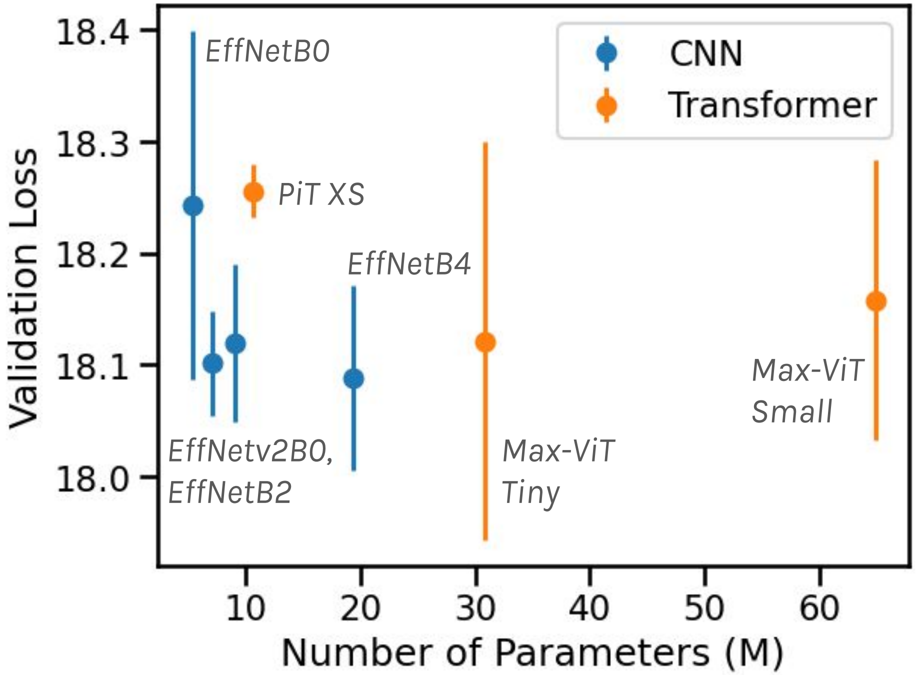
3. Segmentation *pixelwise labels & prediction*

Make it bigger?

Foundation model 'scaling laws' also require data. But we're now

No evidence that larger models work better.

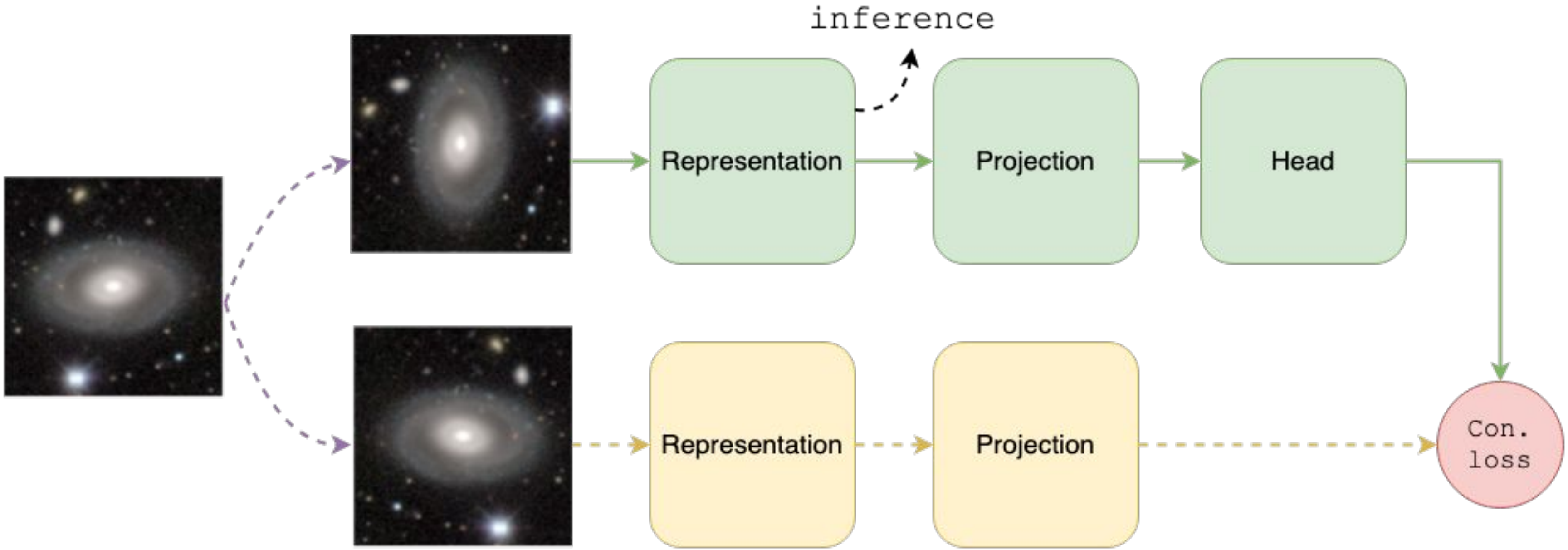
More GPUs are not all you need.



Mike Walmsley et al

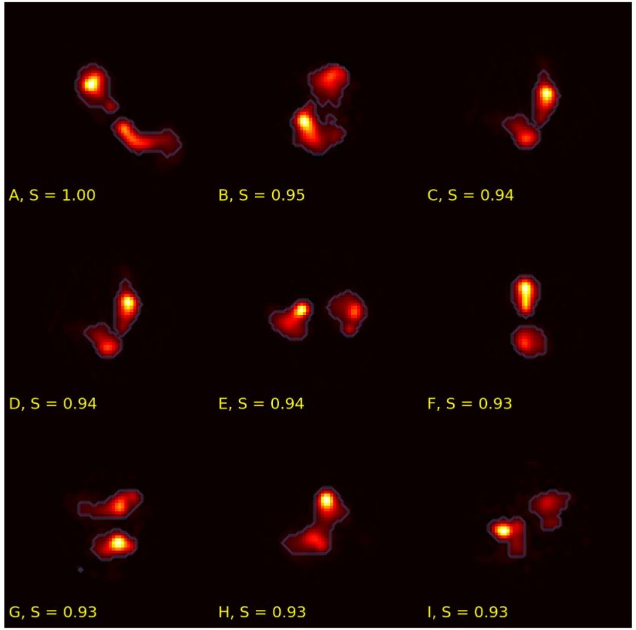
Contrastive Learning with BYOL

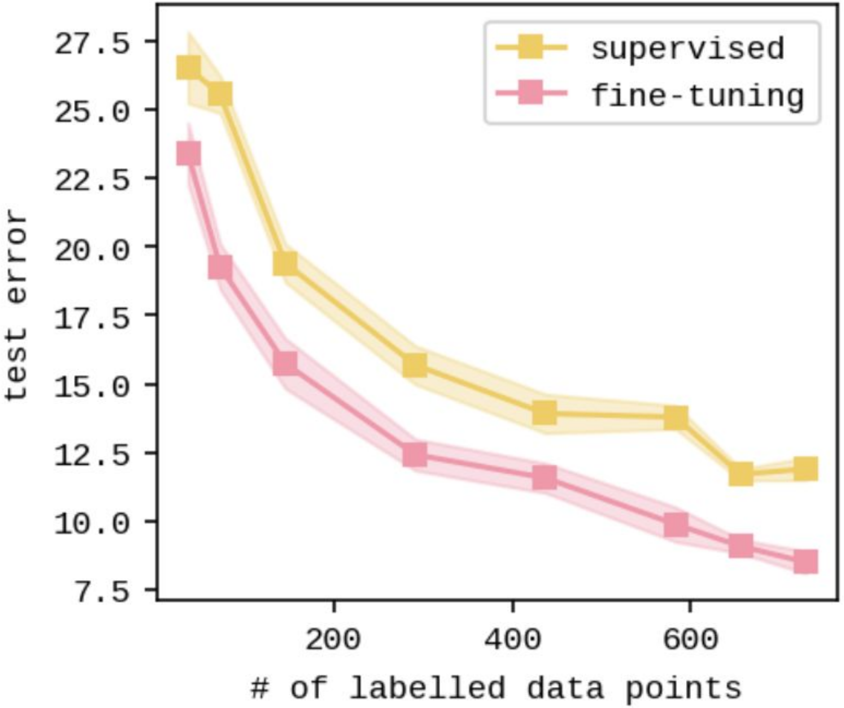
ZOONIVERSE



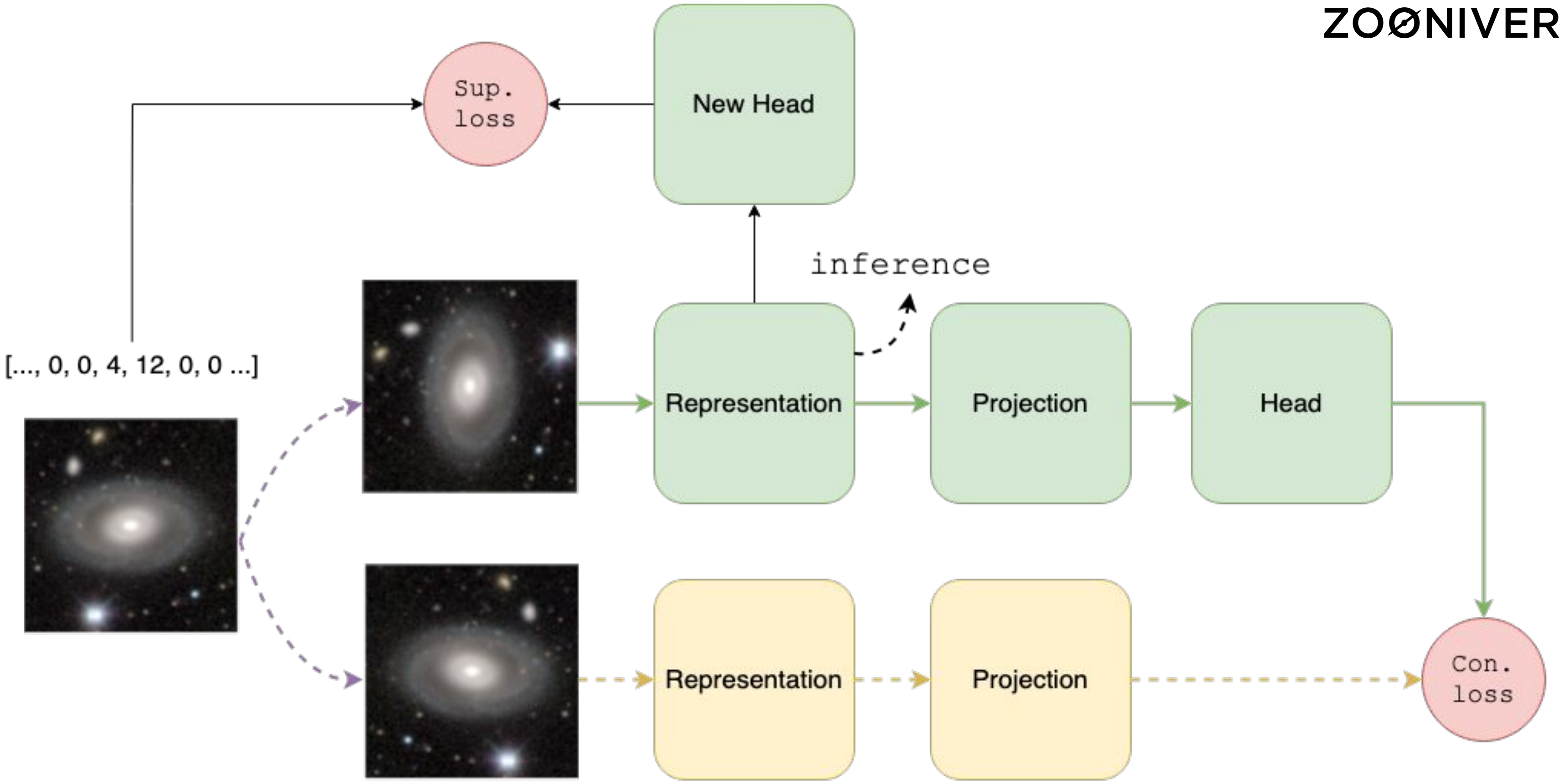
Introduced by Grill (2020)

ZOØNIVERSE





For more, see Slijepcevic et. al. (2023)

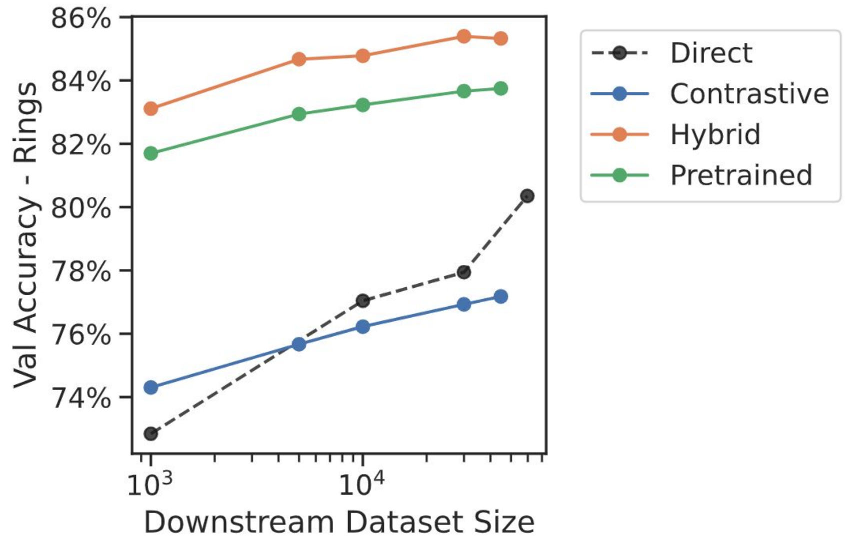


Adding a **new supervised head** to guide the contrastive representation

Mike Walmsley et al

Performance vs. Previous Best Methods

- Contrastive learning beats a few thousand labels
- Large-scale pretraining does better, even with ~50k labels
- Hybrid learning (both contrastive and pretraining) does best



Next - Masked AE

Work by **Micah Bowles**

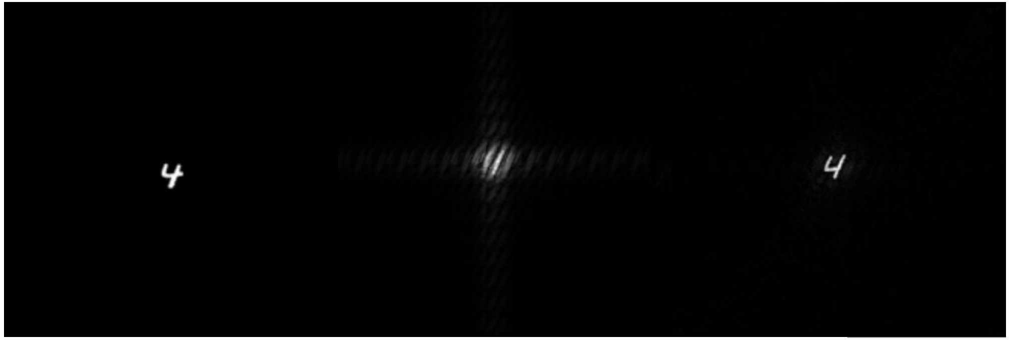
Pixel masking is effective at large-scale pretraining on **dense** images

Fourier masking works better on **sparse** galaxy images

Input | **Pixel** masked | Reconstructed

He 2021

Input | **Fourier** masked | Reconstructed



Bowles (prep.)

1. Multi-task *all GZ questions, all surveys*
2. Pretraining *contrastive + supervised*
3. Segmentation *pixelwise labels & prediction*

ZOONIVERSE

Input

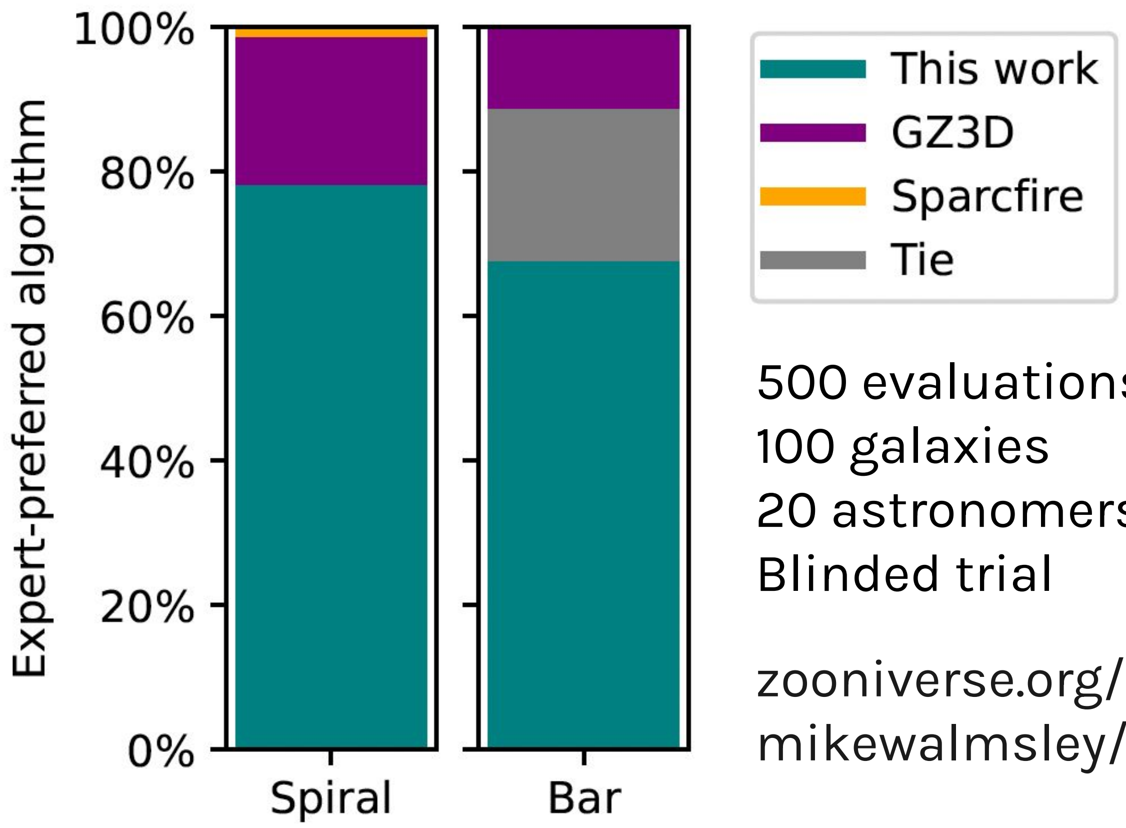
Volunteers

Masters+'21

sparcfire

Davis+'14

Walmsley & Spindler (submitted)



500 evaluations
100 galaxies
20 astronomers
Blinded trial

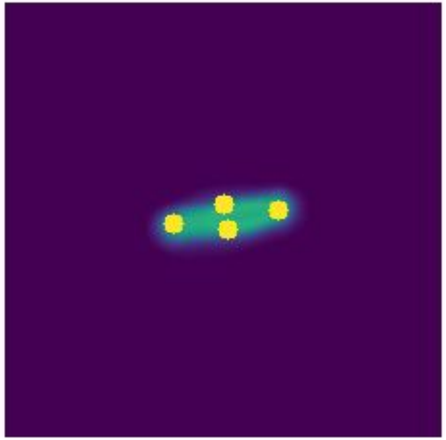
zooniverse.org/projects/mikewalmsley/galaxy-judges

ZOØNIVERSE

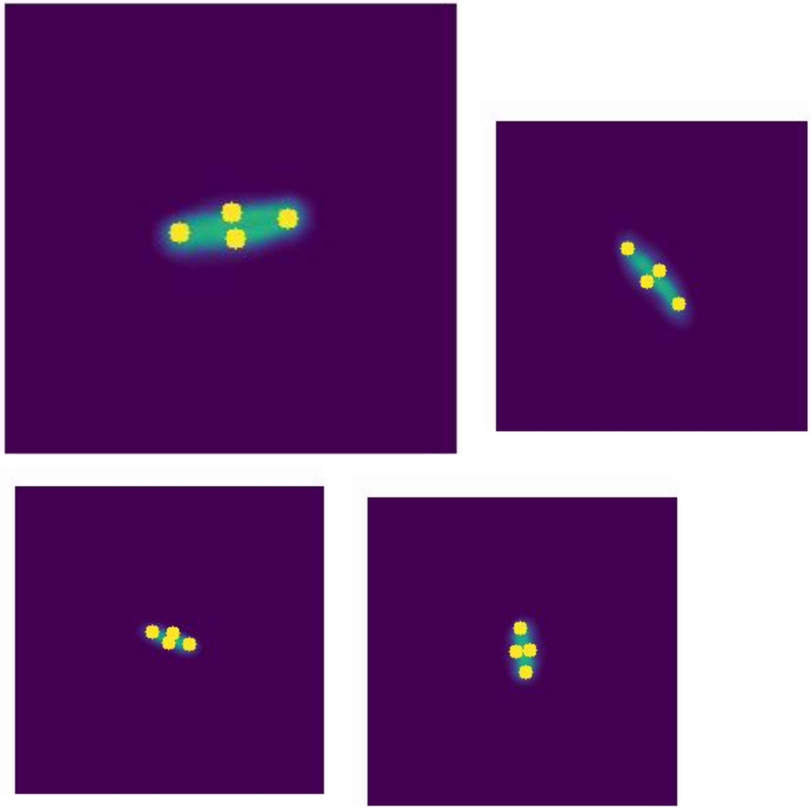
Image

Prediction

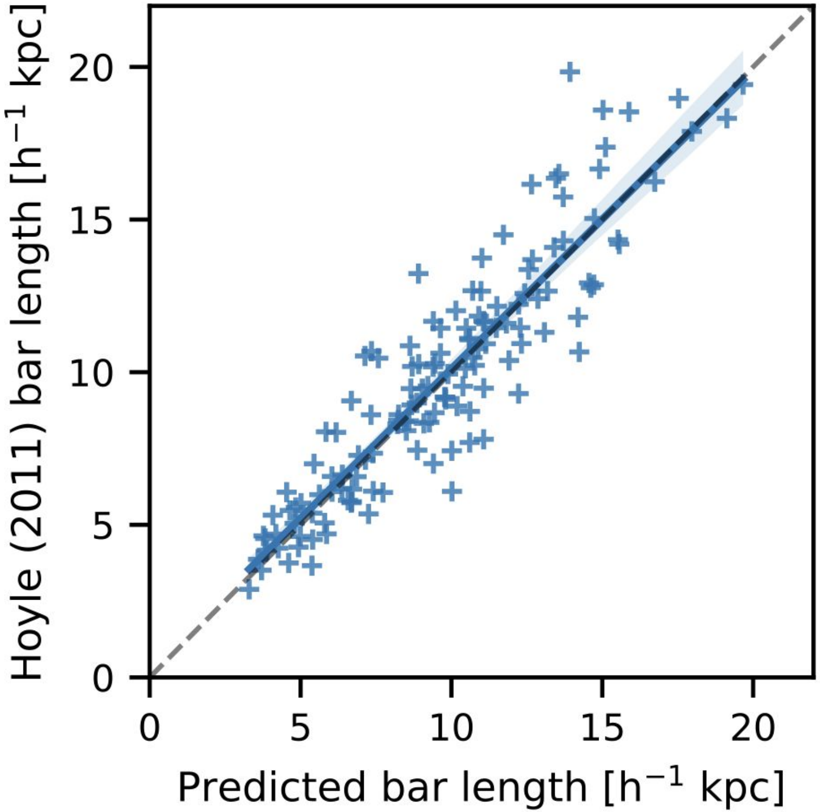
Bar extent



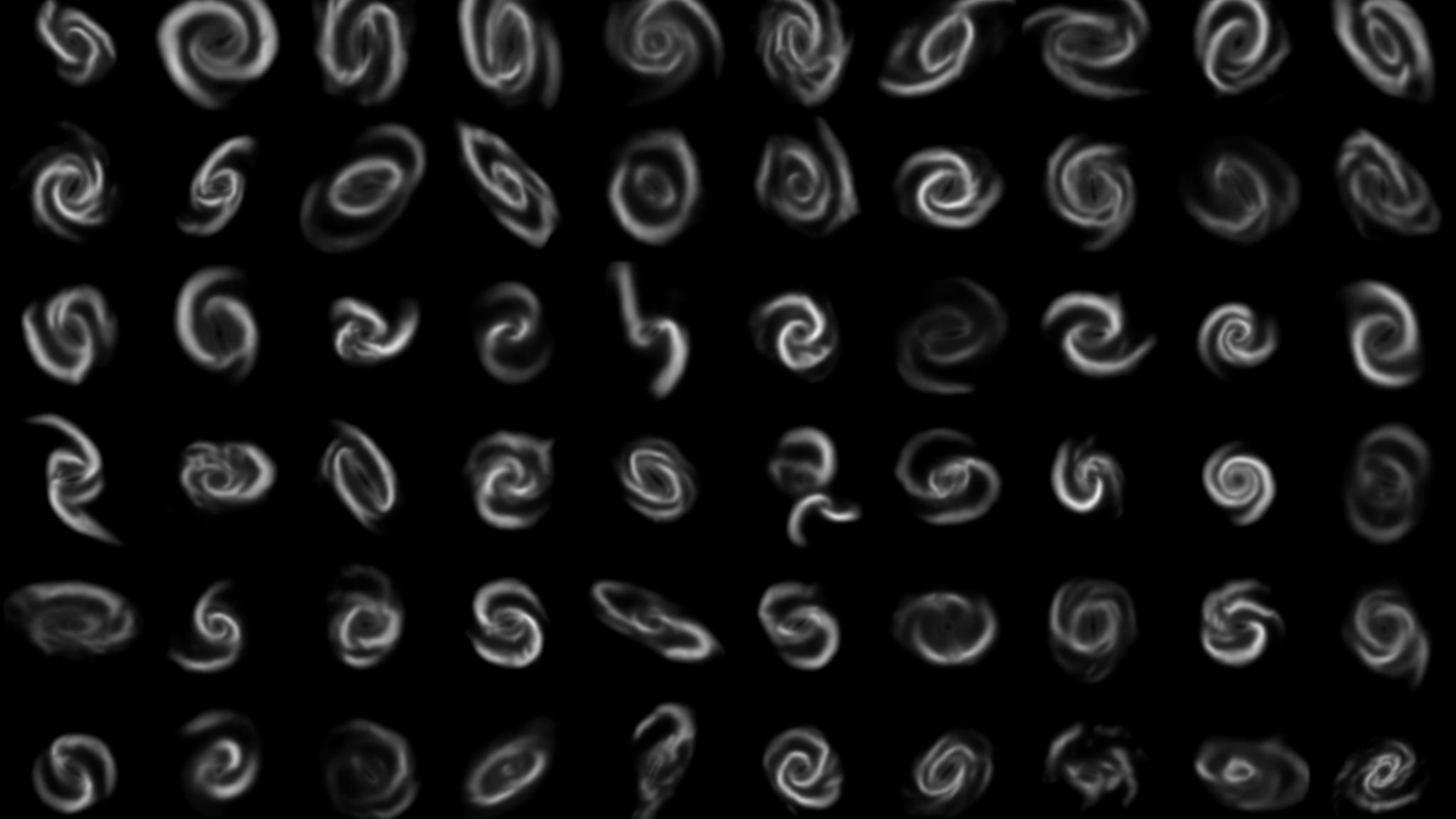
ZOØNIVERSE



Repeat for many bars



Mike Walmsley et al



1. Multi-task *all GZ questions, all surveys*

2. Pretraining *contrastive + supervised*

3. Segmentation *pixelwise labels & prediction*

github.com/mwalmsley/zoobot

Zoobot: Adaptable Deep Learning Models for Galaxy Morphology
Walmsley et. al. (2023), JOSS

Towards Galaxy Foundation Models with Hybrid Contrastive Learning
Walmsley et. al. (2022), ICML ML4Astro

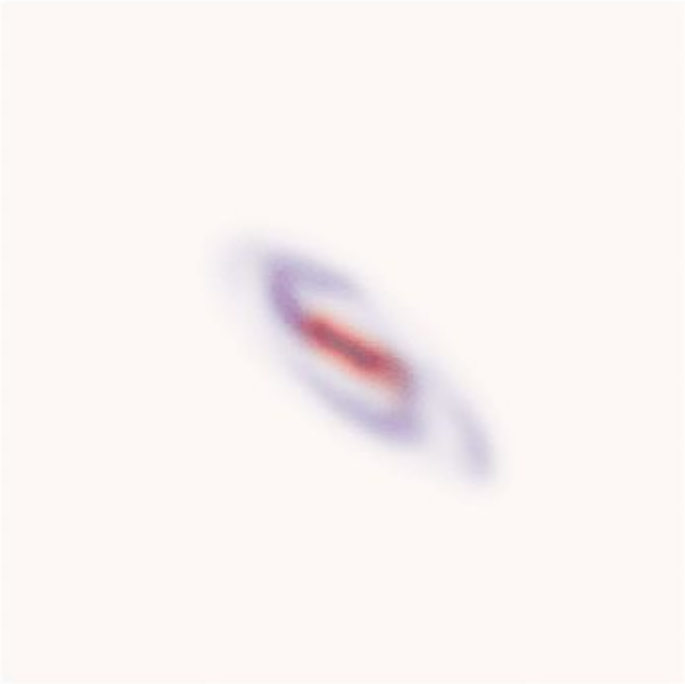
Building a Multi-Purpose Foundation Model for Radio Astronomy
Slijepcevic et. al. (2023), RASTI

Neurips ML4Phys next week: (1) exploring latent space (2) segmentation

Adaptable models

Targeted labelling

SSL Pretraining

Mike Walmsley et al

We are trying to answer the same kind of questions across every survey

Does the galaxy image show {visual feature}
Barred?
How many arms?
Tidal streams?

Estimate {physical property} for this galaxy
Most recent merger?
Bulge mass?
Gas mass?

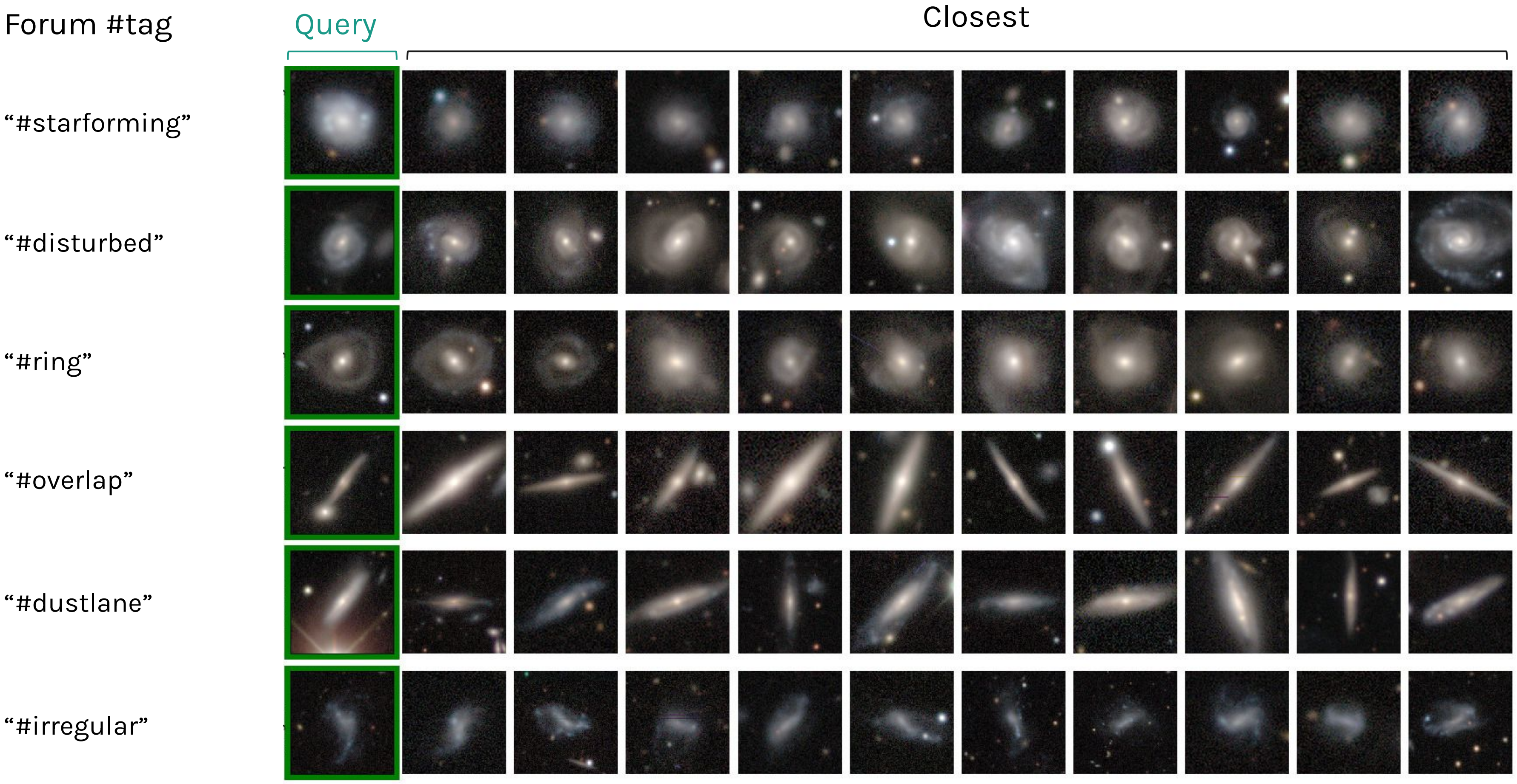
Let's **build a shared tool**

Your survey here

DECALS/DESI, HSC, EUCLID, RUBIN... (and that's just the major ones)

Live Demo

bit.ly/decals_viz



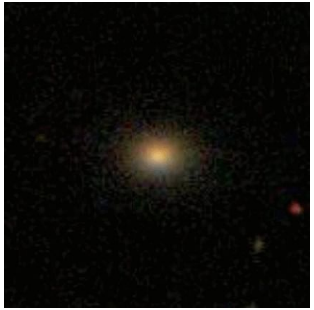
Mike Walmsley et al

ZOØNIVERSE

1. Simple yet extensible foundation *software-like*
2. Pre-trained to solve many tasks *AI-like*
3. The more people use it, the more it learns *AI-like*

arXiv:2206.11927

ZOONIVERSE



GZ2
210k, z < 0.15

Willett (2013)

GZ DECaLS
230k, z < 0.15

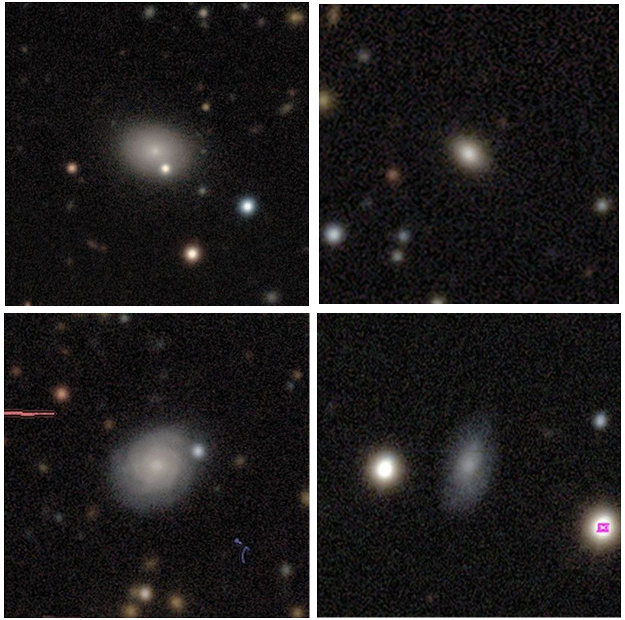
Walmsley (2022)

GZ Hubble
106k, z < 1

GZ CANDELS
50k, 1 < z < 3

Willett (2017)
Simmons (2017)

and



plus **1.3m unlabelled**
from GZ DESI (z < 0.3)

Walmsley (submitted)

README.md

galaxy-datasets

ML-friendly datasets for major Galaxy Zoo citizen science campaigns.

- PyTorch Datasets and PyTorch Lightning DataModules
- TensorFlow tf.data.Dataset's
- Framework-independent (i.e. TensorFlow-friendly) download and augmentation code

You may also be interested in Galaxy MNIST as a simple dataset for teaching/debugging.

Name	Method	PyTorch Dataset	Published	Downloadable	Galaxies
Galaxy Zoo 2	gz2	GZ2	✅	✅	~210k (main sample)
GZ Hubble*	gz_hubble	GZHubble	✅	✅	~106k (main sample)
GZ CANDELS	gz_candels	GZCandels	✅	✅	~50k
GZ DECaLS GZD-5	gz_decals_5	GZDecals5	✅	✅	~230k (GZD-5 only)
GZ Rings	gz_rings	GZRings	❎	✅	~93k
GZ DESI	gz_desi	GZDesi	❎	WIP	WIP
CFHT Tidal*	tidal	Tidal	✅	✅	1760 (expert)

Shared Datasets

Convenient access to GZ images for ML practitioners

torch.Dataset, tf.data.Dataset, pl.DataModule

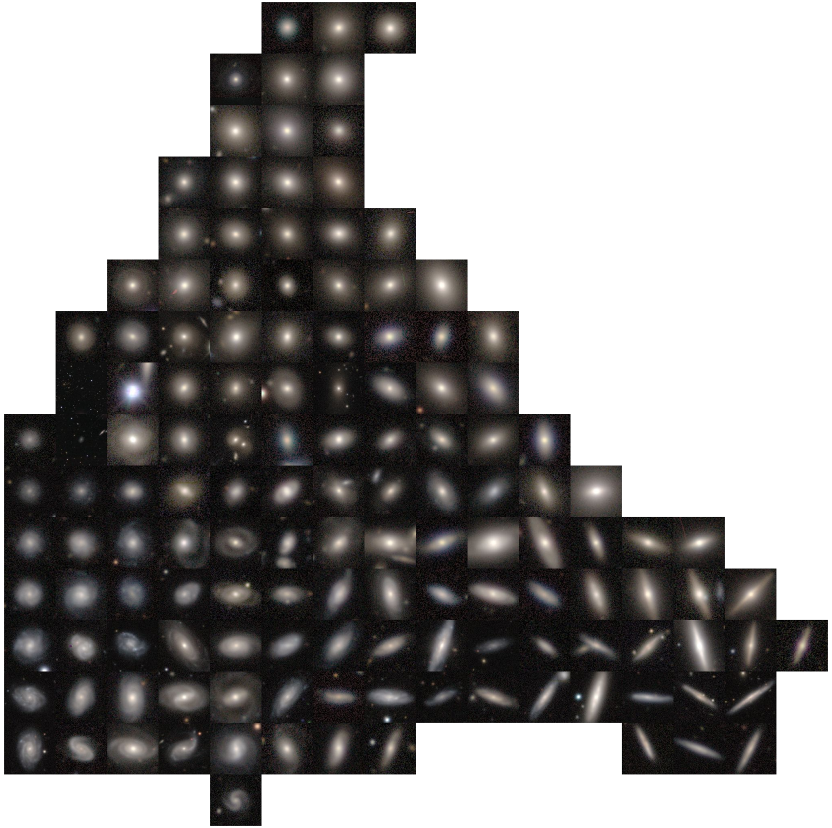
Self-downloading

See also GalaxyMNIST

Meaningful Internal Representation

ZOONIVERSE

Learned representation
(features before dense layers, PCA+UMAP)



Galaxies arranged by representation

Mike Walmsley et al

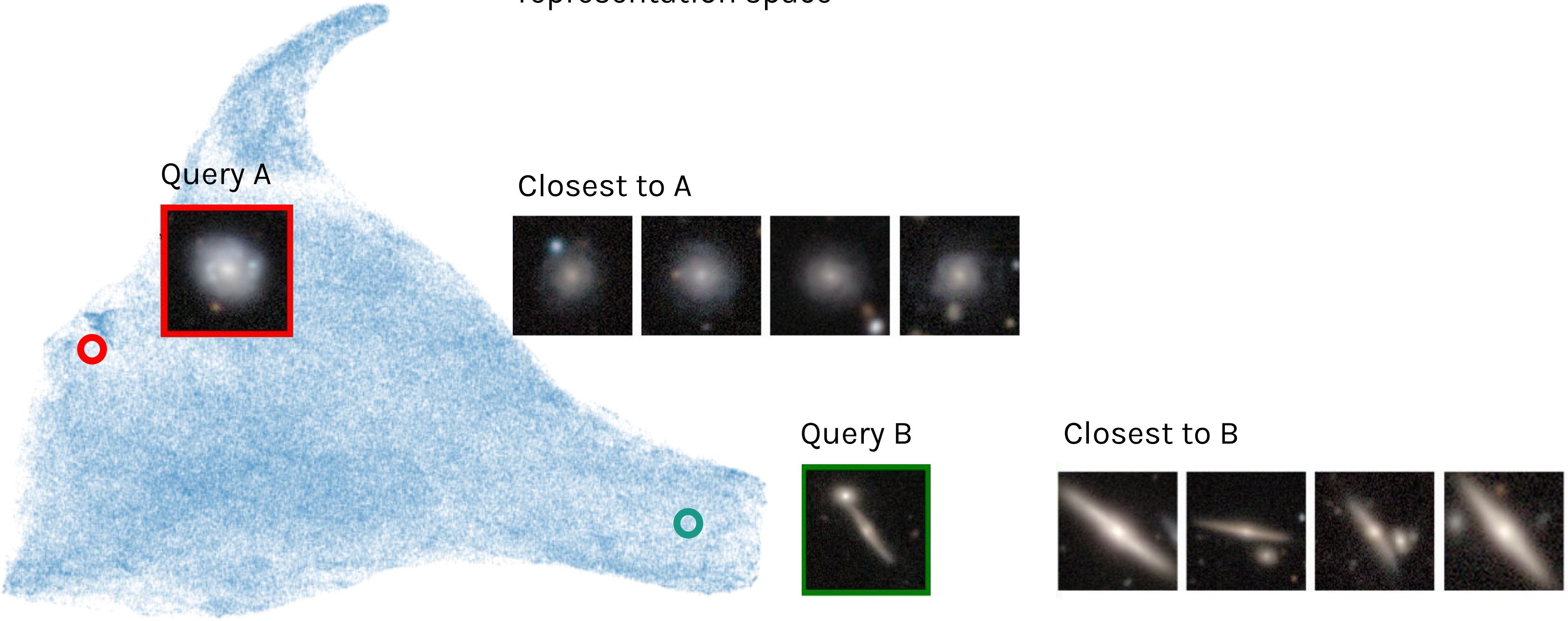
Similarity Search

ZOONIVERSE

Pick a galaxy...

...show the closest galaxies in representation space

Query A

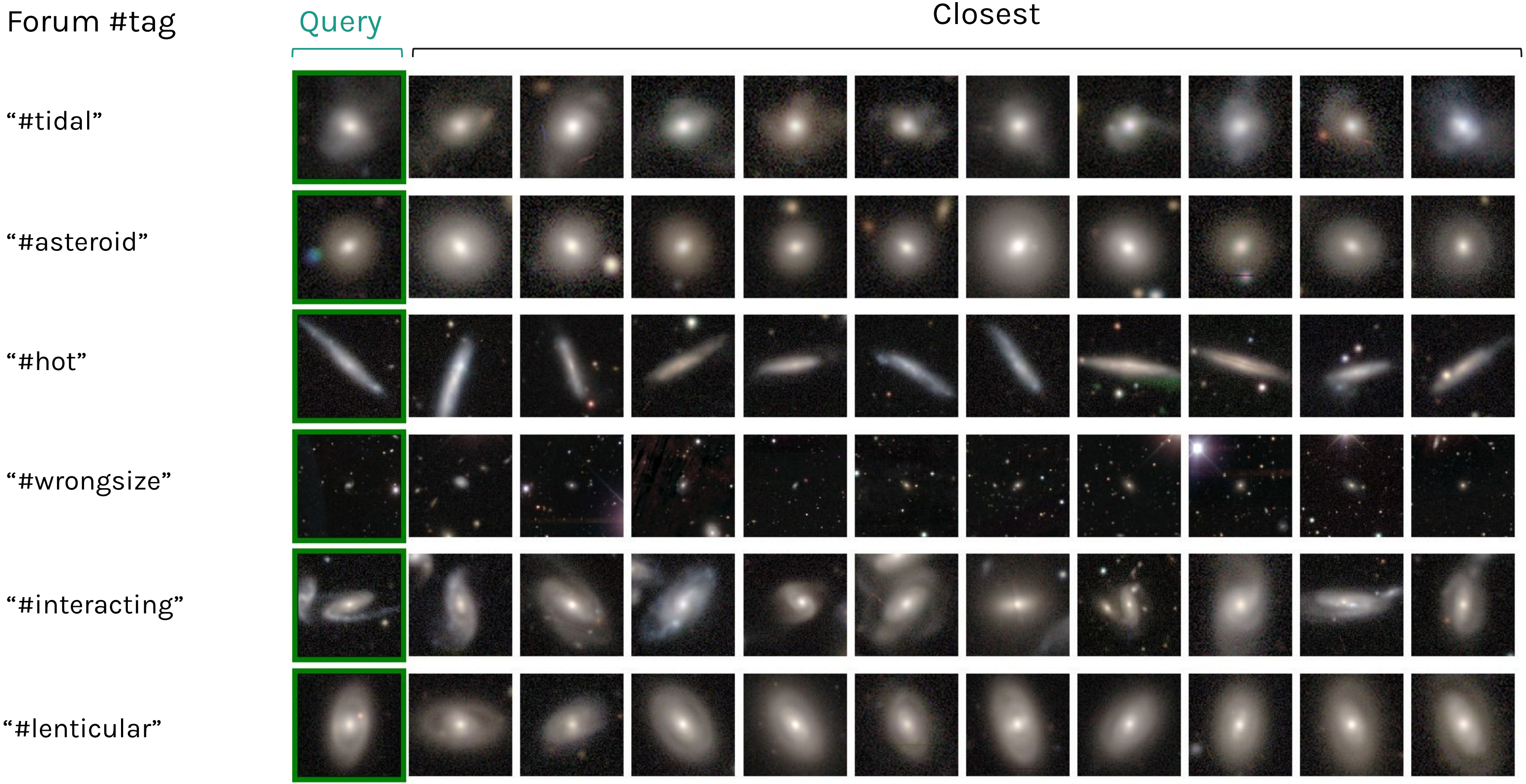


Closest to A

Query B

Closest to B

Mike Walmsley et al



Mike Walmsley et al

Try it yourself: **bit.ly/decals_viz**

ZOONIVERSE

Choose Your Galaxies

Bar?

Answer

Strong

Posterior Mean

0.00 1.00

0.00 1.00

Has spiral arms?

Answer

Yes

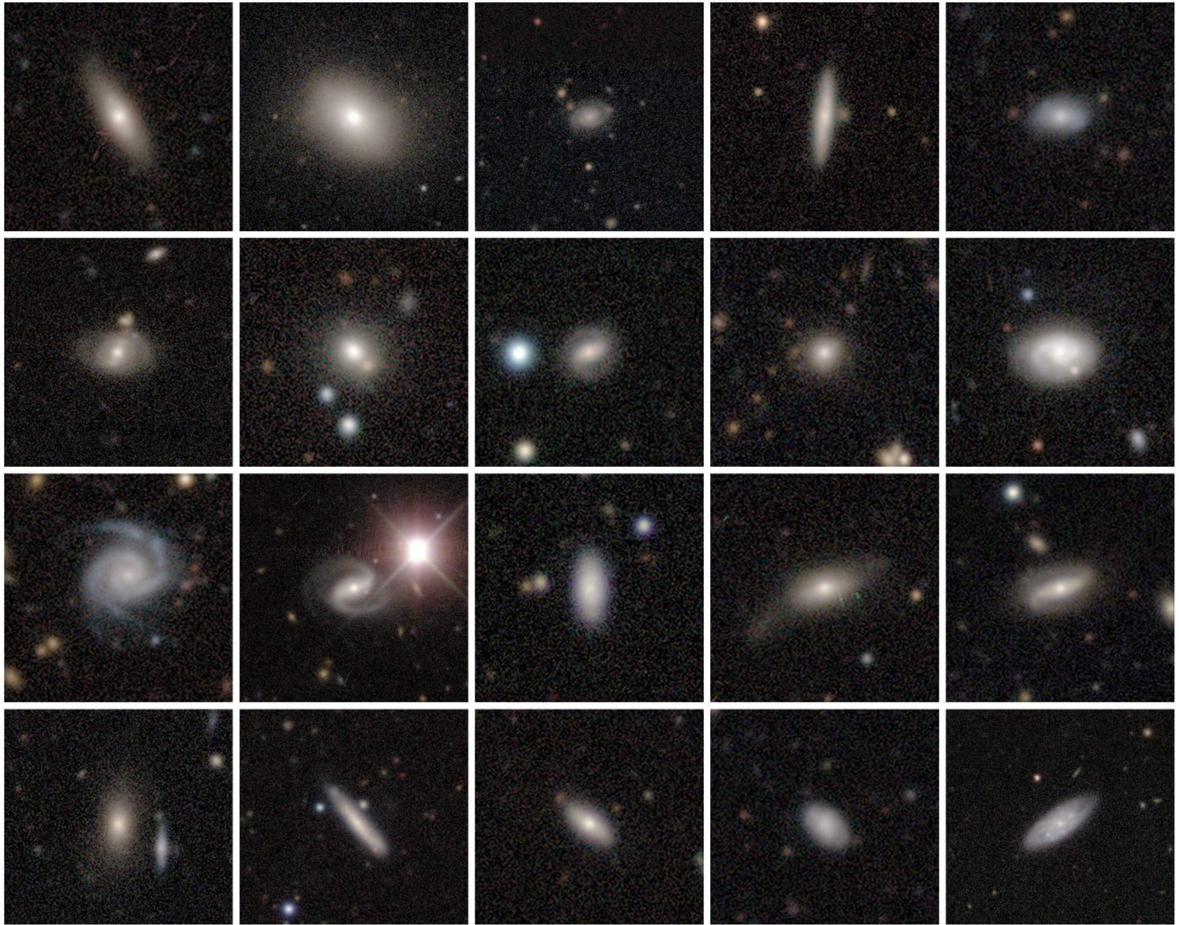
Posterior Mean

0.00 1.00

0.00 1.00

Spiral arm count?

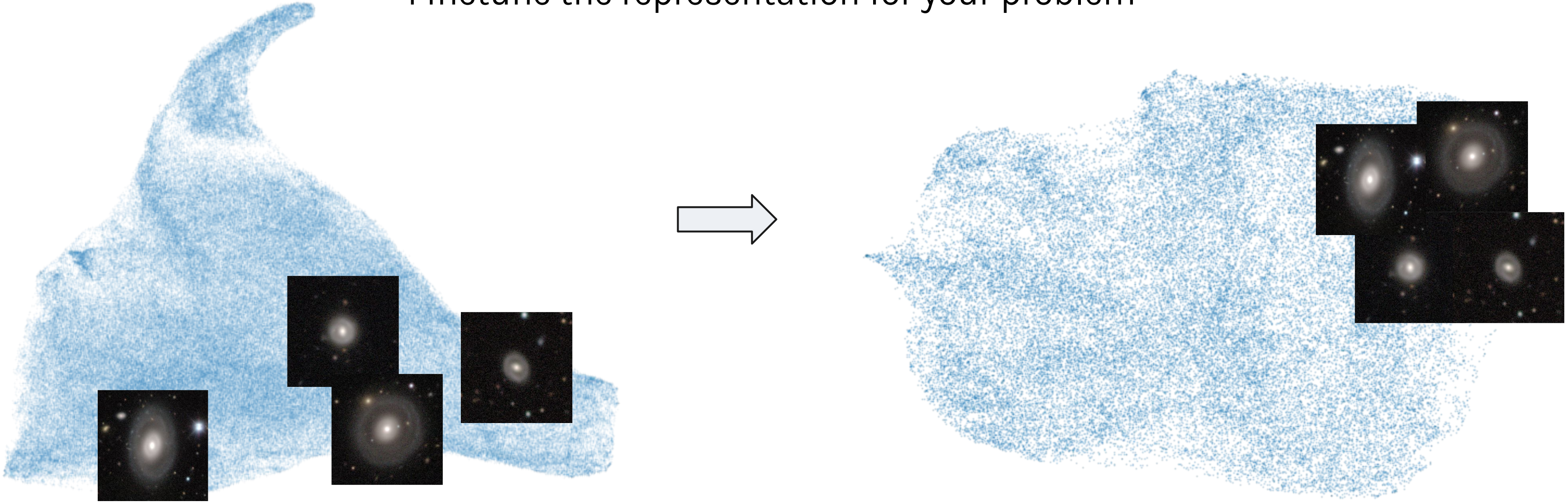
To use this filter, set "Has Spiral Arms = Yes"to > 0.5



Transfer Learning

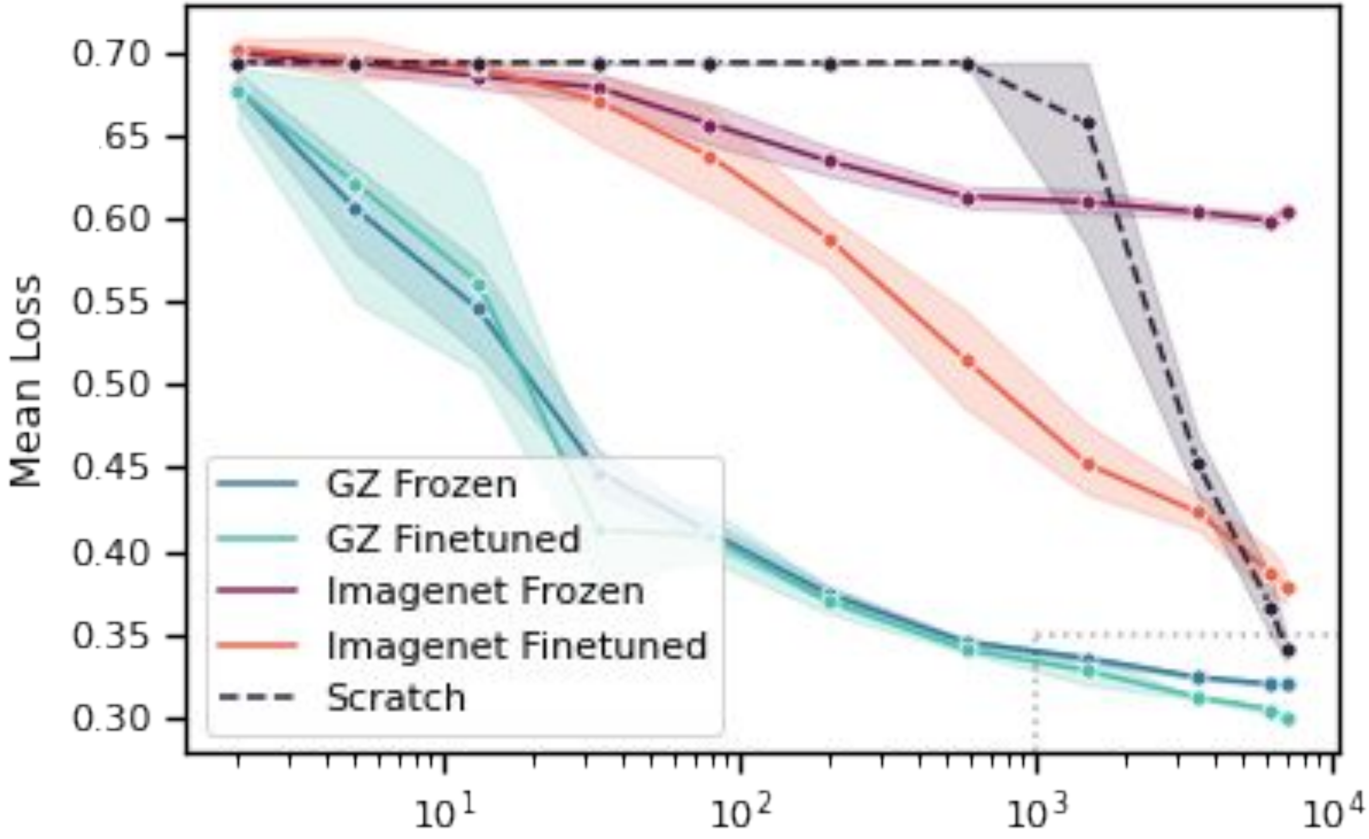
Start with a few hundred labelled examples

Finetune the representation for your problem



(illustrative figures only)

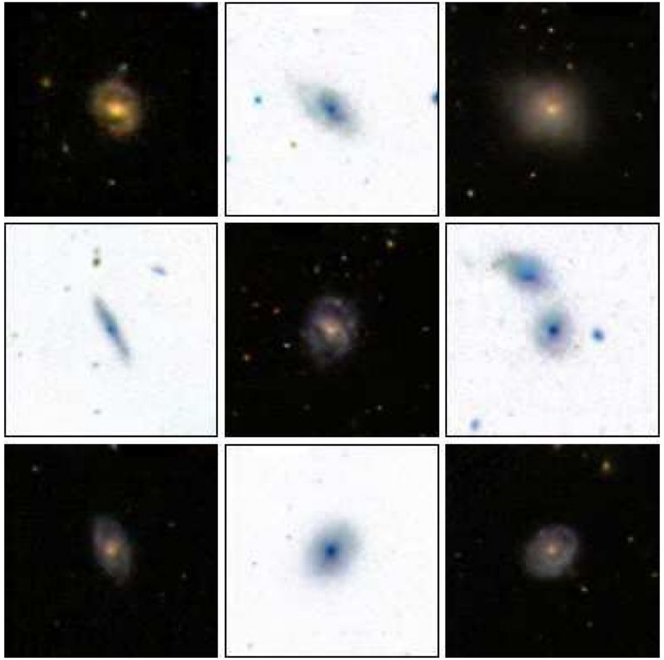
Mike Walmsley et al



Ring Galaxies to Train On

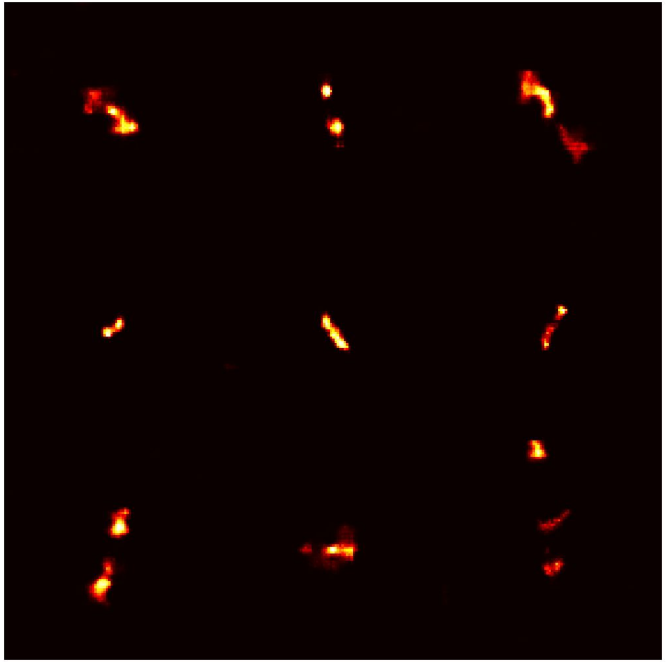
Mike Walmsley et al

Generative Learning



GAN-created SDSS galaxies.
Not real!

Fussell & Moews (2018)

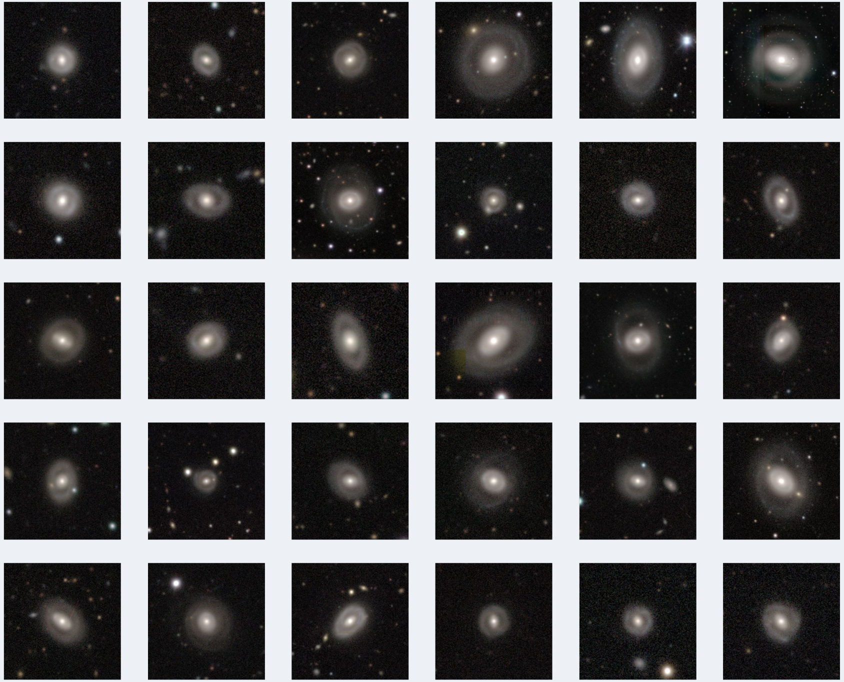


GAN-created radio galaxies.
Not real!

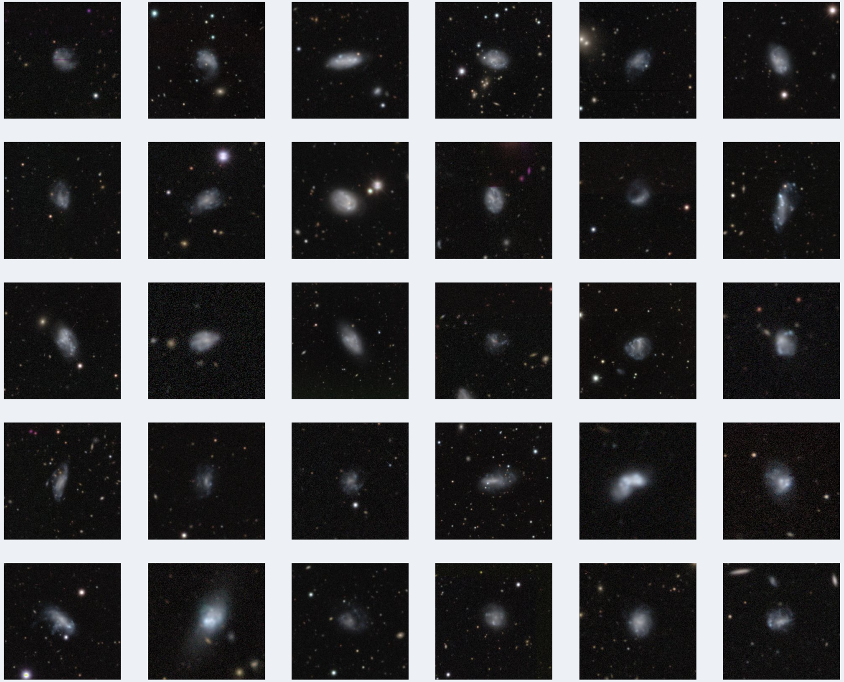
Inigo Val (in prep.)

Transfer Learning

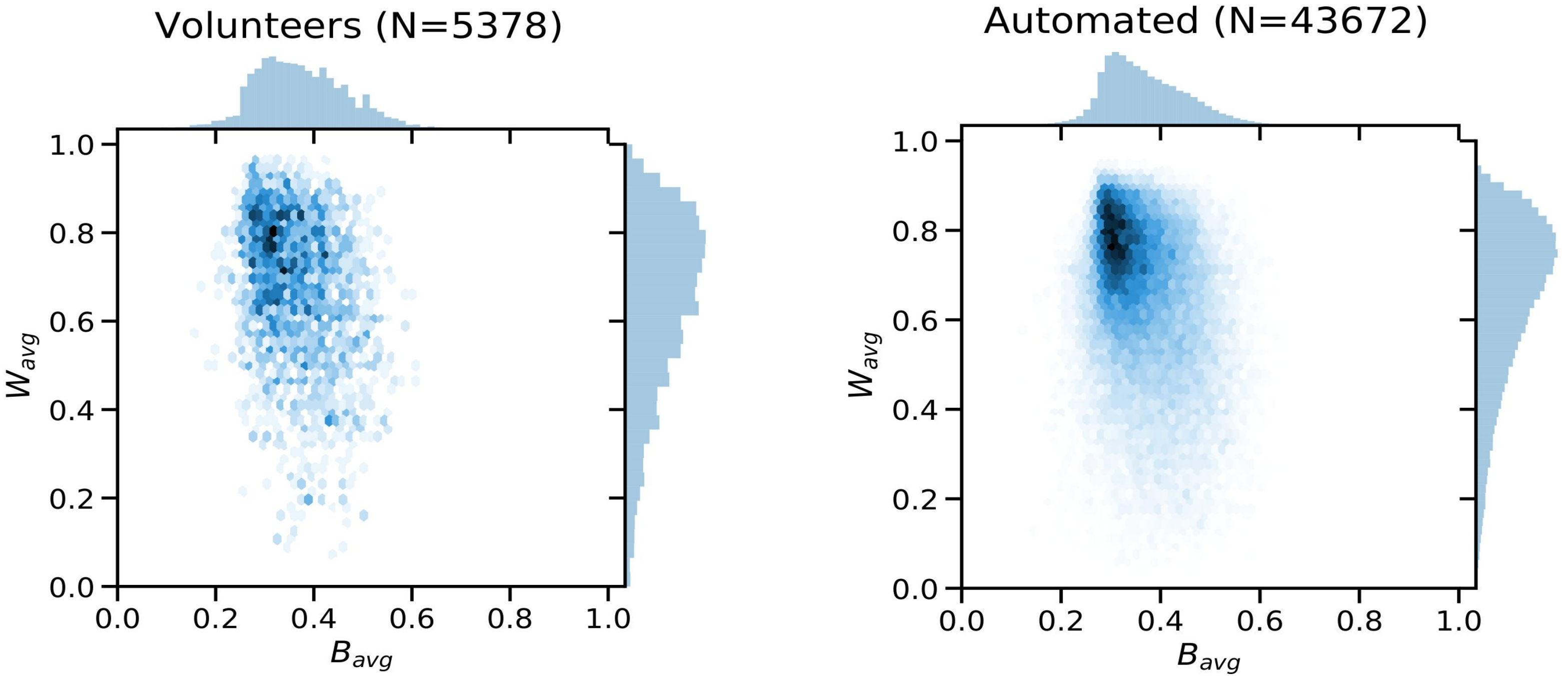
Automatic selection cuts: featured > 0.6, face-on > 0.7, has spiral arms < 0.5.
Training set: ***212 rings***



Max prob. "ring", validation set



Min prob. "ring", validation set



Winding angle vs. bulge size, measured by volunteers or deep learning

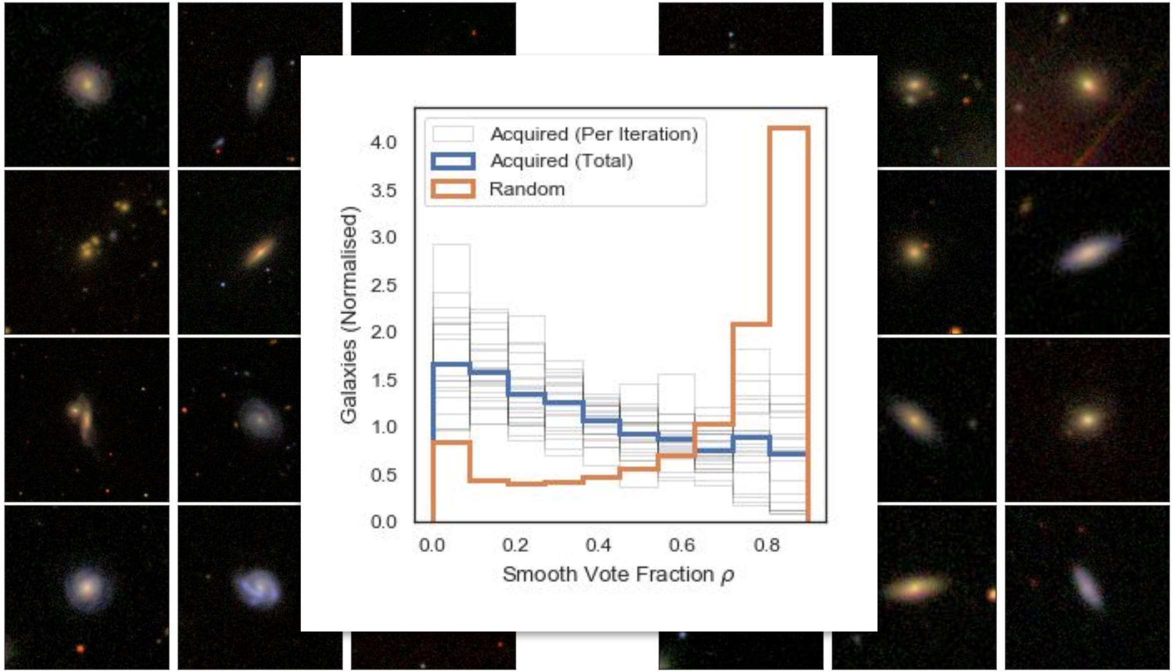
Dataset: zenodo.org/record/4196267

1. Need calibrated predictions from **noisy labels**
    a. Active learning

2. The **tasks keep changing**
    a. Multi-task learning

3. 99% of new data will remain **unlabelled**
    a. Hybrid supervised-contrastive learning

Selected Galaxies for “Smooth?”



High mutual information

Low mutual information

Mike Walmsley et al

Try it yourself: bit.ly/decals_viz

ZOONIVERSE

Choose Your Galaxies

Bar?

Answer

Strong

Posterior Mean

0.75 1.00

0.00 1.00

Has spiral arms?

Answer

Yes

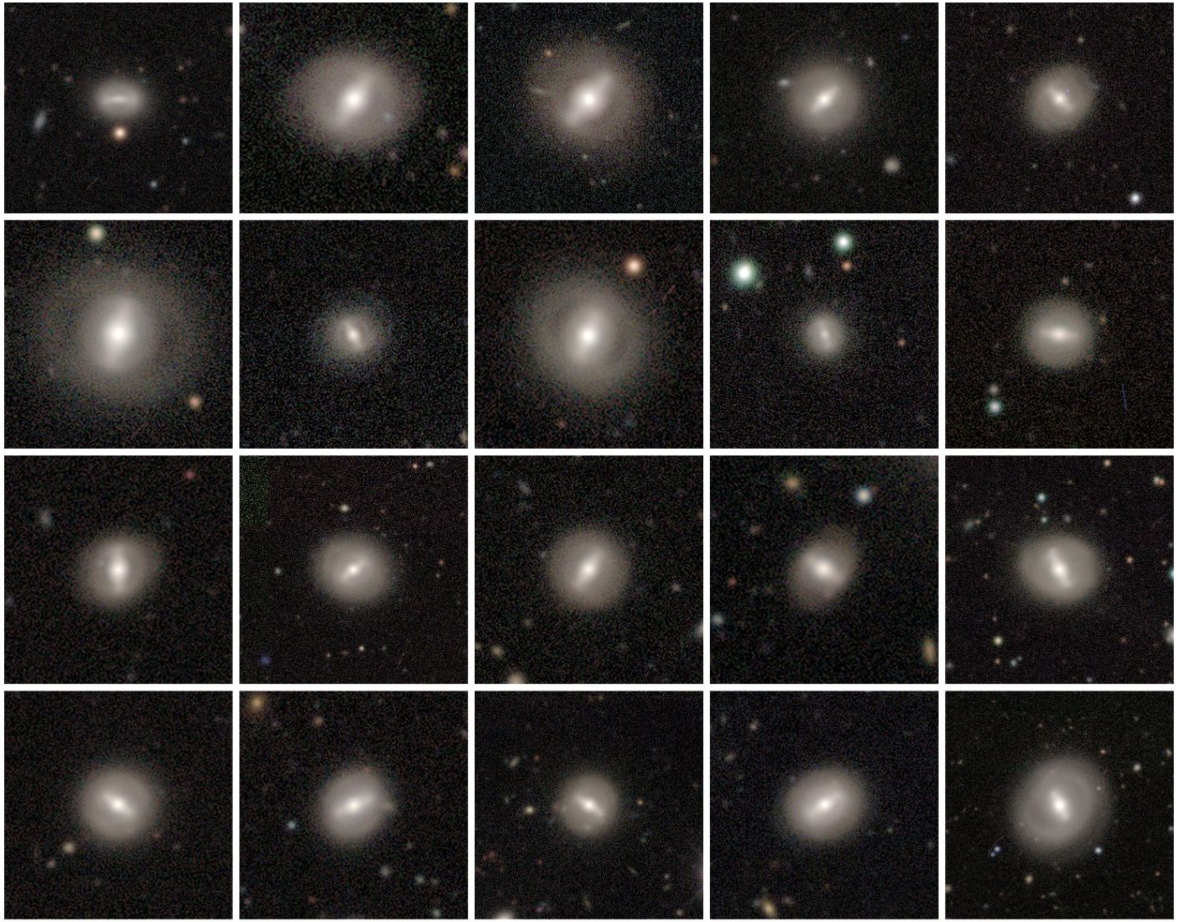
Posterior Mean

0.00 0.24

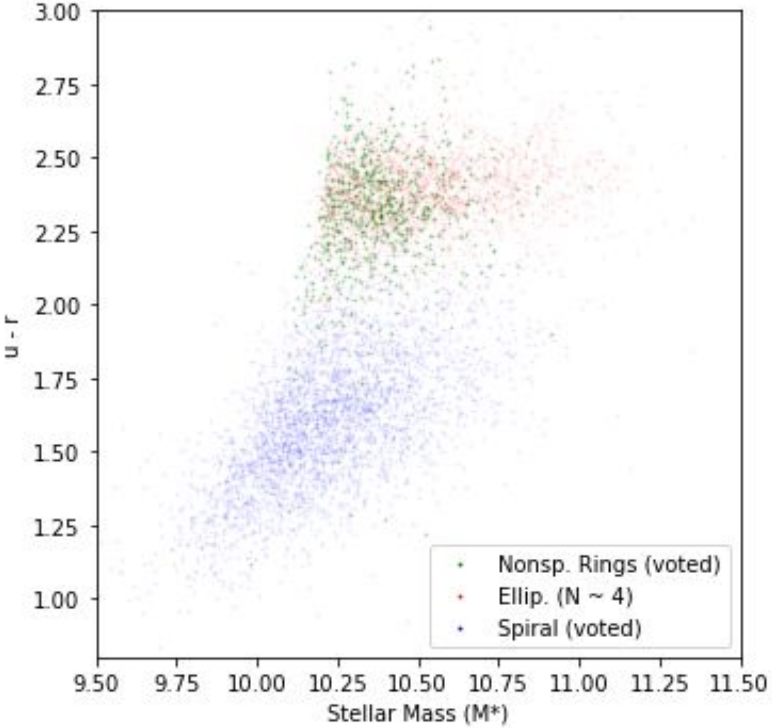
0.00 1.00

Spiral arm count?

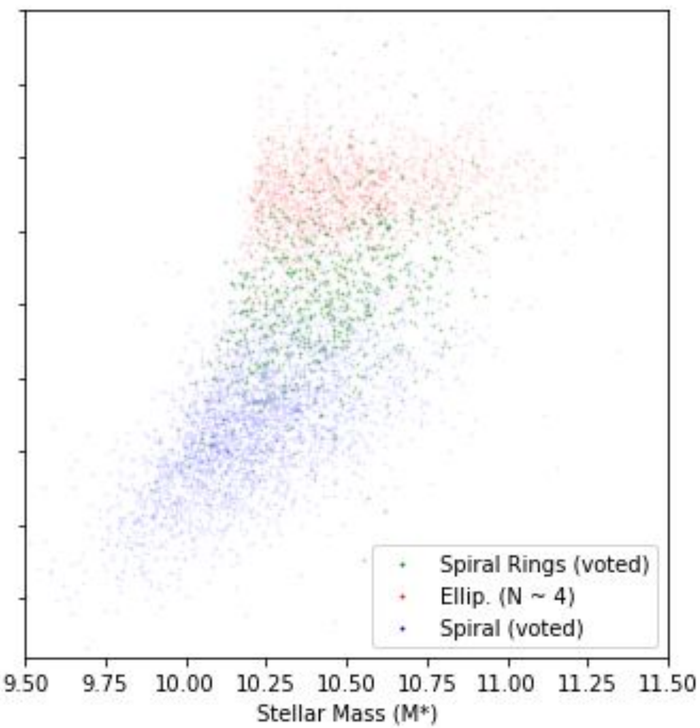
To use this filter, set "Has Spiral Arms = Yes"to > 0.5



ZOØNIVERSE

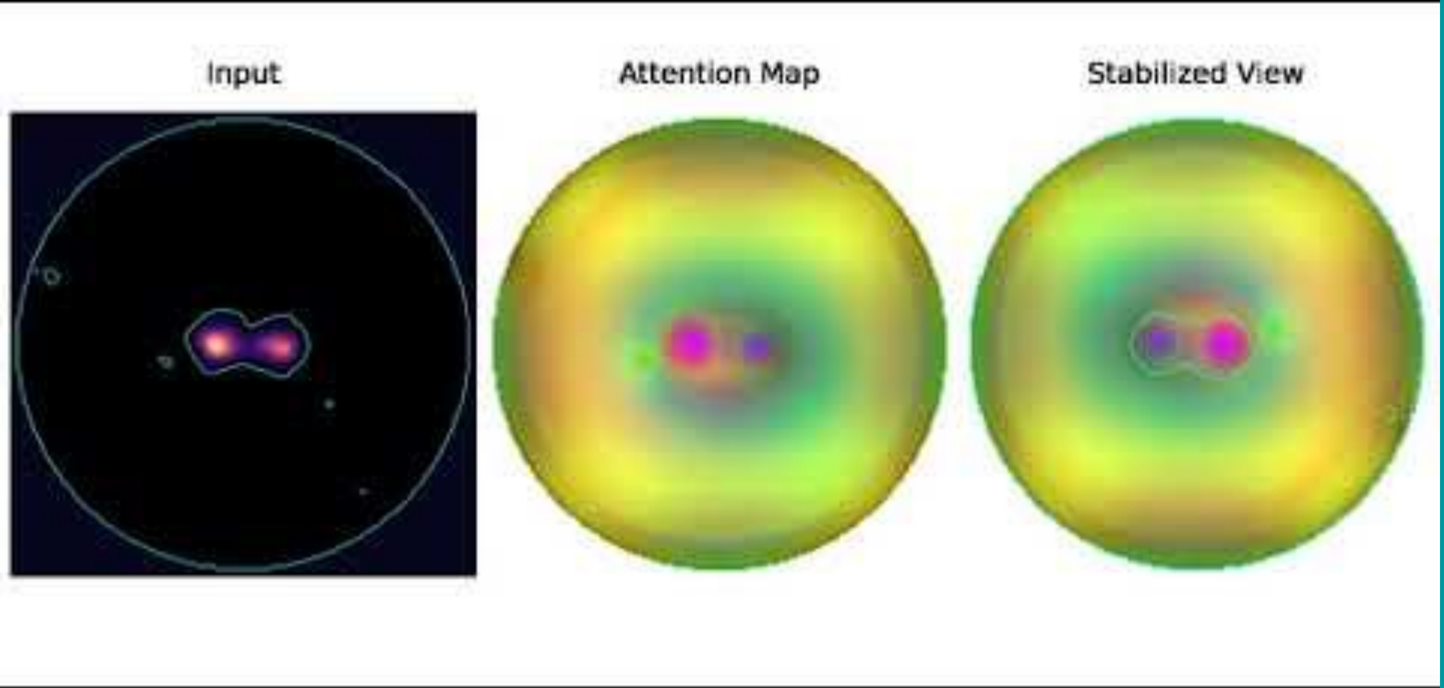


Non-spiral rings
(green)



Spiral rings
(also green)

Mike Walmsley et al



ZOØNIVERSE

Use Symmetry

Helps constrain model parameters

More constraints = less training data needed

Micah Bowles
micah.bowles@
postgrad.manchester
.ac.uk

bit.ly/decals_similarity

Similarity Search

by Mike Walmsley (@mike_walmsley_)

RA (deg)

184.6750

Dec (deg)

11.7309

Search

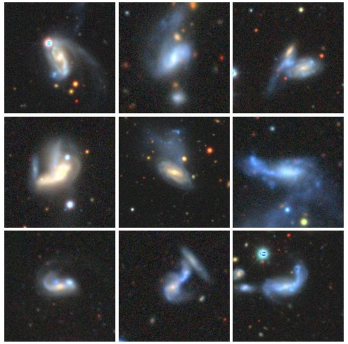
Important Notes +

Closest Galaxy

RA: 184.67458. Dec: 11.73206. Search Vizier

Similar Galaxies

Show table +



RA (deg)

190.975458

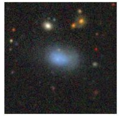
Dec (deg)

16.547361

Search

Important Notes

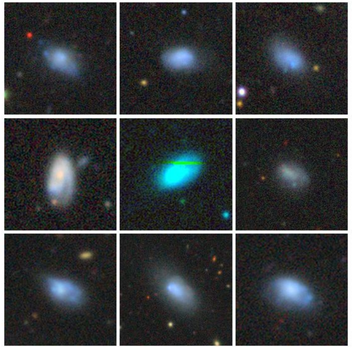
Closest Galaxy



RA: 190.97546. Dec: 16.54736. Search Vizier

Similar Galaxies

Show table



RA (deg)

124.417040

Dec (deg)

4.497872

Search

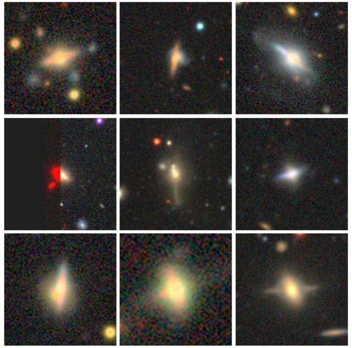
Important Notes +

Closest Galaxy

RA: 124.41704. Dec: 4.49787. Search Vizier

Similar Galaxies

Show table +



RA (deg)

180.431121

Dec (deg)

0.183135

Search

Important Notes +

Closest Galaxy

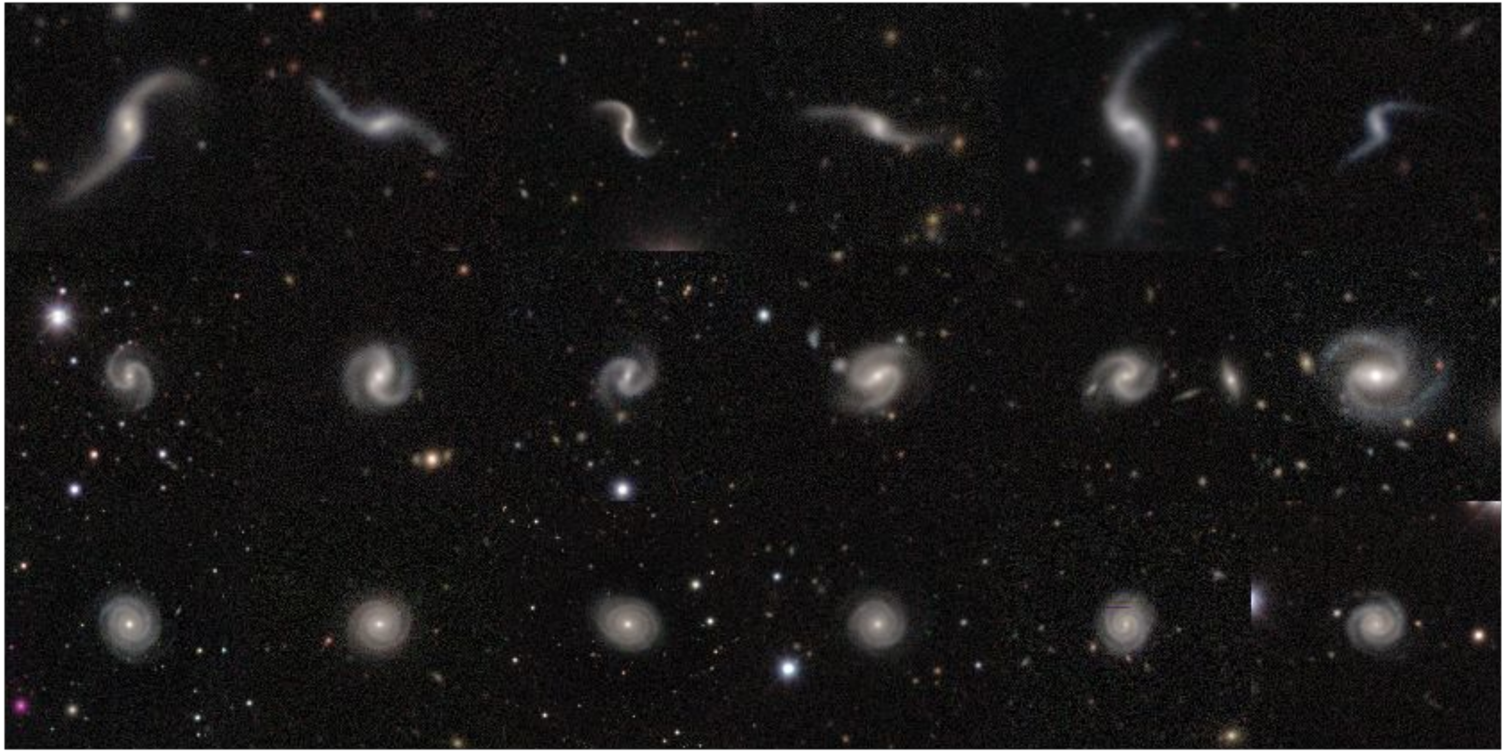
RA: 180.43112. Dec: 0.18313. Search Vizier

Similar Galaxies

Show table +

Mike Walmsley et al

ZOØNIVERSE



*Galaxies with **posteriors** for loose (upper), medium (centre) or tightly-wound (lower) spiral arms*

Mike Walmsley et al

(Supervised) Deep Learning in One Slide

Model

- Some function f(image)
- f has learnable parameters aka "weights"
- **Optimise** the weights for **max performance** on **training images**

Convolutional Neural Network

- Specific type of **black box** model
- **Millions** of weights

What if I get stuck in a local minima?

How do we define max performance? (aka the "loss function")

How do I know it learned what I want?

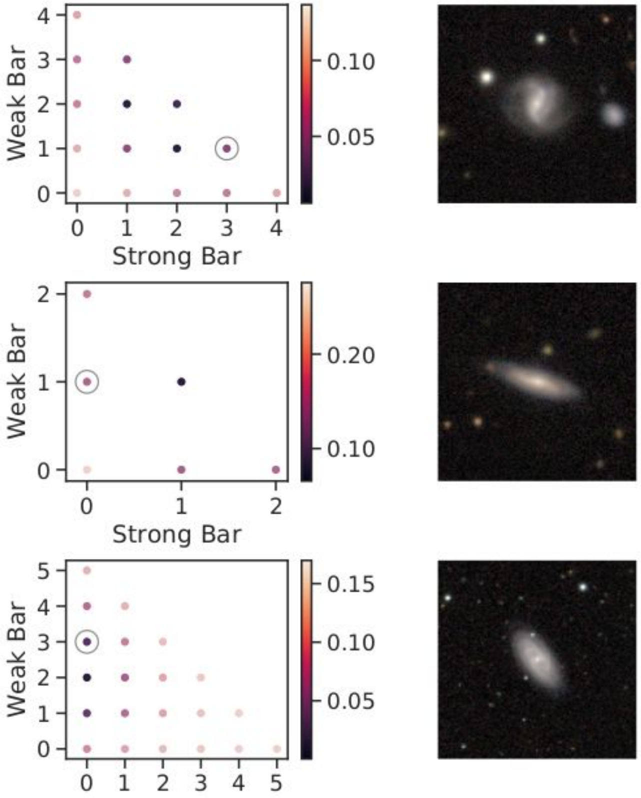
How do I avoid learning spurious correlations?

Multiple Answers

$$\mathcal{L} = \int \mathrm{Beta}(\rho|\alpha, \beta)\mathrm{Bin}(k|\rho, N)d\alpha d\beta$$

Add a few dimensions…

$$\mathcal{L}_q = \int \mathrm{Dirichlet}(\vec{\rho}|\vec{\alpha})\mathrm{Multi}(\vec{k}|\vec{\rho}, N)d\vec{\alpha}$$



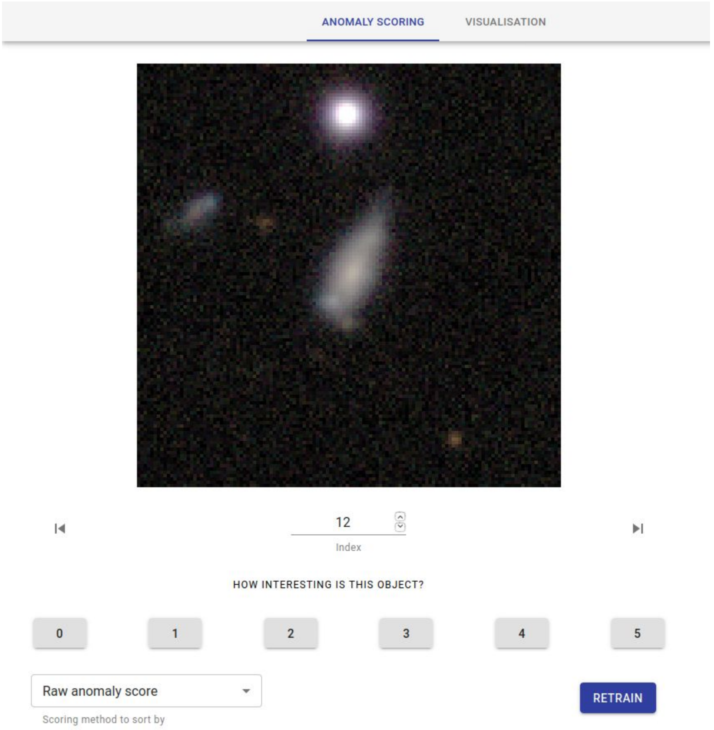
Mike Walmsley et al

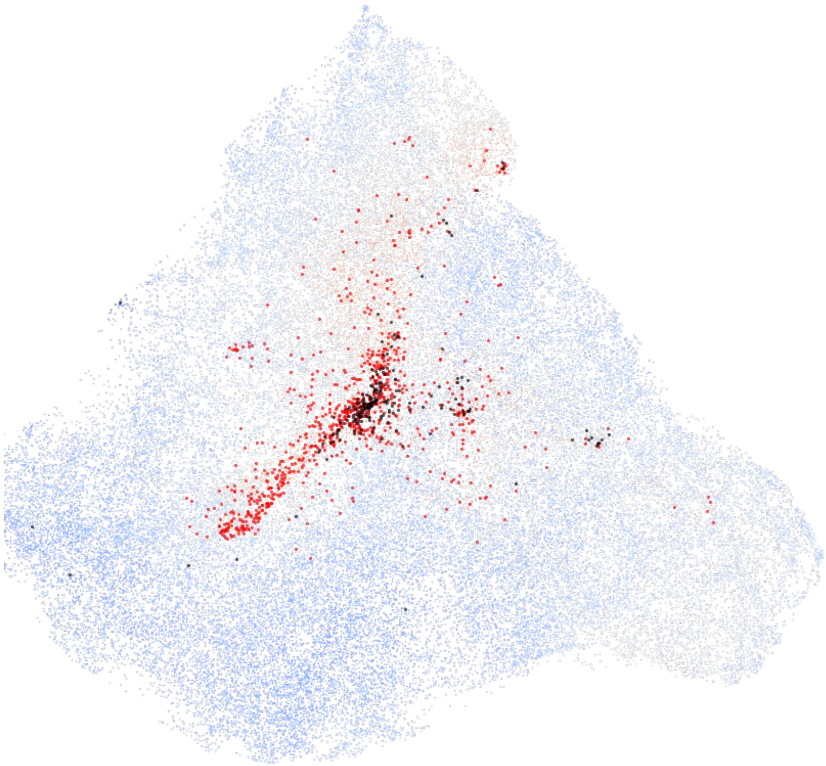
Anomaly Finding

ZOONIVERSE

CNN Representation +

Astronomaly UI (Lochner+ 21)

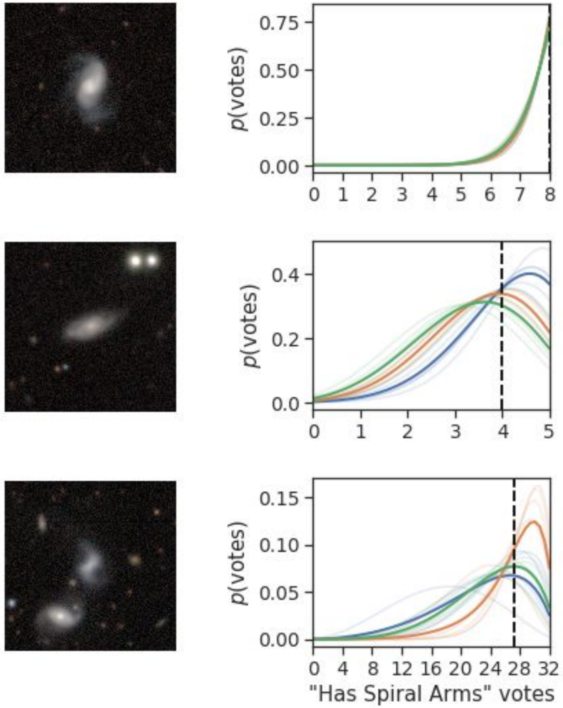




Mike Walmsley et al

Loss for All Questions

$$\mathcal{L}_q = \int \mathrm{Beta}(\rho|\alpha, \beta)\mathrm{Bin}(k|\rho, N)d\alpha d\beta$$

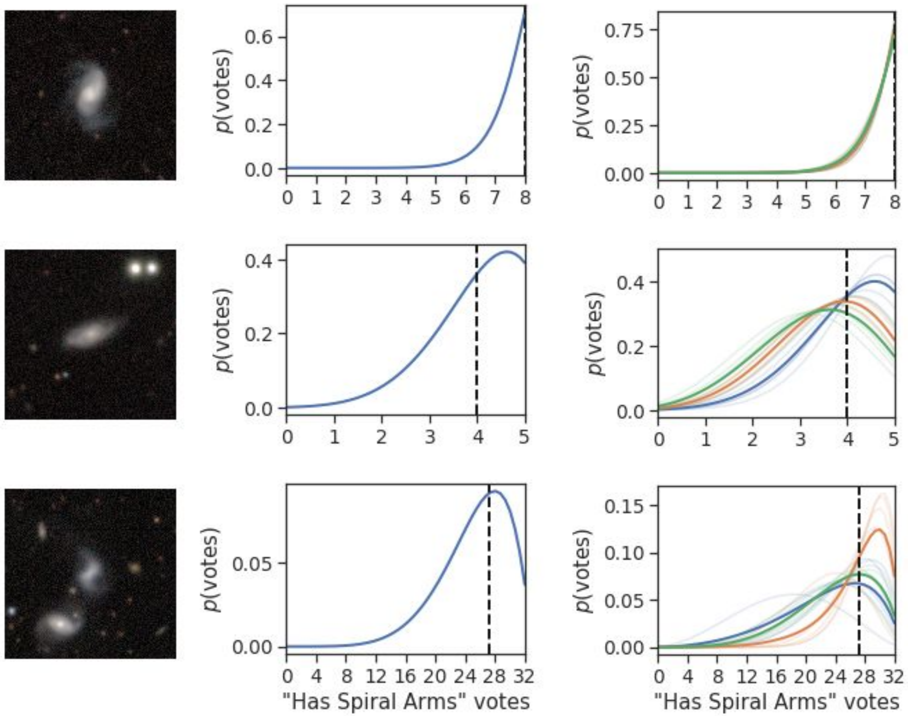


Posteriors for Votes

- Our CNN can learn from uncertain labels and make probabilistic predictions $p(k|w)$
- Marginalising over weights (BCNN) lets us predict votes *over all CNN we might have trained*

$$p(k|D) = \int p(k|w)\ p(w|D)dw$$

Train many models
Dropout on each



1 Model   15 "Models" (BCNN)

For more, see arxiv:2102.08414

Deep Learning in One Slide

Machine Learning Model

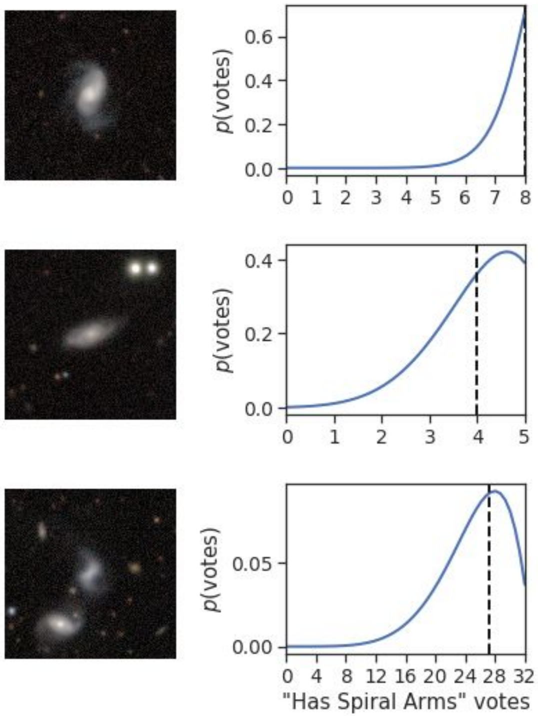
- Some function f(image)
- f has learnable parameters aka weights
- **Optimise** the weights for **max performance** on training images

Convolutional Neural Network ("CNN")

- Specific type of **black box** model
- **Millions** of weights ("deep")

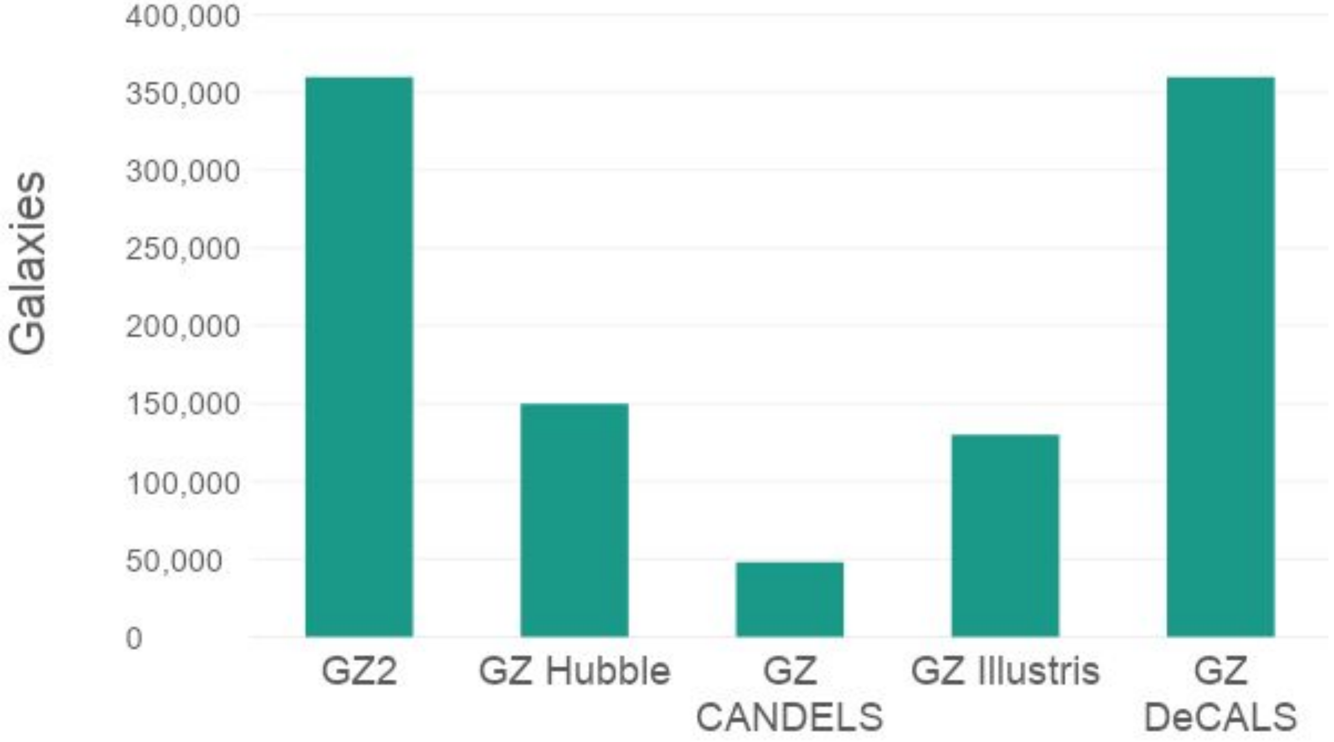
Posteriors for Votes

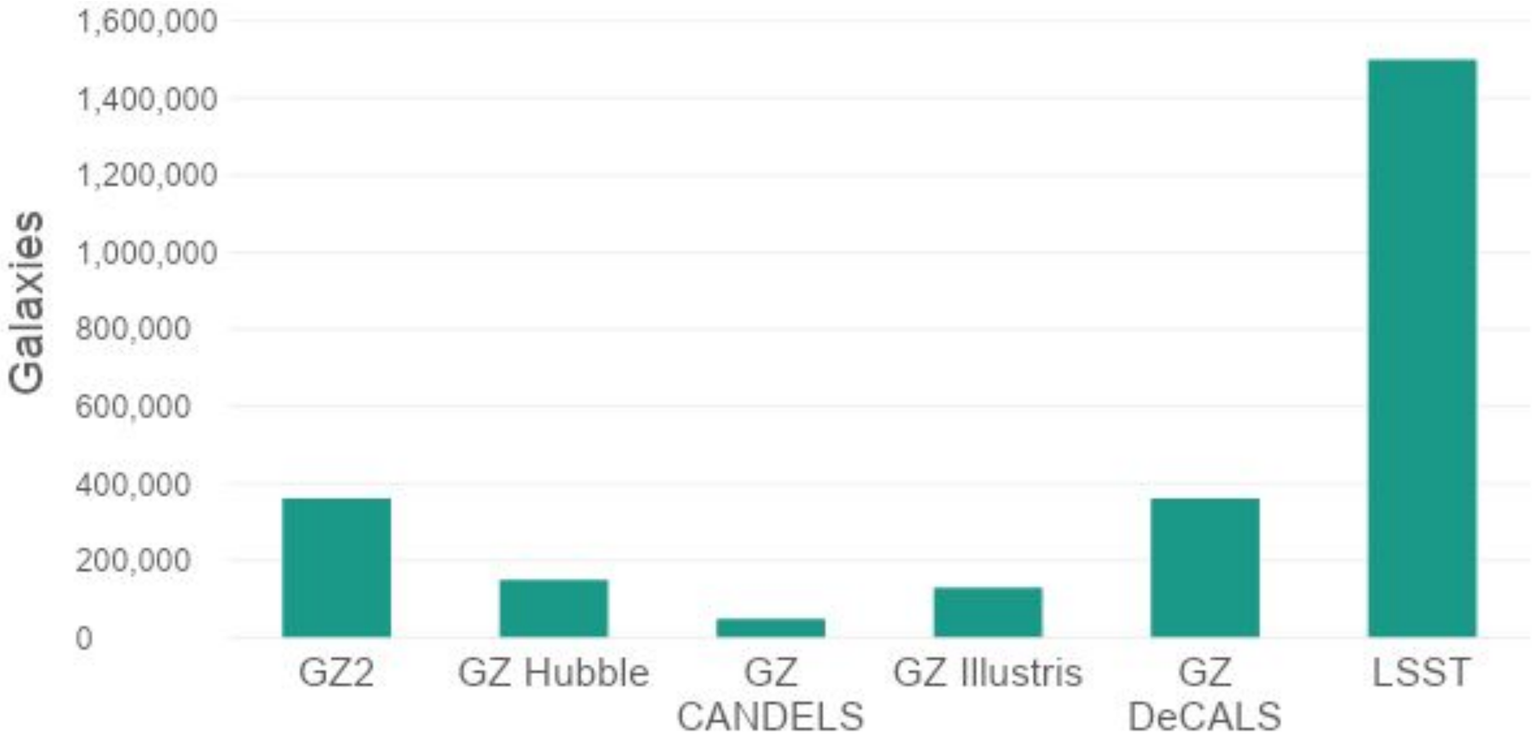
- Our CNN can learn from uncertain labels and make probabilistic predictions $p(k|w)$



1 Model

For more, see arxiv:2102.08414


Galaxies
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
GZ2
GZ Hubble
GZ CANDELS
GZ Illustris
GZ DeCALS



Mike Walmsley et al

Mike Walmsley et al

Performance

~ 99% accurate on every question for galaxies where the volunteers are confident

Question	Count	Accuracy	Precision	Recall	F1
Smooth Or Featured	3495	0.9997	0.9997	0.9997	0.9997
Disk Edge On	3480	0.9980	0.9980	0.9980	0.9980
Has Spiral Arms	2024	0.9921	0.9933	0.9921	0.9924
Bar	543	0.9945	0.9964	0.9945	0.9951
Bulge Size	237	1.0000	1.0000	1.0000	1.0000
How Rounded	3774	0.9968	0.9968	0.9968	0.9968
Edge On Bulge	258	0.9961	0.9961	0.9961	0.9961
Spiral Winding	213	0.9906	1.0000	0.9906	0.9953
Spiral Arm Count	659	0.9863	0.9891	0.9863	0.9871
Merging	3108	0.9987	0.9987	0.9987	0.9987

Classification metrics on confident galaxies

Probabilistic to Bayesian CNN

Galaxy x
CNN weights w
Training data D_{train}
CNN output $f^{w}(x)$
Dropout dist. q_{θ}^{*}
Forward pass t of T

What about the models we might have trained, but didn't?

$$p(y = c \mid x, D_{train}) = \int f^{w}(x)\, p(w \mid D_{train})\, dw$$

Unknown!

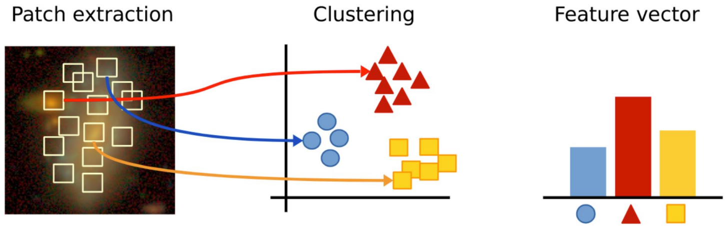
Approximate $p(w \mid D_{train})$ with Dropout

$$\approx \int q_{\theta}^{*}(w)\, dw$$

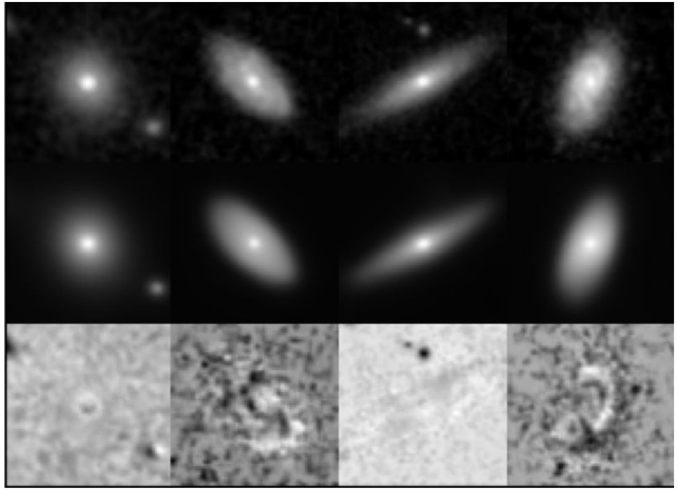
$$\approx \frac{1}{T} \sum_{t=1}^{T} f^{w_t}(x)$$

See Y. Gal et al (2016)

No Labels Needed?



Clustering image patches
Martin (2020)
See also Hocking (2017)
See also Self-Organising Maps

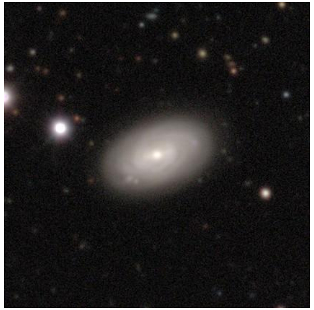


Learning to reconstruct images
Spindler (2020)
See also Gheller (2022)

ZOONIVERSE

GZ2
210k, $z < 0.15$

Willett (2013)



GZ DECaLS
230k, $z < 0.15$

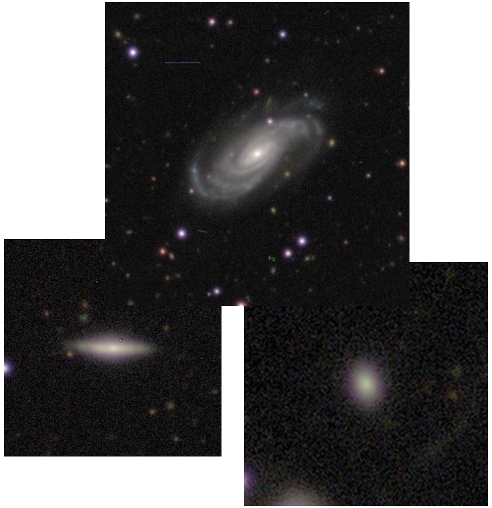
Walmsley (2022)

GZ Hubble
106k, $z < 1$

GZ CANDELS
50k, $1 < z < 3$

Willett (2017)
Simmons (2017)

and



GZ DESI
375k, $z < 0.3$

plus **1.3m unlabelled**

Walmsley (in prep.)

1. Build a Bayesian Galaxy Zoo model

2. Mess around

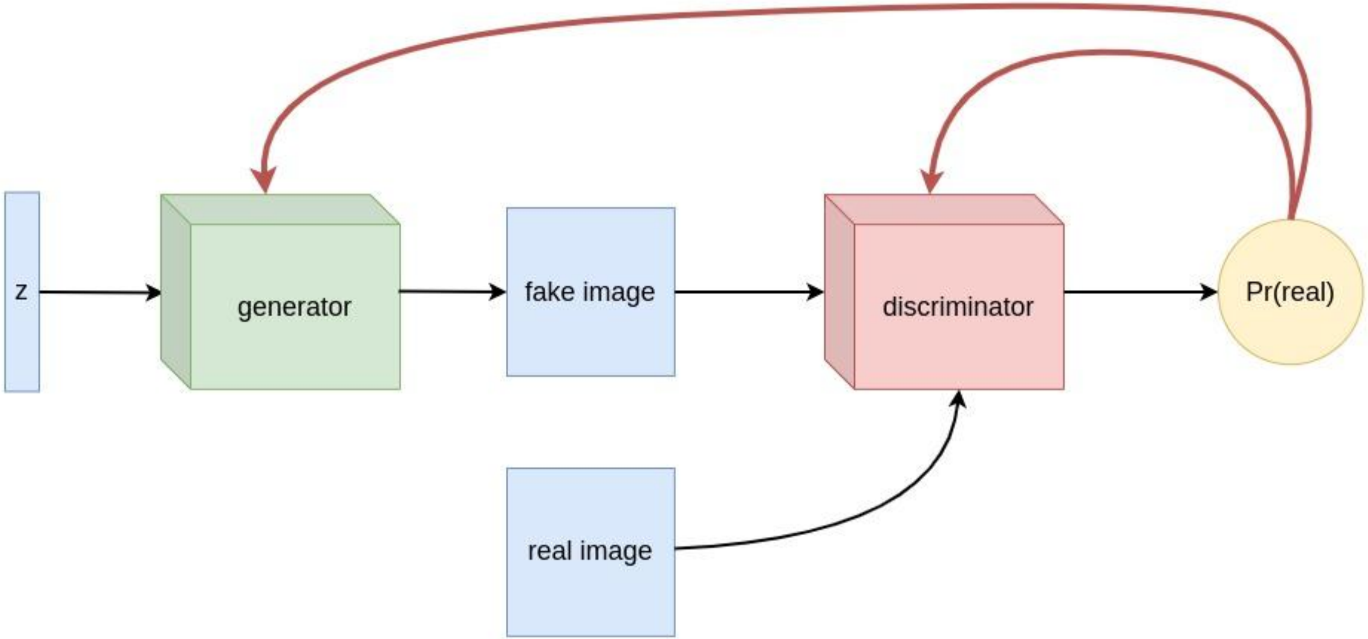


Diagram of a generative adversarial network (GAN)

- Generative adversarial networks (GANs) can generate semantically different yet realistic looking data.
- We can create pseudo-infinite number of realistic images by feeding in a different random vector.
- All we need to do is feed in the data we wish to imitate - no need for labels or physical parameters.