ZOØNIVERSE

Detailed Morphology with Rubin

Mike Walmsley (he/him), Anna Scaife, and the Galaxy Zoo team

University of Manchester (soon Toronto)

zooniverse.org

@mike_walmsley_

Live Demo

bit.ly/decals_viz

ZOØNIVERSE

Choose Your Galaxies

Bar?

Answer

Strong

Posterior Mean

0.77 1.00

0.00 1.00

Has spiral arms?

Answer

Yes

Posterior Mean

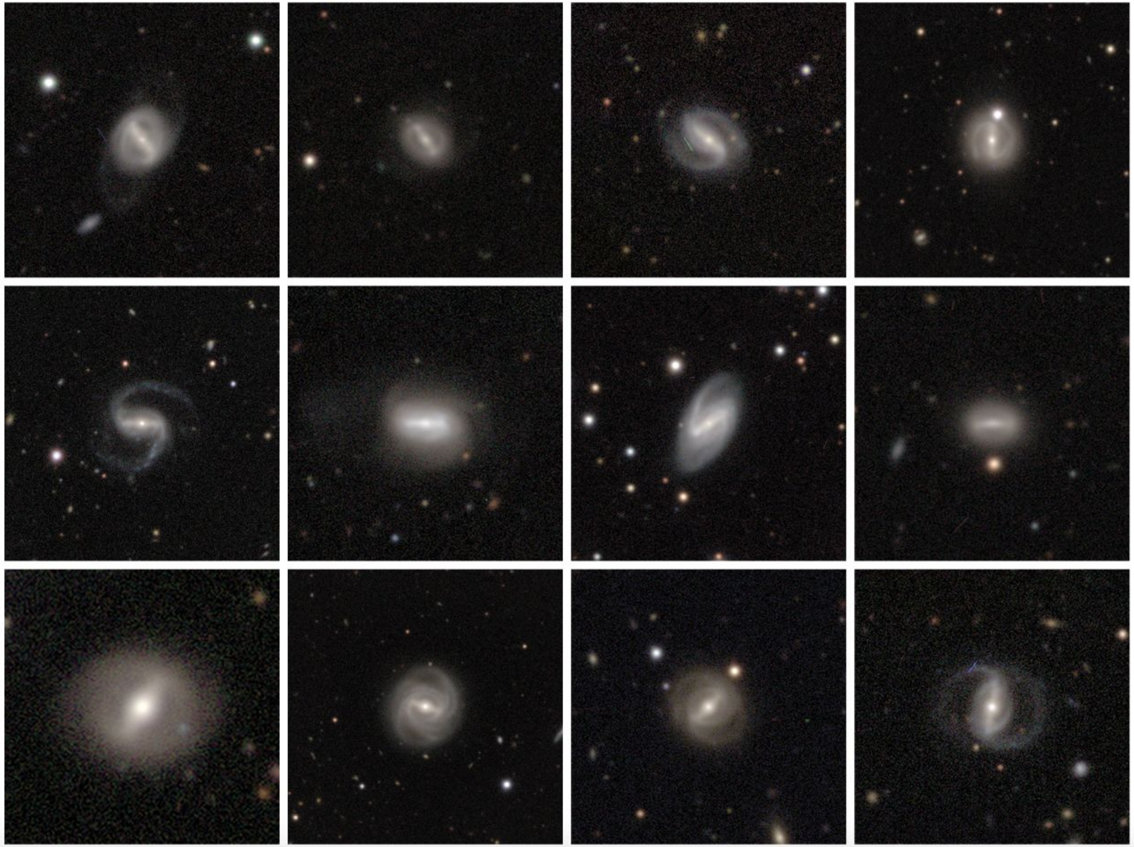
0.00 1.00

0.00 1.00

Spiral arm count?

Launchpad

247 of 253,286 galaxies match your criteria.



ZOØNIVERSE

Answer

Yes

Posterior Mean

0.74 1.00

0.00 1.00

Spiral arm count?

Answer

2

Posterior Mean

0.72 1.00

0.00 1.00

Spiral winding?

Answer

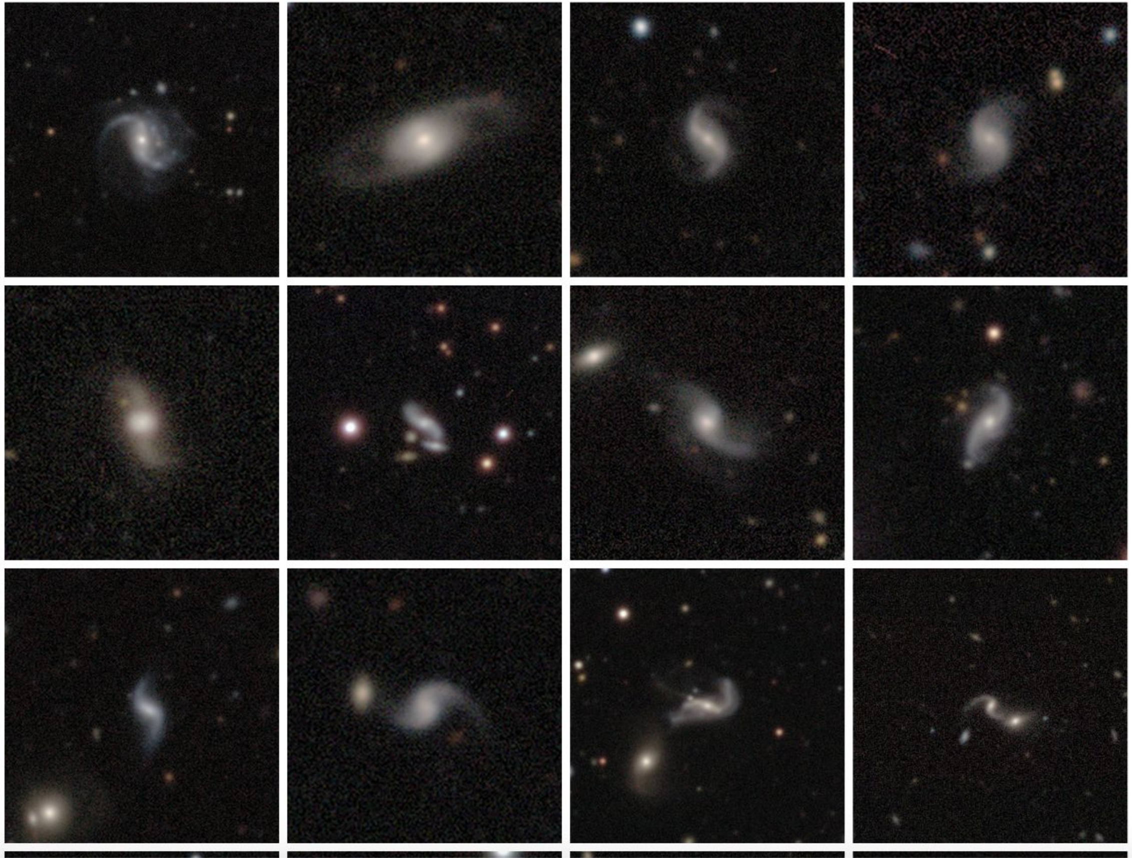
Loose

Posterior Mean

0.70 1.00

0.00 1.00

1,055 of 253,286 galaxies match your criteria.



Unauthorized Roadmap to Rubin Morphologies

GZ DECaLS →	GZ DESI →	GZ HSC
		a.k.a "Cosmic Dawn"
Train models to answer GZ questions	Train really good models to answer every GZ question	Adapt models to a new survey in a few months
Experiment with active learning	First model-only catalog	Simple, effective active learning deployed

Plus ***many*** *other projects! e.g.*

- *Clump Scout to locate starforming clumps within galaxies (Adams, Dickinson)*
- *The Merger Challenge competition to benchmark merger classifiers (Margalef, Wang)*
- *Building models for low surface brightness tidal features (Gordon, Ferguson, Mann)*

Simple Active Learning for HSC

Each week:

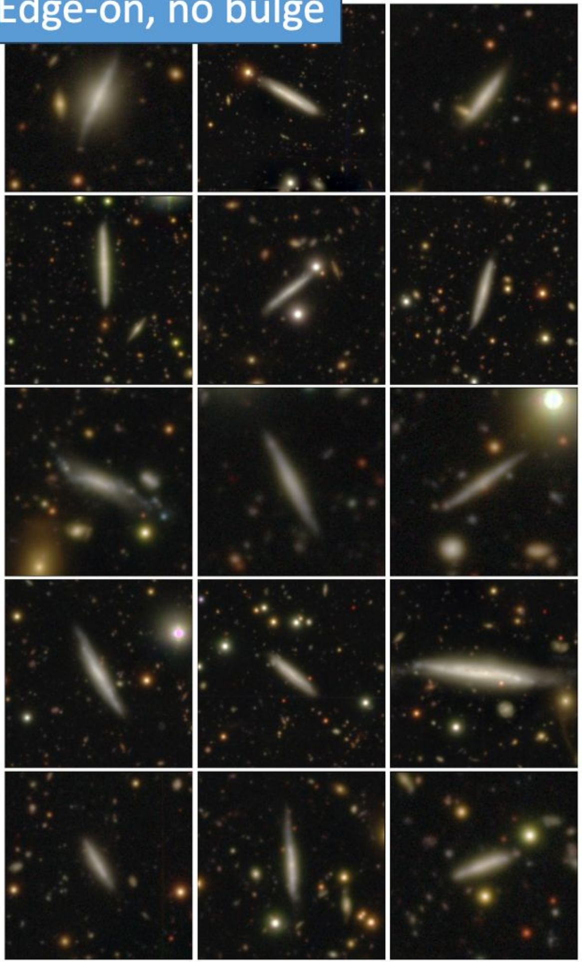
- Predict morphology of every galaxy
- Galaxies not confidently smooth are shown to volunteers
- Retrain model on all (new + existing) volunteer labels
- Check new model is as good or better

Running serverside

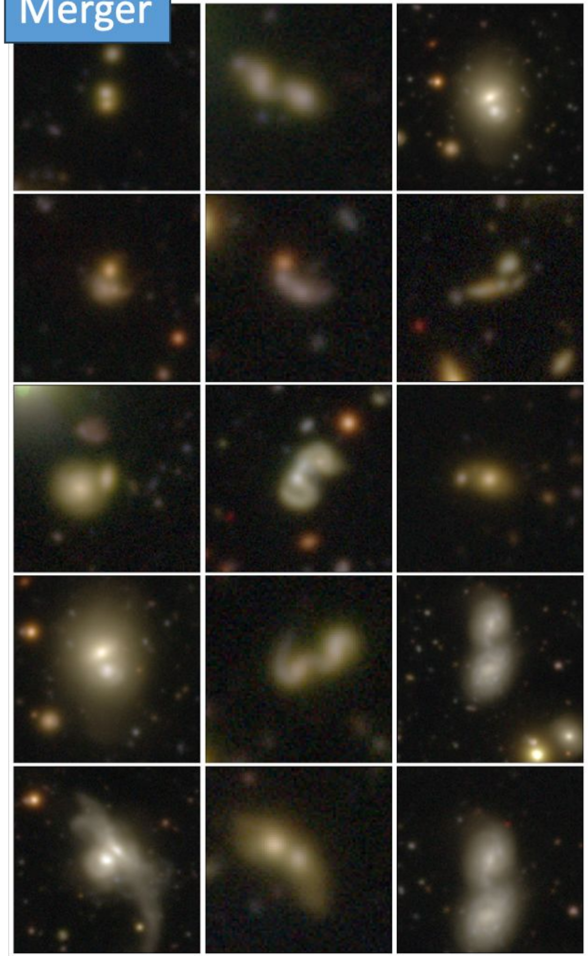
Courtesy James Pearson - lead researcher on Galaxy Zoo Cosmic Dawn

ZOONIVERSE

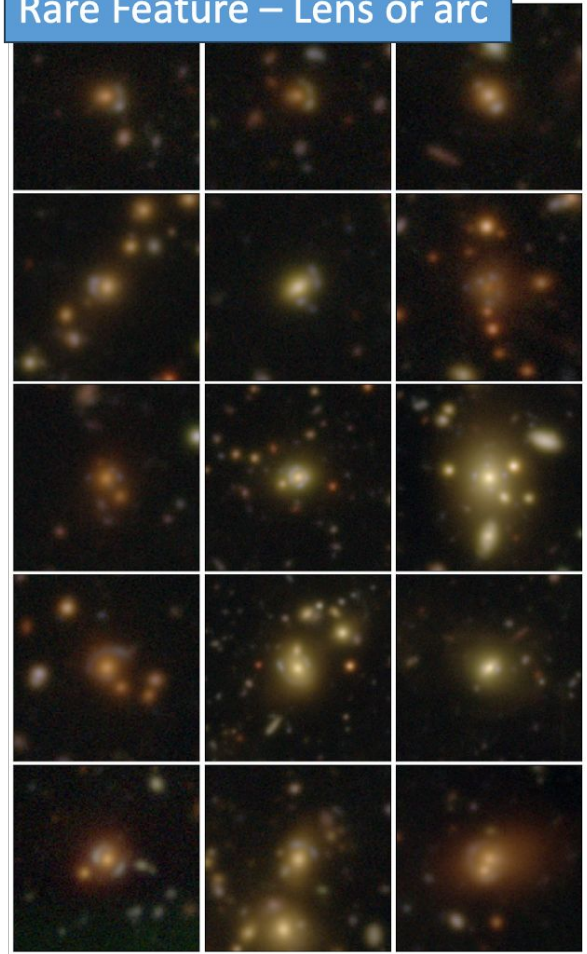
Edge-on, no bulge



Merger



Rare Feature – Lens or arc



Live Demo #2

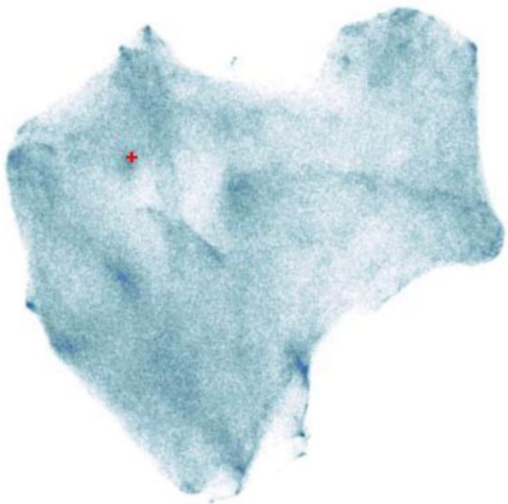
bit.ly/gz-explorer

Move Around

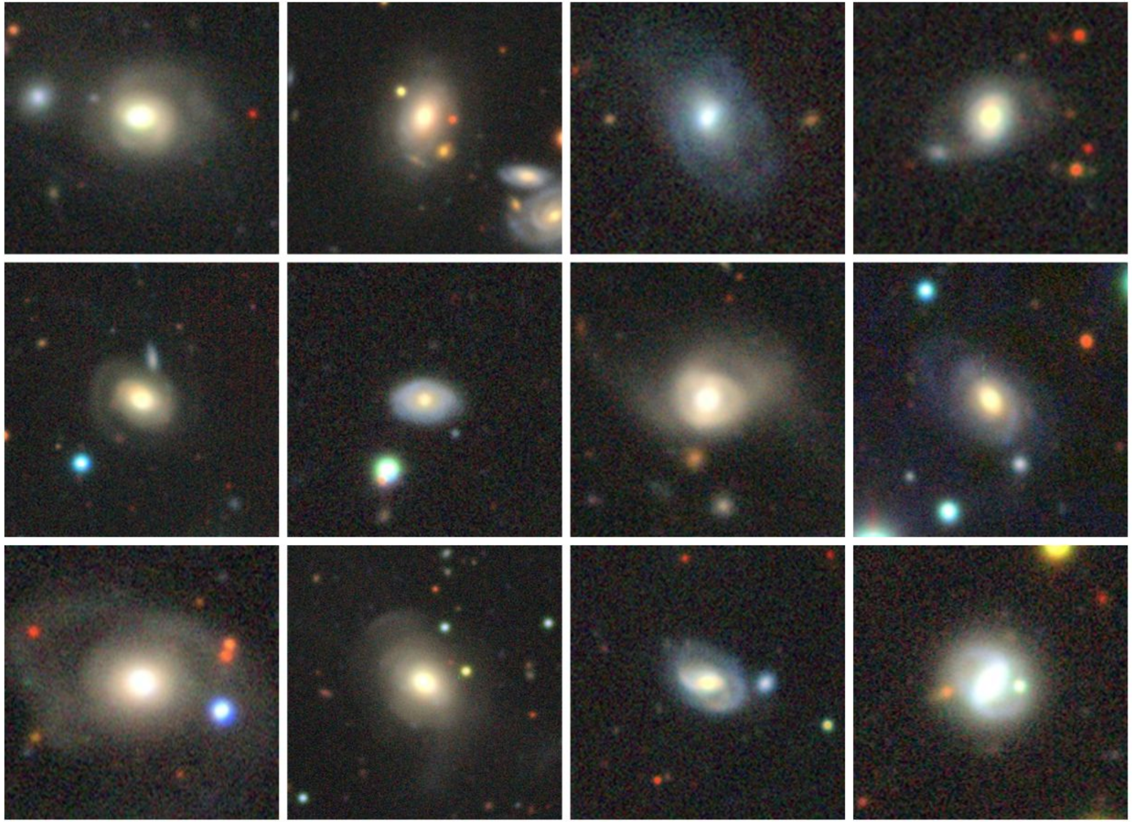
Click anywhere on the latent space to view galaxies.

Select Reduction

Featured v2



Location: (-0.109, 6.410)



Download CSV of the 1000 galaxies closest to your search

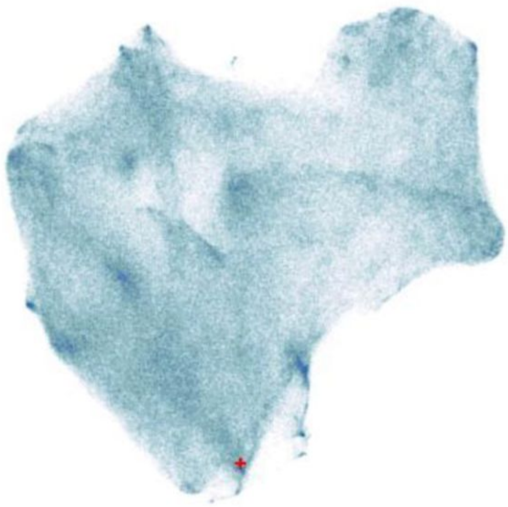
ZOONIVERSE

Move Around

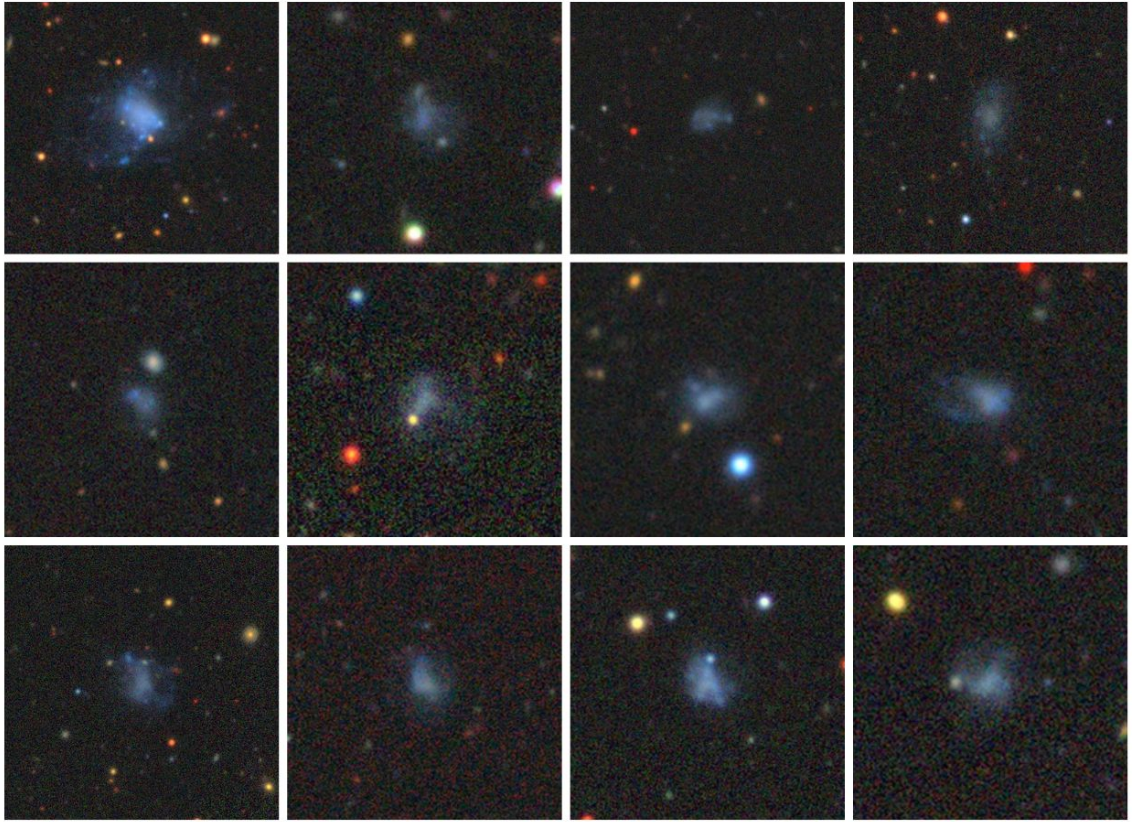
Click anywhere on the latent space to view galaxies.

Select Reduction

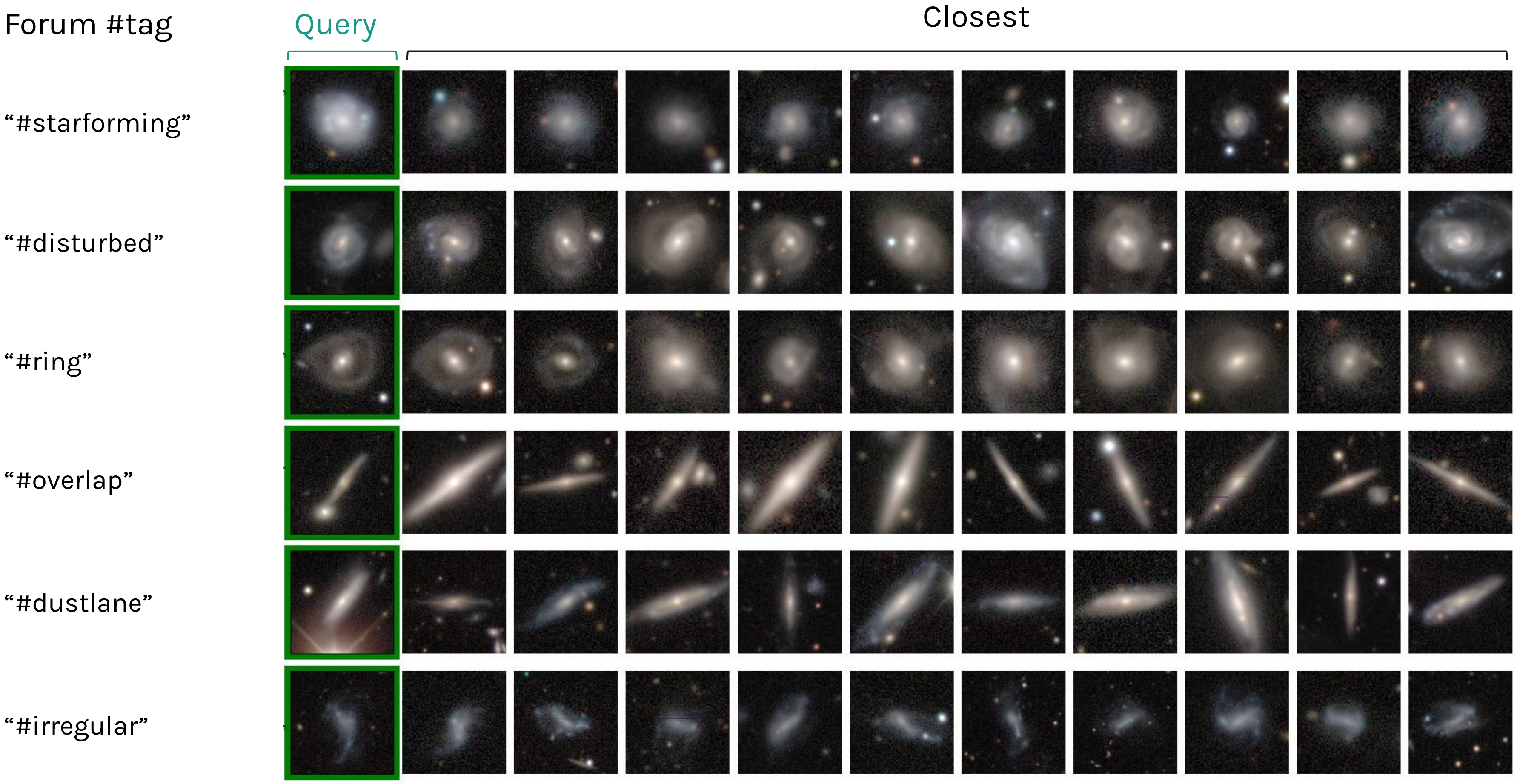
Featured v2



Location: (2.809, 0.812)



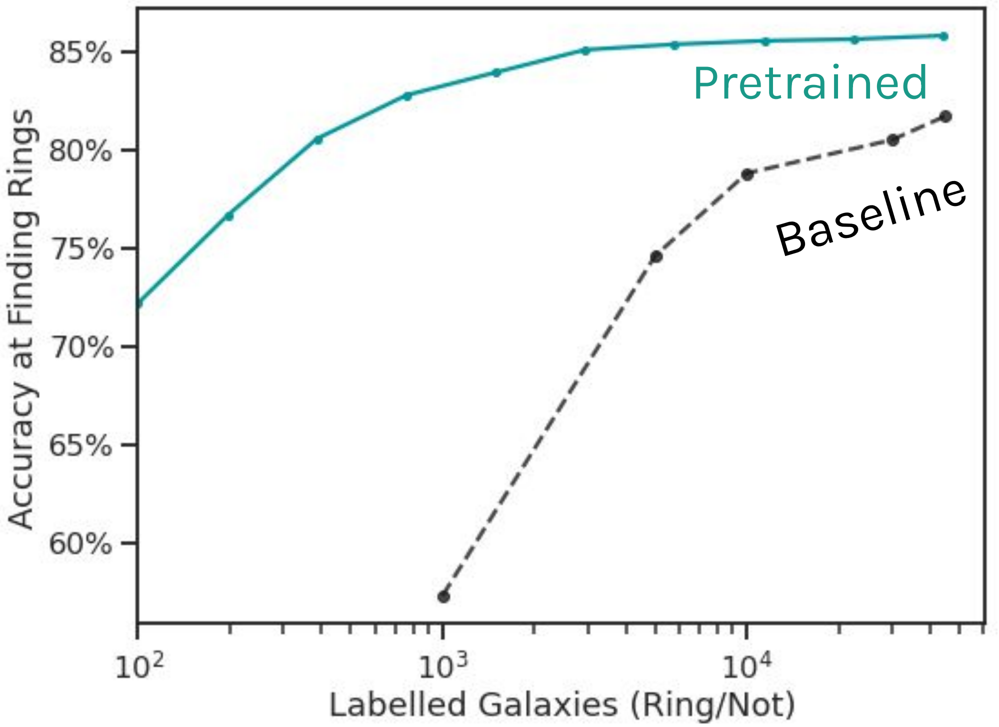
Download CSV of the 1000 galaxies closest to your search



Mike Walmsley et al

Adapt to your problem #1

ZOONIVERSE



Pretraining on Galaxy Zoo allows good performance with just a few hundred labels

Mike Walmsley et al

GZ Rings

Fine-tune Zoobot to find rings

40,000 ringed galaxies in DESI

6x more than all previous work combined

...plus another 39,700 or so

Mike Walmsley et al

```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***

ZOØNIVERSE

Unauthorized Roadmap to Rubin Morphologies

GZ DECaLS → GZ DESI → GZ HSC - - - - - →

a.k.a "Cosmic Dawn"

GZ DECaLS	GZ DESI	GZ HSC
Train models to answer GZ questions	Train really good models to answer every GZ question	Adapt models to a new survey in a few months
Experiment with active learning	First model-only catalog	Simple, effective active learning deployed

*Plus **many** other projects! e.g.*

- *Clump Scout to locate starforming clumps within galaxies (Adams, Dickinson)*
- *The Merger Challenge competition to benchmark merger classifiers (Margalef, Wang)*
- *Building models for low surface brightness tidal features (Gordon, Ferguson, Mann)*

Unauthorized Roadmap to Rubin Morphologies

GZ Euclid → GZ Rubin

GZ Euclid

Run everything at extreme scale

Calibrate for redshift visibility

Calibrate for volunteer individuality

Push ML boundaries

GZ Rubin

Watch this space

Vastly helped by Rubin’s infrastructure, Zooniverse connection

mwalmsley.dev/postdoc

GZ DECaLS:

arxiv: 2102.08414

zenodo: 4573248

Representations:

arxiv: 2110.12735

Large-Scale Learning

arxiv: 2206.11927

github.com/mwalmsley

/zoobot

/galaxy-datasets

bit.ly

/decals_viz

/galaxy_explorer

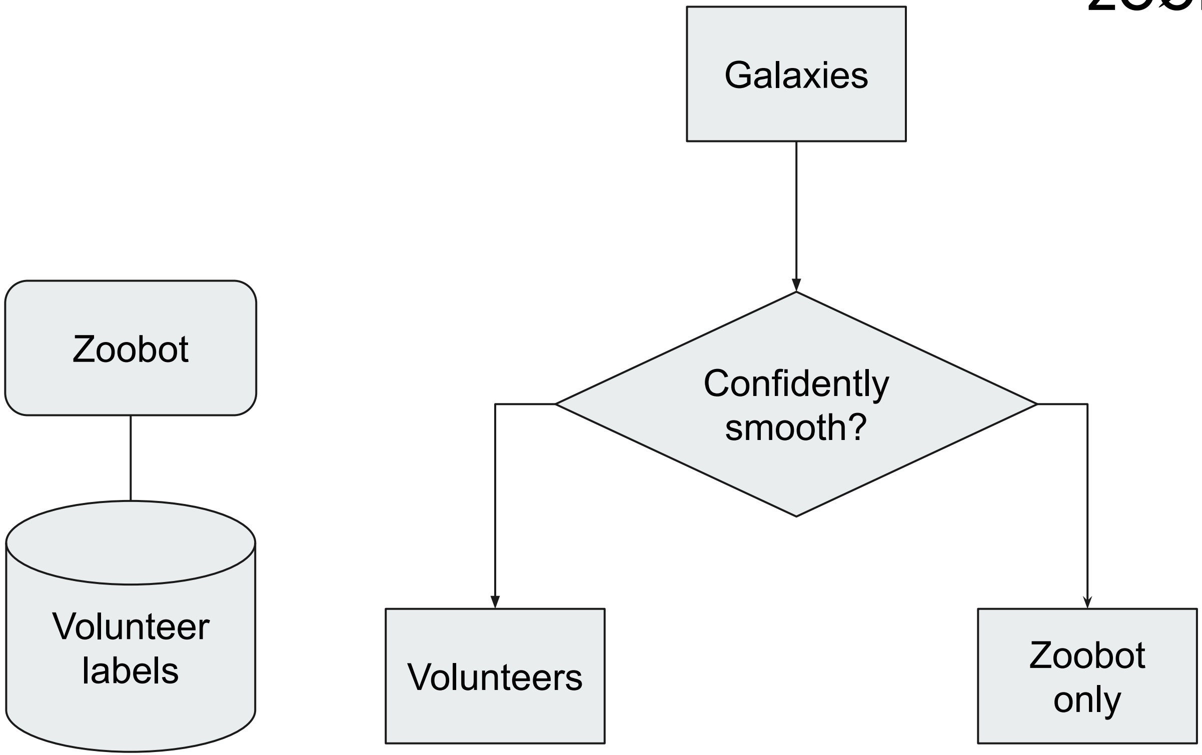
ZOONIVERSE

Technical

- How do we run our code on Rubin data?
- Benchmarking: what works, what doesn’t?
- Controlling for impact of redshift on detections
- Raising the bar on what ML can do

Human

- Link to Euclid? Euclid Q1 release is first. Joint DDP?
- How does this fit in with Rubin’s citizen science plans?



Active learning diagram

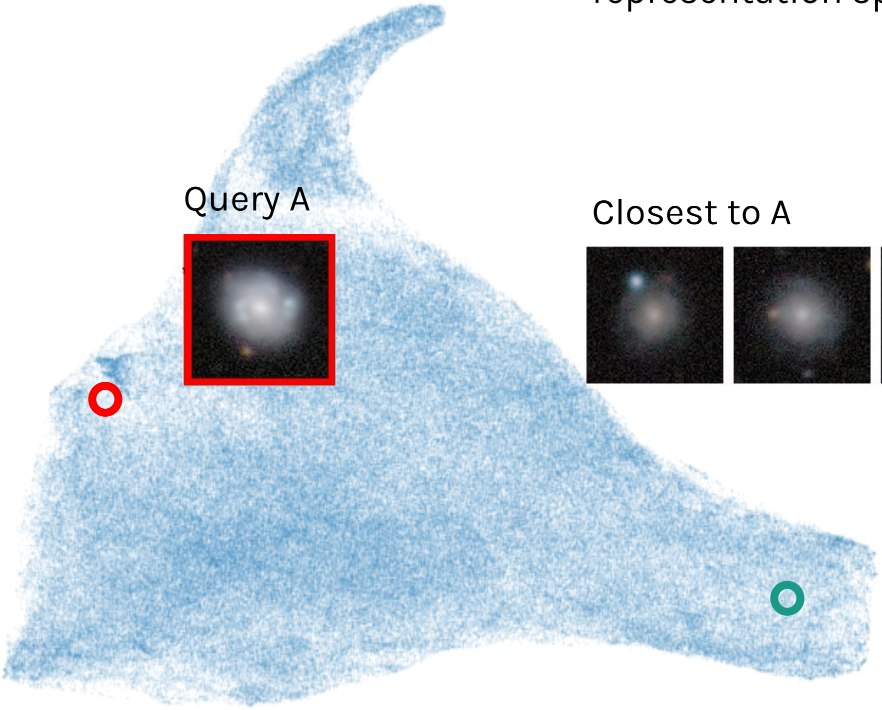
Similarity Search

ZOONIVERSE

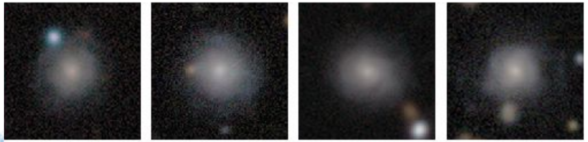
Pick a galaxy…

…show the closest galaxies in representation space

Query A



Closest to A



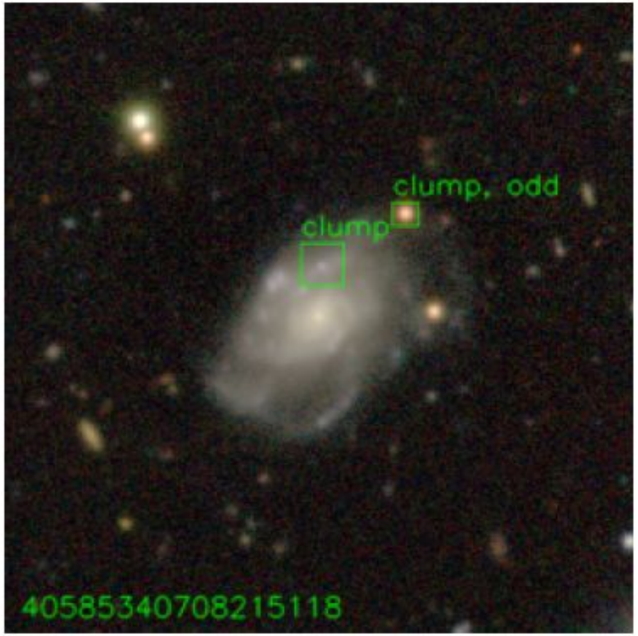
Query B

Closest to B

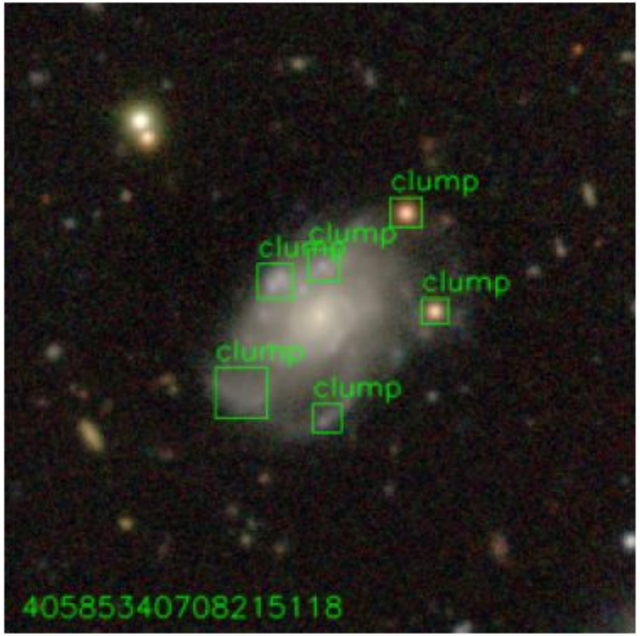
Mike Walmsley et al

ZOØNIVERSE

Without Zoobot



With Zoobot



Faster-RCNN clump detection in HSC

Project by **Jürgen Popp** (OU)

DESI 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

ZOØNIVERSE

Challenges:

- How do we run our code on Euclid data?
- Benchmarking for reliable performance

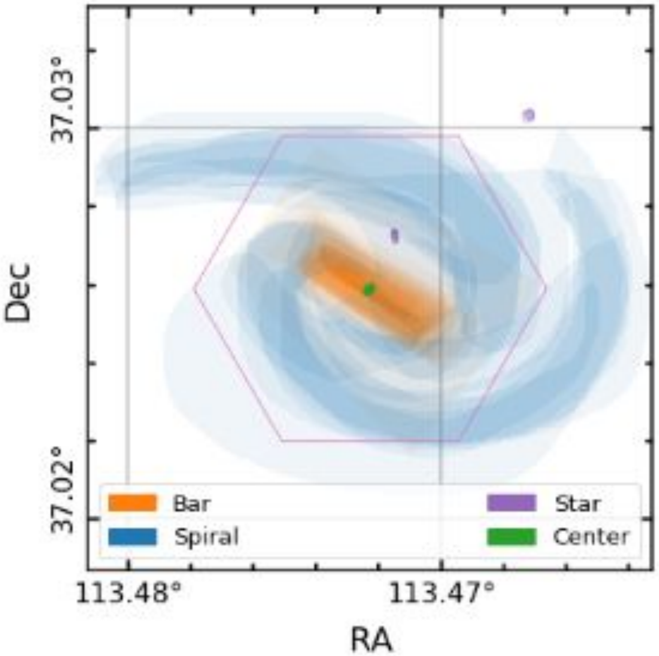
Opportunities:

- **Localise** galaxy features
- **Adapt models** to answer your legacy science questions

Pixel Segmentation

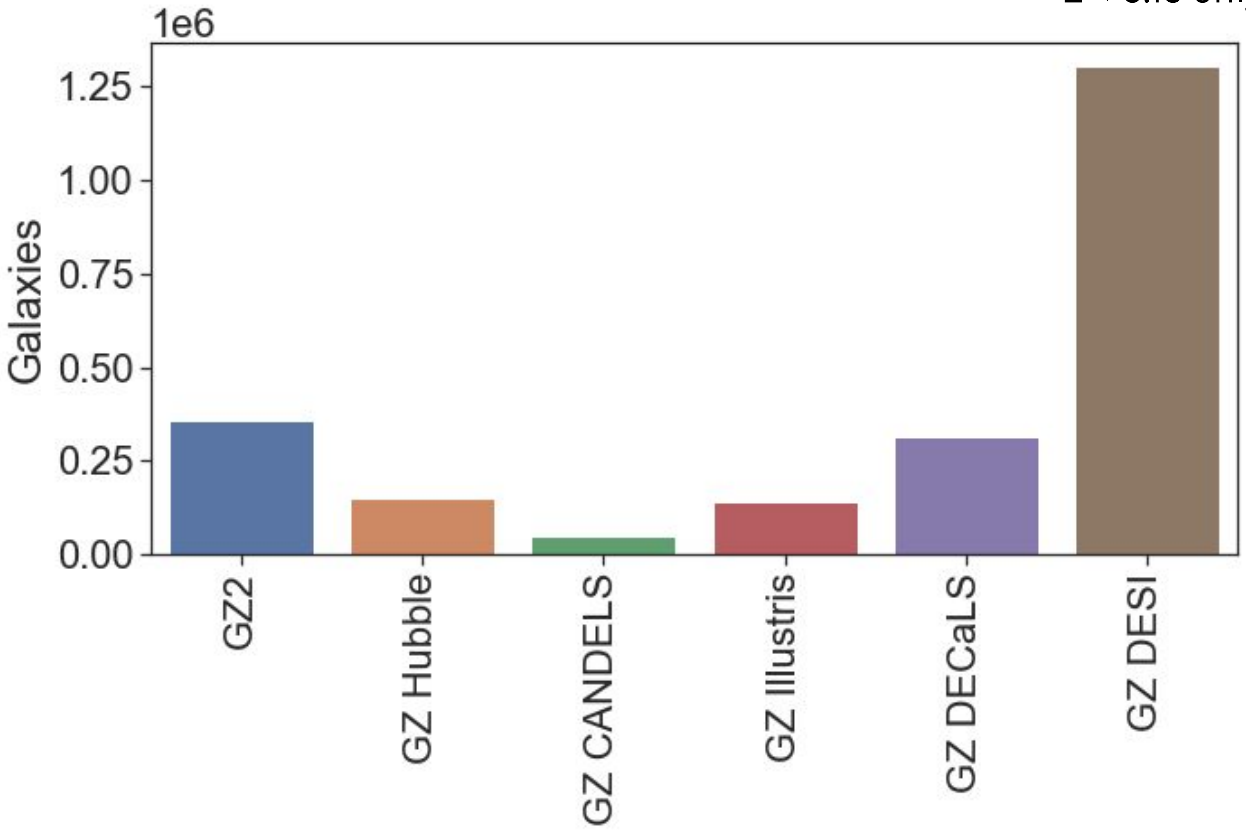
Identify pixels of features
Calculate shapes, SFR, etc.

Tidal tails, streams, shells
Spiral arms
Bars
So much more…

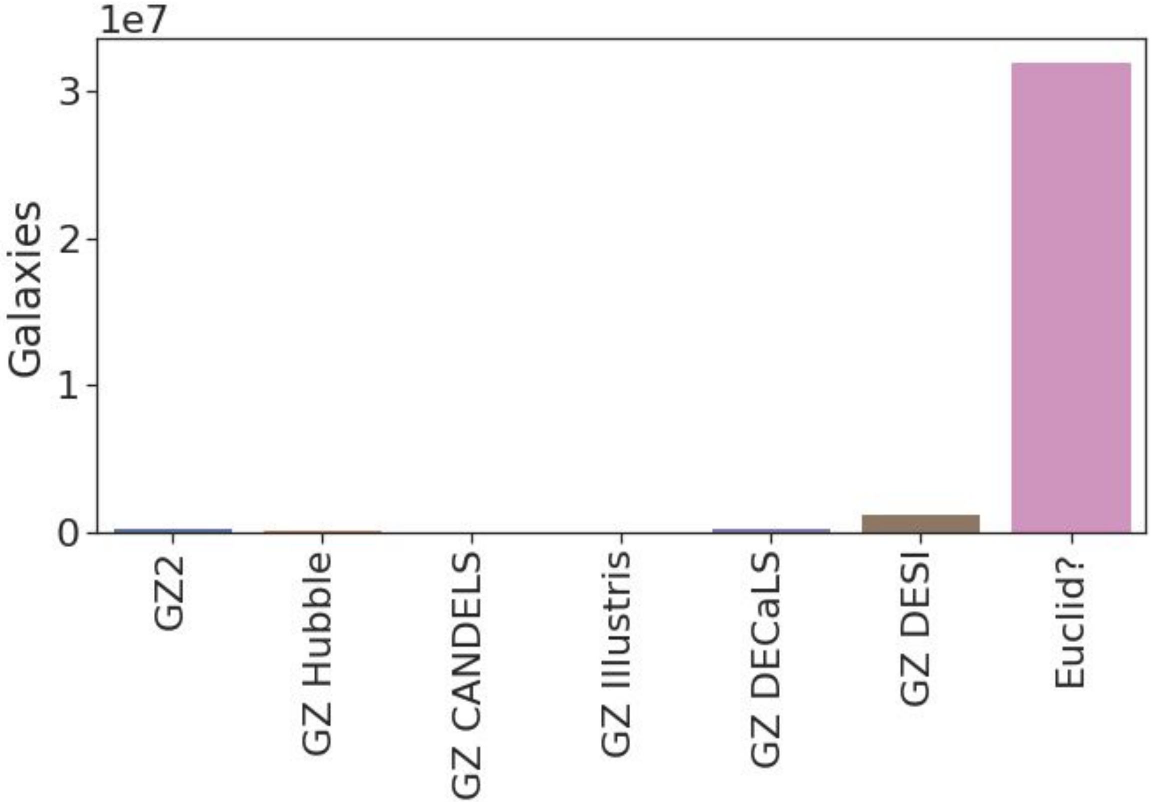


Volunteer annotations
Masters et. al. 2021

Z < 0.15 only, 8.7M full sample



Mike Walmsley et al

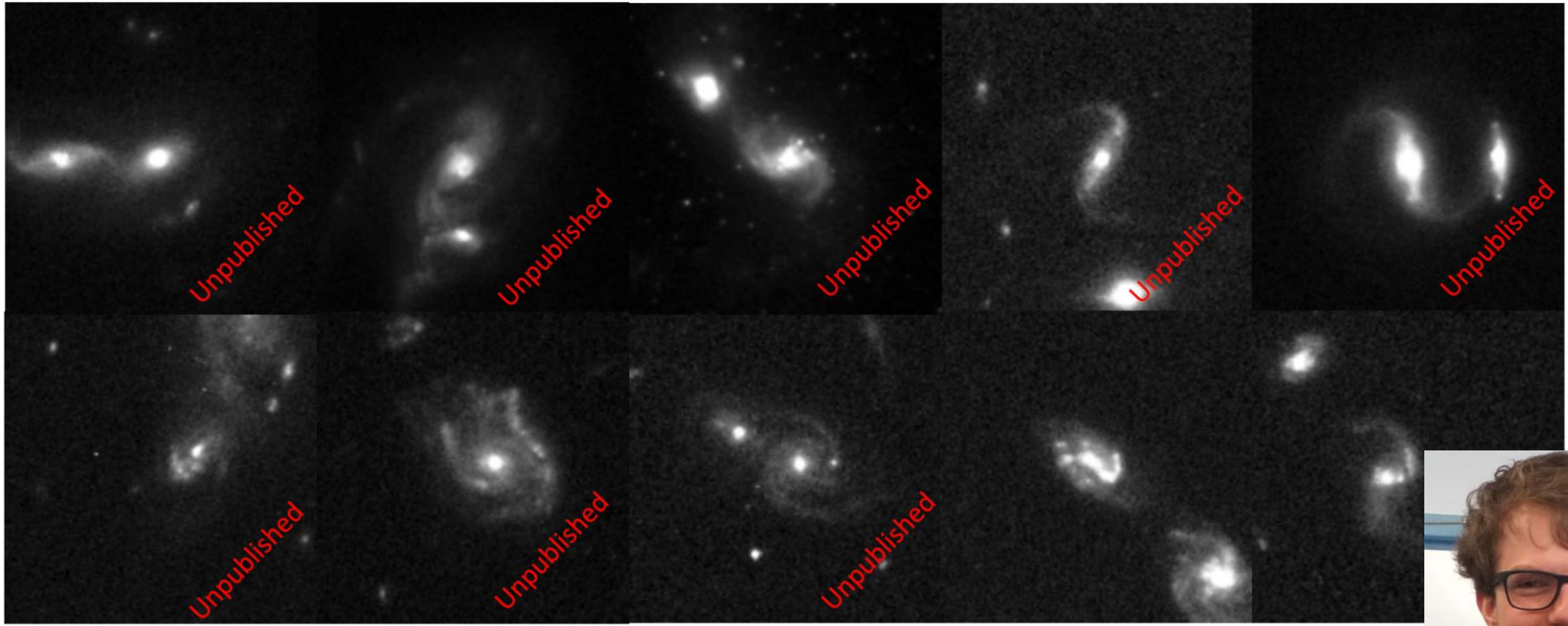


~32M galaxies
with
VIS mag < 22
Half-light > 0.4”

And they’re at *much* higher z!

Mike Walmsley et al

Results I: The Unknown Gems of the Archive



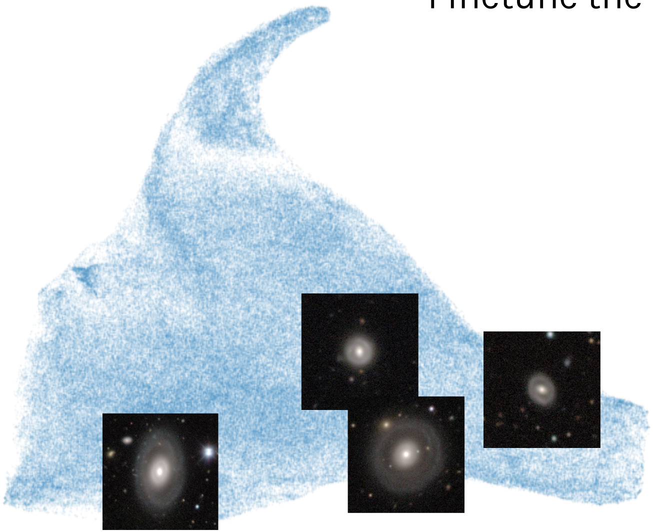
Project by **David O'Ryan** (Lancaster) during **3 month ESA internship**

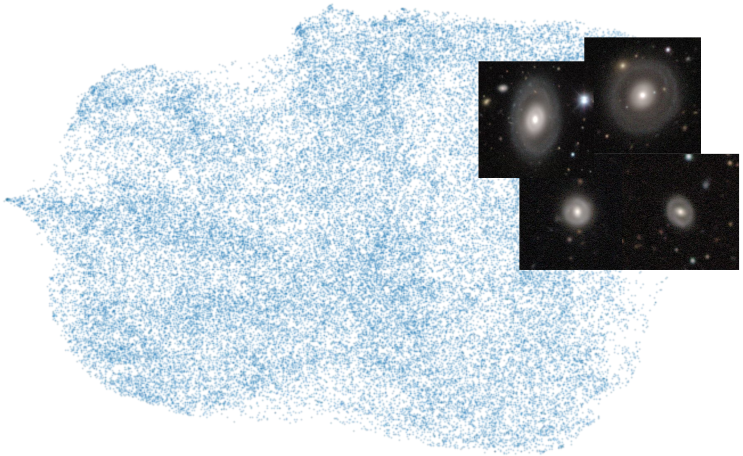
See e.g. Dominguez-Sanchez+19 for astro transfer learning background

Transfer Learning

Start with a few hundred labelled examples

Finetune the representation for your problem

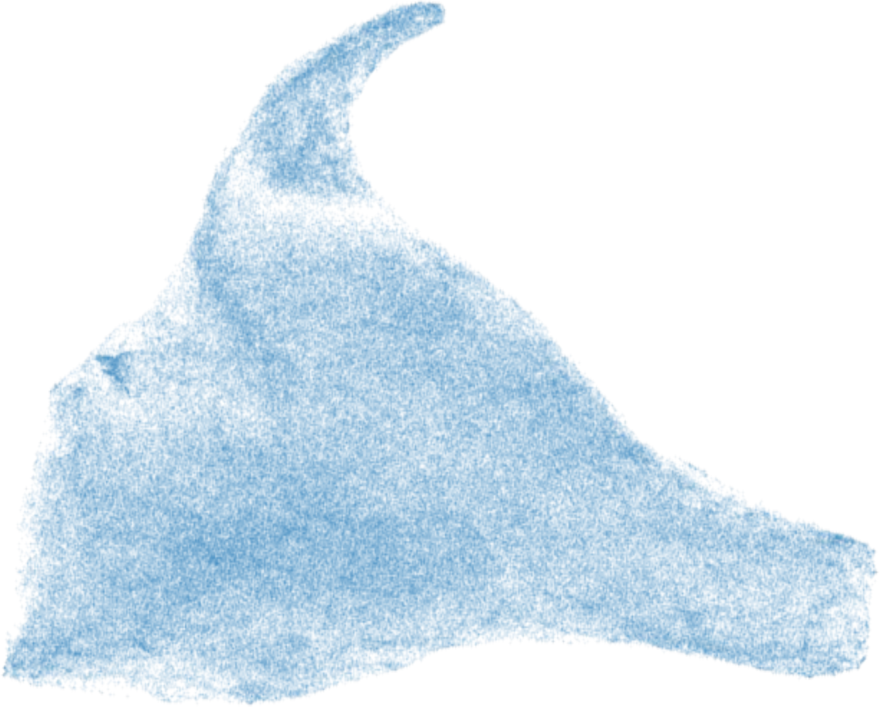




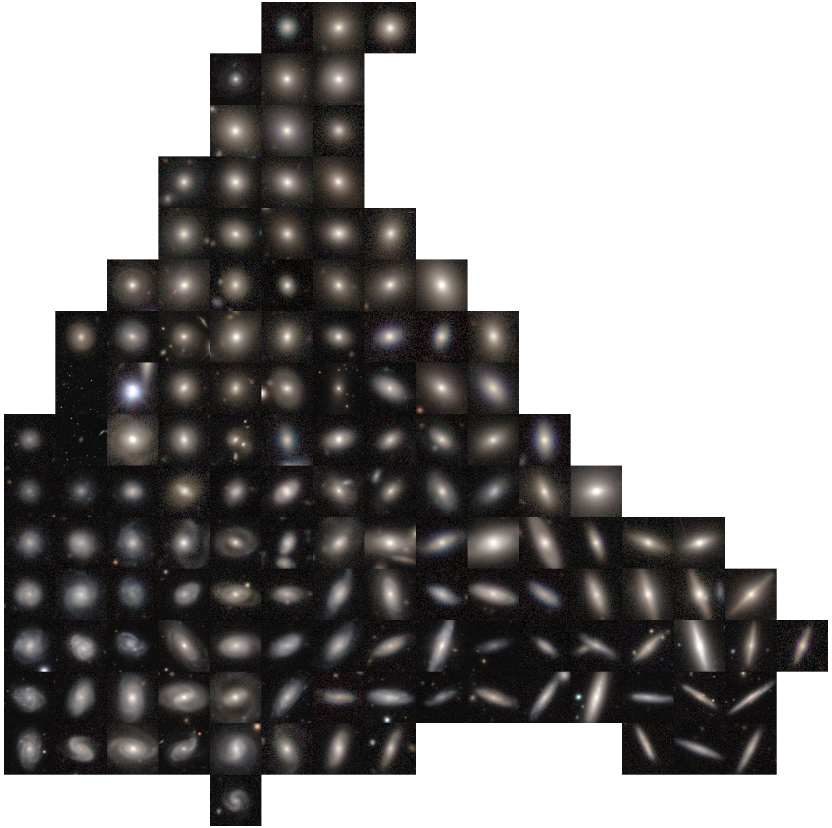
(illustrative figures only)

Meaningful Internal Representation

ZOØNIVERSE



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



Galaxies arranged by representation

Mike Walmsley et al

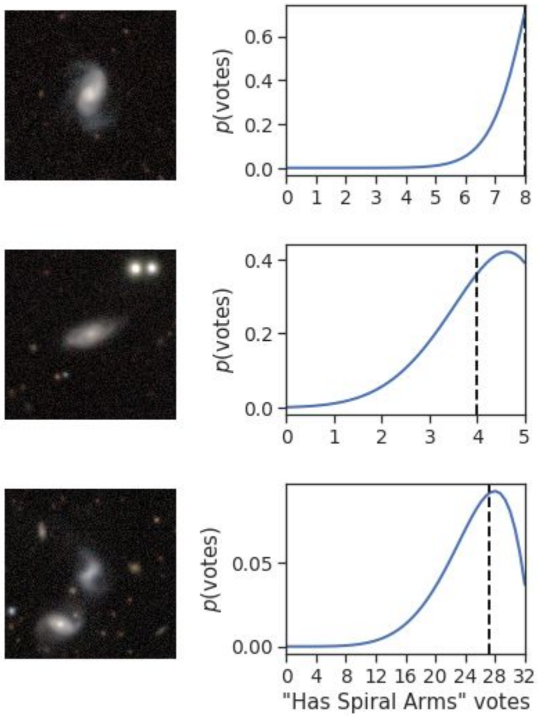
ZOØNIVERSE

Opportunities

- Easy access: include GZ DESI in DESI query service
- Links between spectra-derived parameters (everything!) and morphology

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

Probabilistic CNN

Volunteers N
Responses k
Typical vote prob. ρ
Galaxy x
CNN output $f^w(x)$

N volunteers and k responses ≈ N trials and k successes

How fair might the coin be? $\mathrm{Beta}(\rho|\alpha, \beta)$

Toss N times, get k heads $\mathrm{Bin}(k|\rho, N)$

How likely is each ρ given observed k, N?

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

Predict $f^w(x) = \alpha, \beta$ and maximise the likelihood of α, β

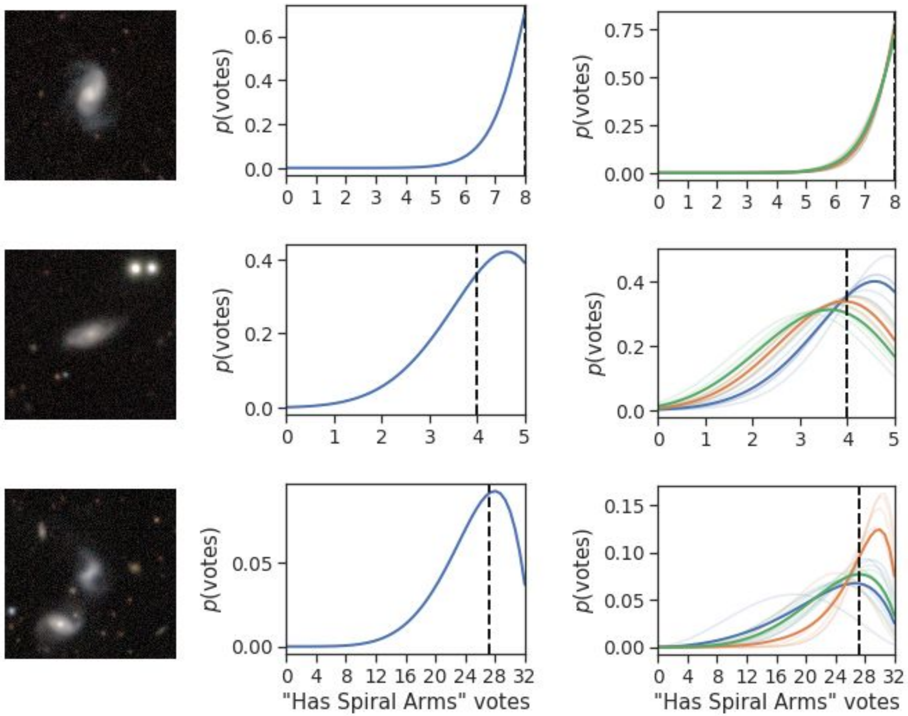
Mike Walmsley et al

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

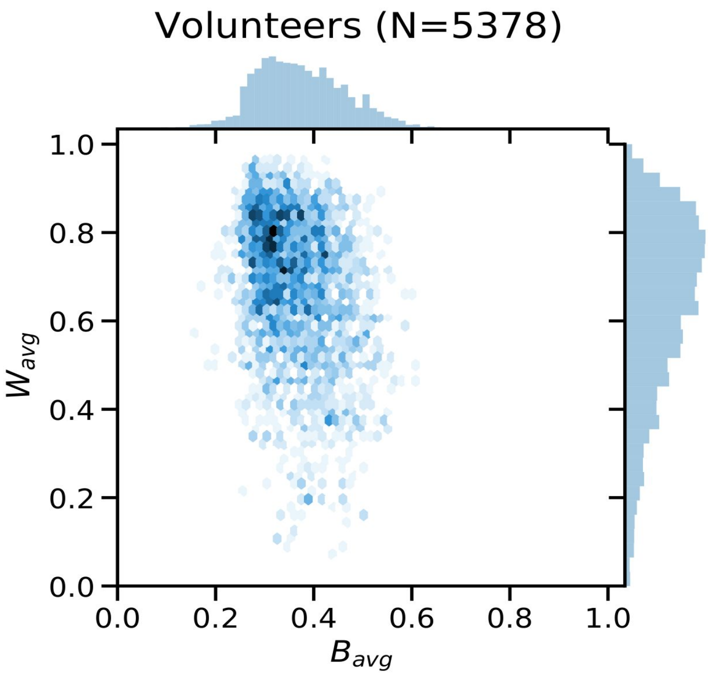


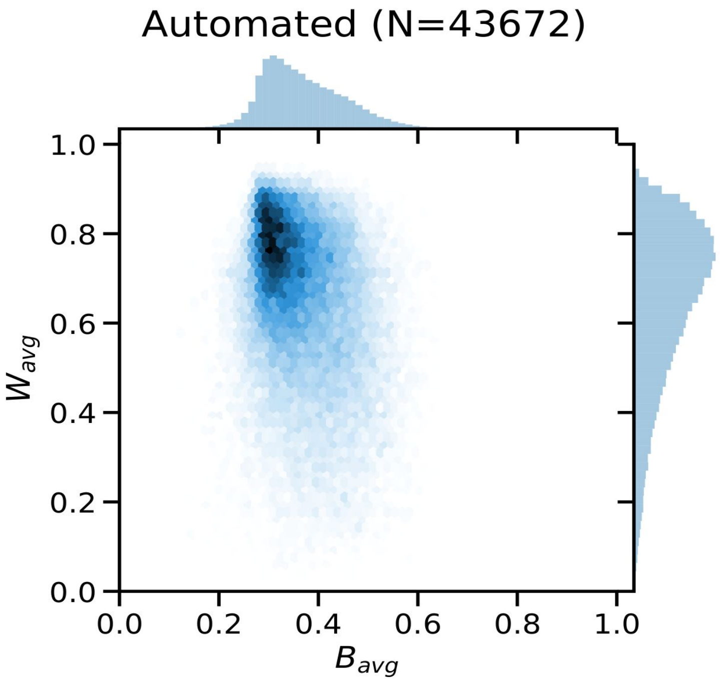
1 Model   15 "Models" (BCNN)

For more, see arxiv:2102.08414

For more, see arxiv:2102.08414

ZOONIVERSE





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

zenodo.org/record/4196267

Mike Walmsley et al

ZOØNIVERSE

Galaxy Zoo DESI: Detailed Morphology Measurements for 8.7M Galaxies in the DESI Legacy Imaging Surveys

ABSTRACT

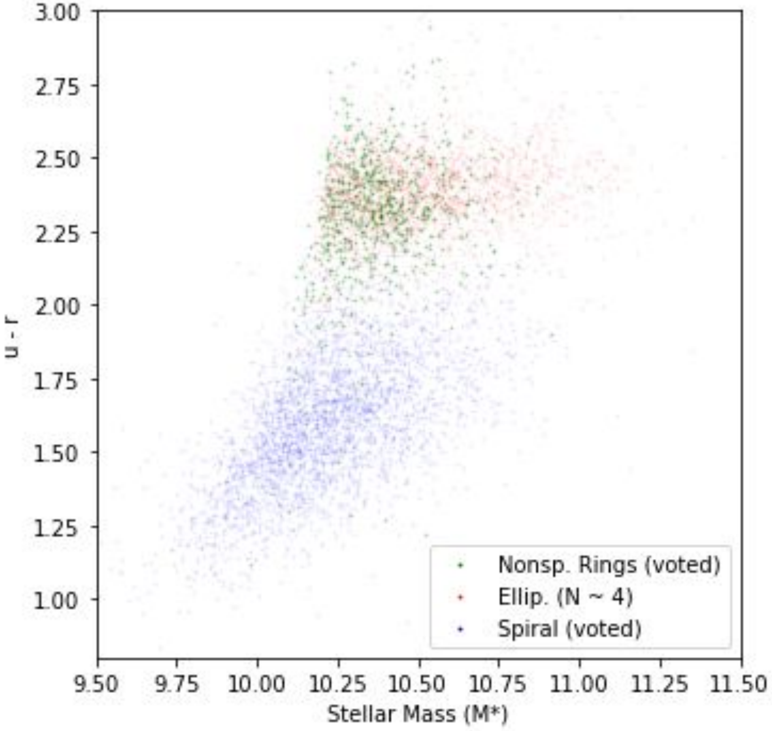
We present detailed morphology measurements for 8.67 million galaxies in the DESI Legacy Imaging Surveys (DECaLS, MzLS, and BASS, plus DES). These are automated measurements made by deep learning models trained on Galaxy Zoo volunteer votes. Our models typically predict the fraction of volunteers selecting each answer to within 5-10% for every answer to every GZ question. The models are trained on newly-collected votes for DESI-LS DR8 images as well as historical votes from GZ DECaLS. We also release the newly-collected votes. Extending our morphology measurements outside of the previously-released DECaLS/SDSS intersection increases our sky coverage by a factor of 4 (5,000 to 19,000 deg^2) and allows for full overlap with complementary surveys including ALFALFA and MaNGA.

Key words: catalogues, software: data analysis, methods: statistical, galaxies: bar, galaxies: interaction, galaxies: general

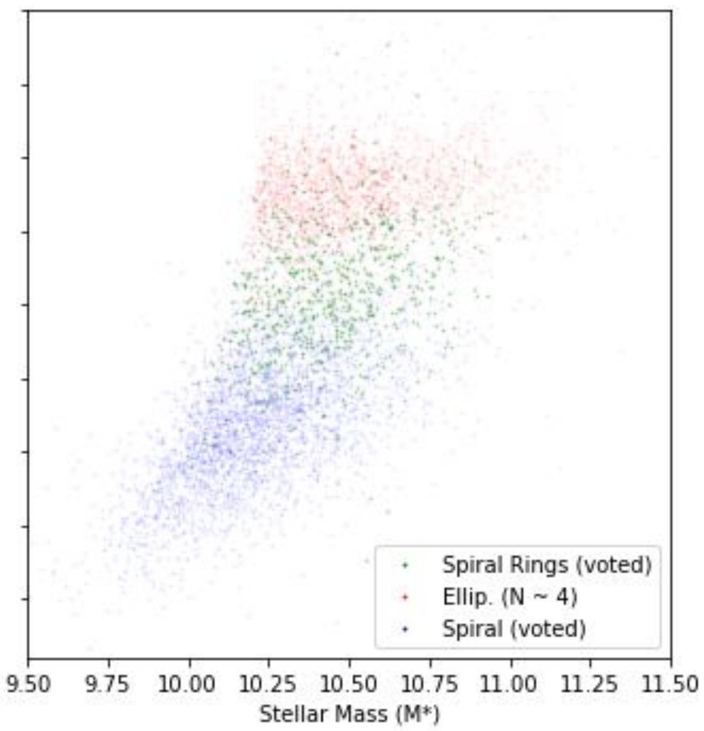
Mergers and Tidal Features in HSC

Project by **Kiyoaki Omori** (Kavli IPMU)

ZOØNIVERSE



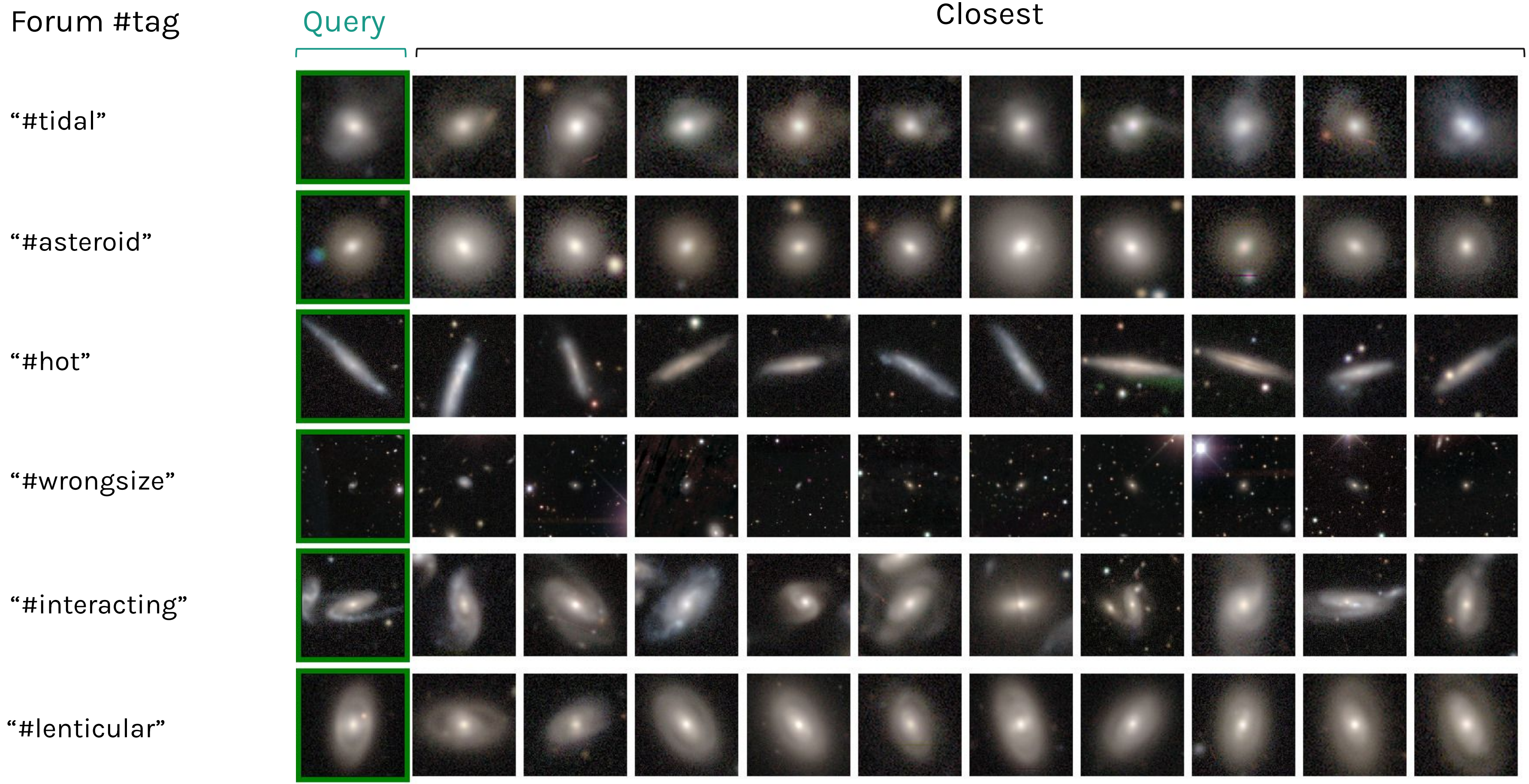
Non-spiral rings
(green)



Spiral rings
(also green)

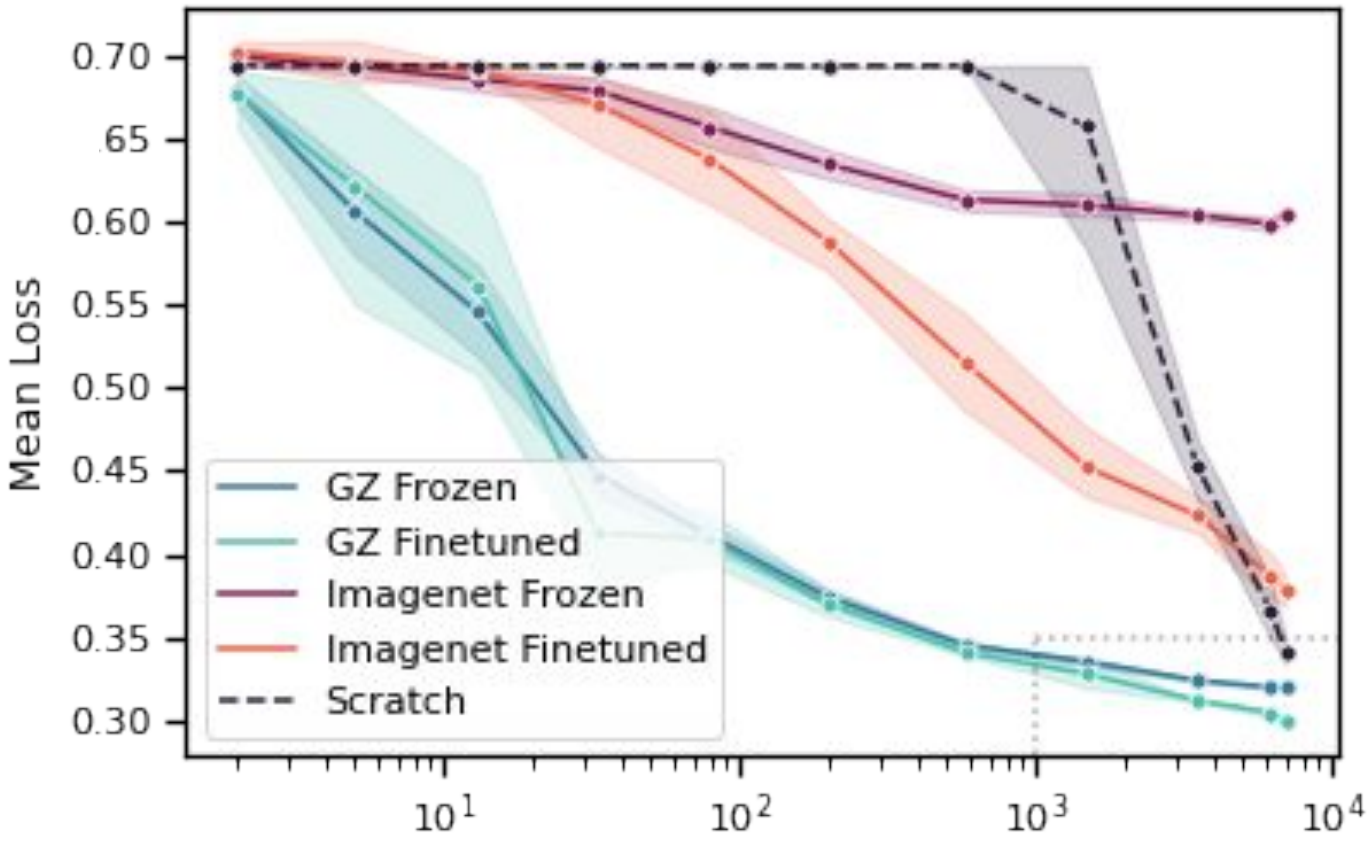
Mike Walmsley et al

Forum #tag



Mike Walmsley et al

ZOØNIVERSE



Ring Galaxies to Train On

Mike Walmsley et al

Work by Ben Aussels, Sandor Kruk

"Euclid" Performance

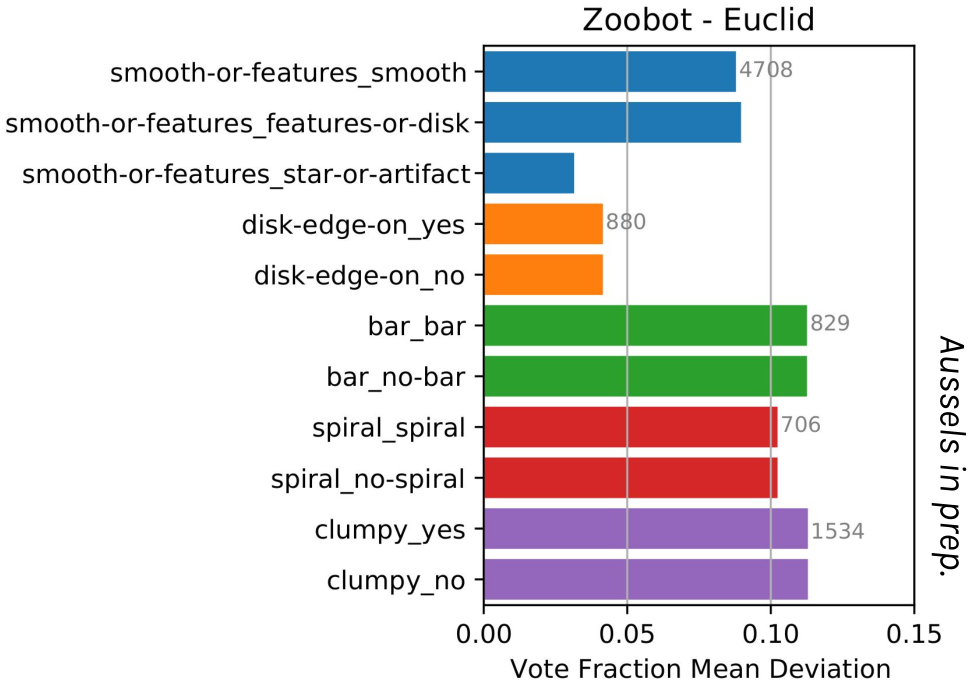
via HST with Euclid PSF

Predicts the fraction of GZ volunteers giving each answer

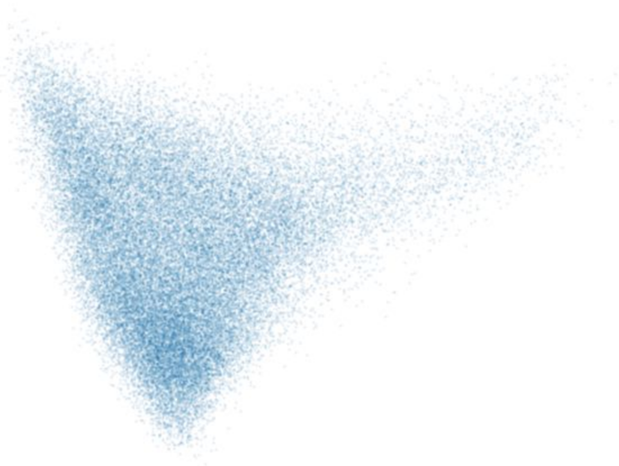
Typically **accurate to ~10% of the human answers**

For example :

Ask 10 people
Model predicts 6 say "bar"
5 actually say "bar"



Finding Interesting Anomalies



Representation to explore

Rate a galaxy* by interest

Train regression model** on your interests

Coloured by expected interest

*Active learning for highest expected improvement; see active-learning.net

**Gaussian process, as uncertainties are useful for active learning

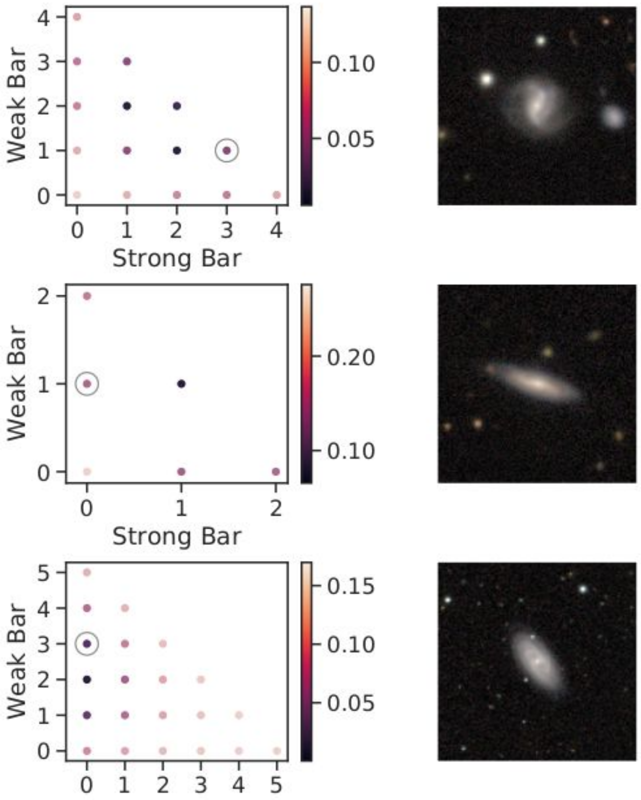
Mike Walmsley et al

Multiple Answers

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

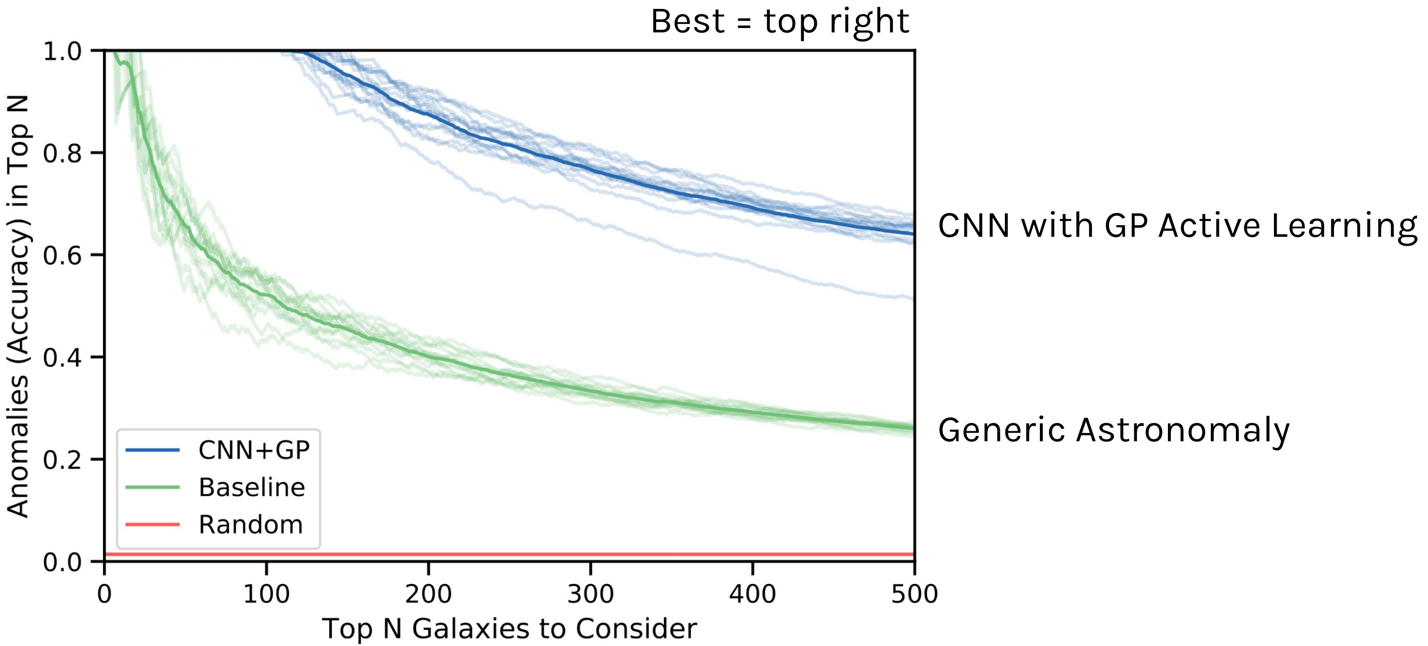
Add a few dimensions…

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



Mike Walmsley et al

More anomalies faster via deep representation + active learning



See Lochner and Bassett (2021) for motivation and baseline, Walmsley (2021) for CNN+GP

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

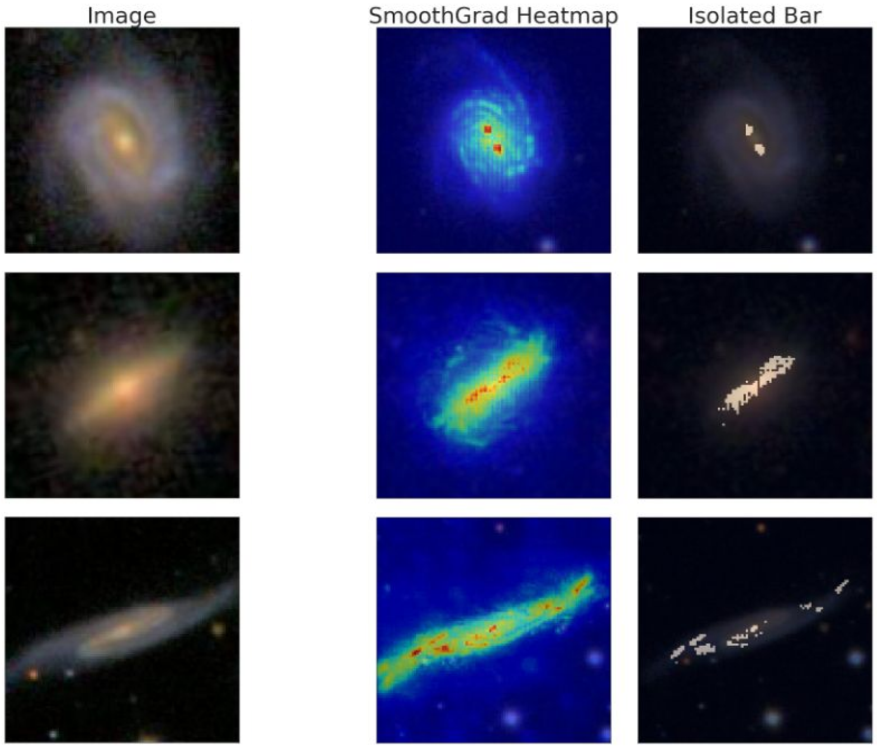
Convolutional Neural Network ("CNN")

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

Prabh Bhambra
Supervisors:
Ofer Lahav
Benjamin Joachimi

Bhambra+ 22
MNRAS 511 4

Without Zoobot

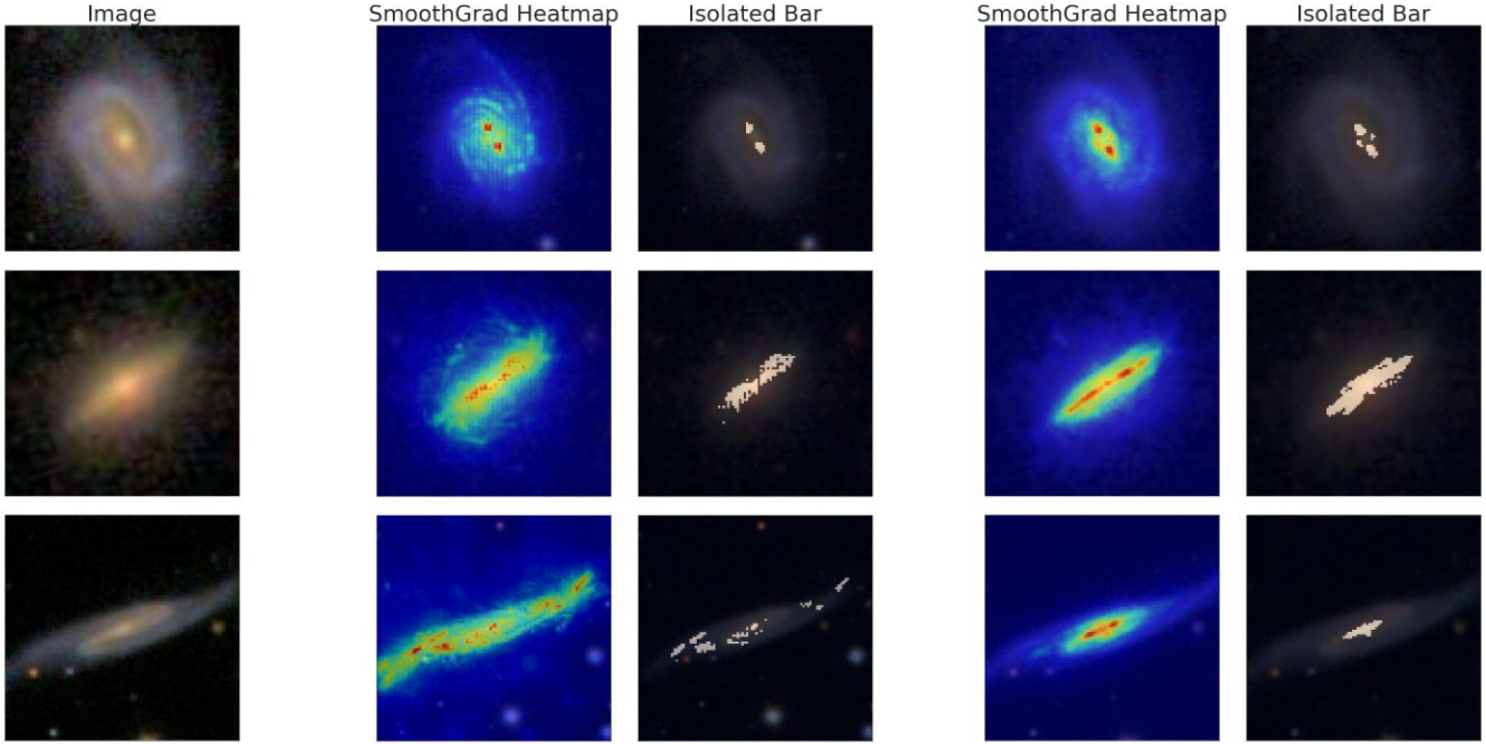


Prabh Bhambra
Supervisors:
Ofer Lahav
Benjamin Joachimi

Bhambra+ 22
MNRAS 511 4

From Scratch

With Zoobot



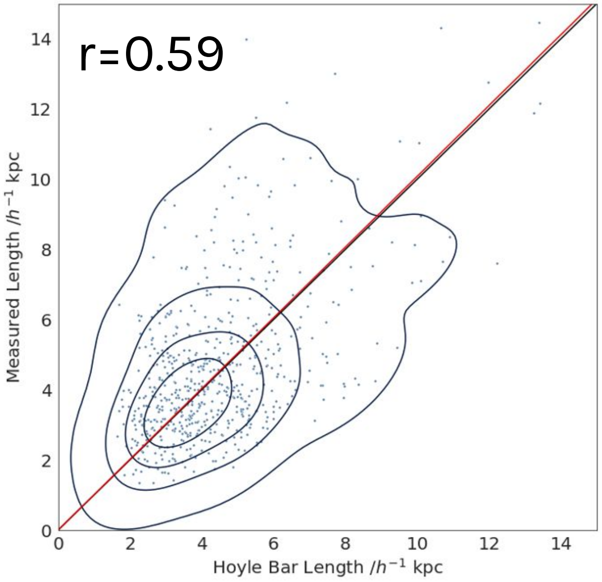
Prabh Bhambra
Supervisors:
Ofer Lahav
Benjamin Joachimi

Bhambra+ 22, MNRAS 511 4

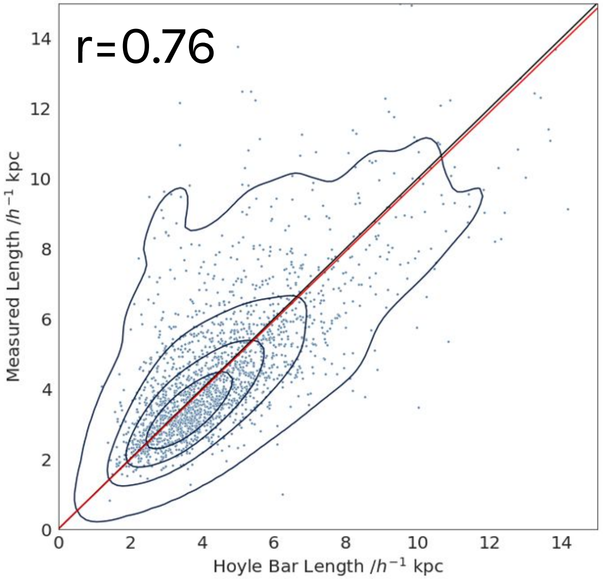
ZOØNIVERSE

Figures courtesy Prabh Bhambra

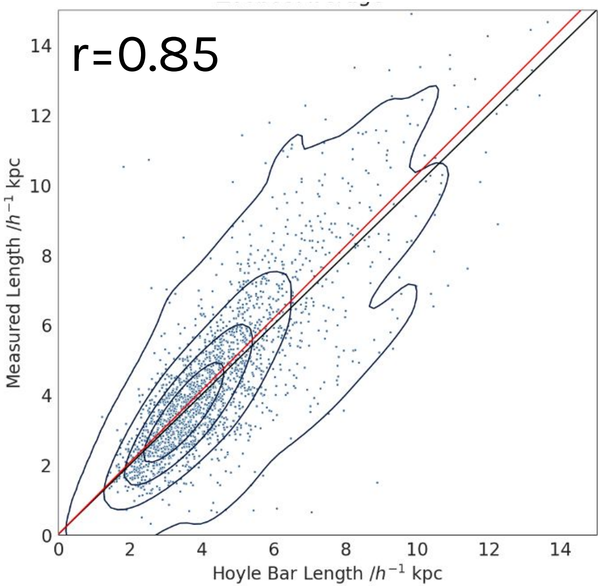
Direct Deep Regression



From Scratch



With Zoobot



Estimated vs. True Bar Length

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

- 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?

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?

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

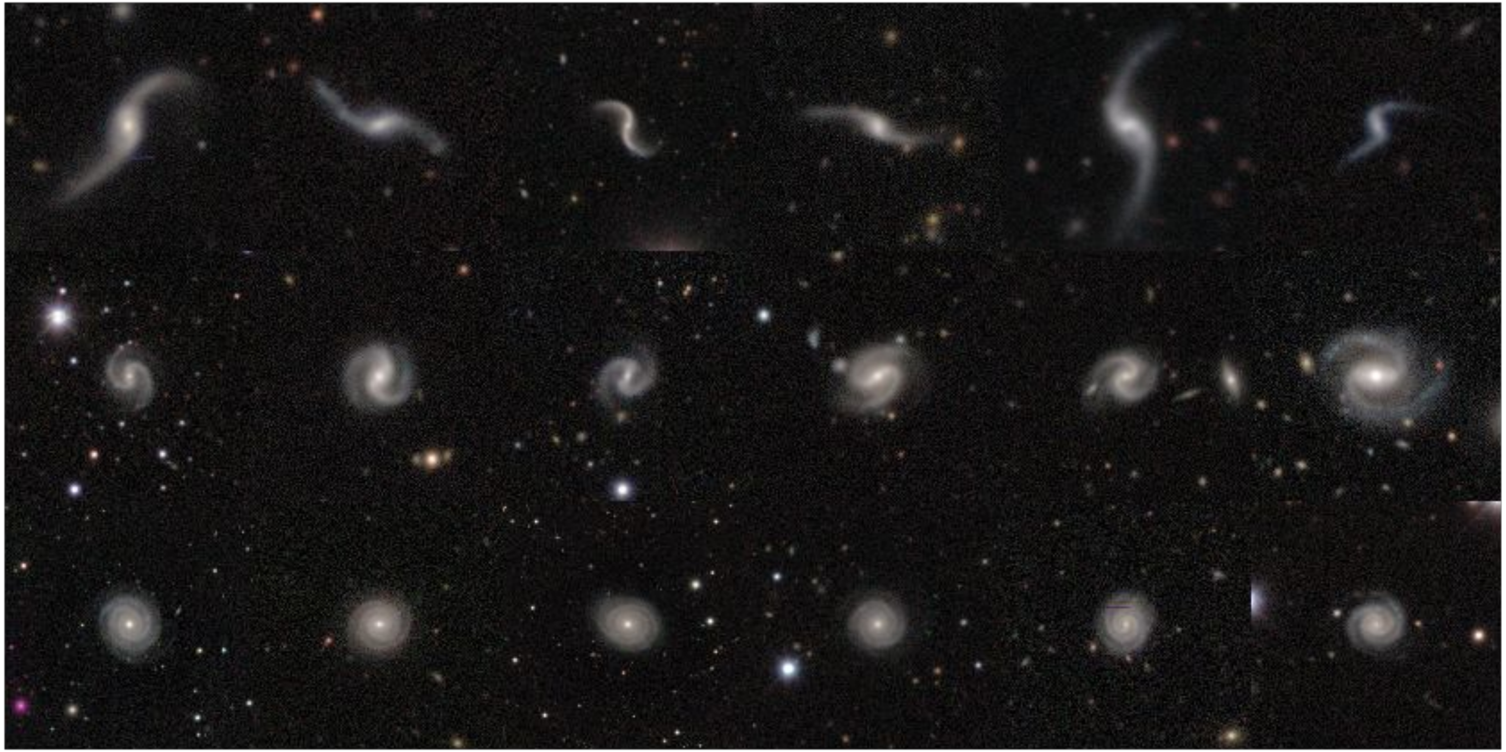
Unknown!

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

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

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

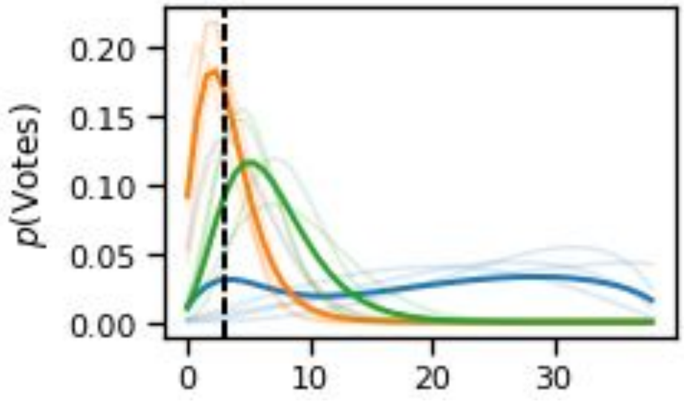
See Y. Gal et al (2016)



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

Mike Walmsley et al

See Houlsby (2014)



Featured?

Pick galaxies where the models **confidently** disagree.

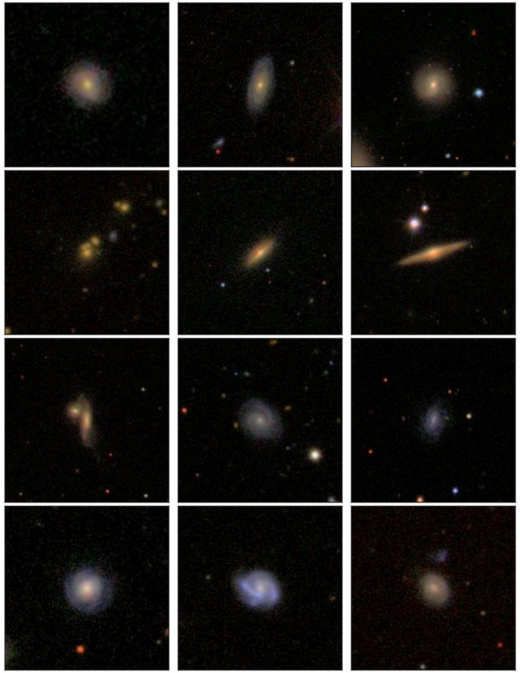
$$I = -\int H[p(k|w)]\ p(w|D)\ dw + H\left[\int p(k|w)\ p(w|D)\ dw\right]$$

Each model is confident…

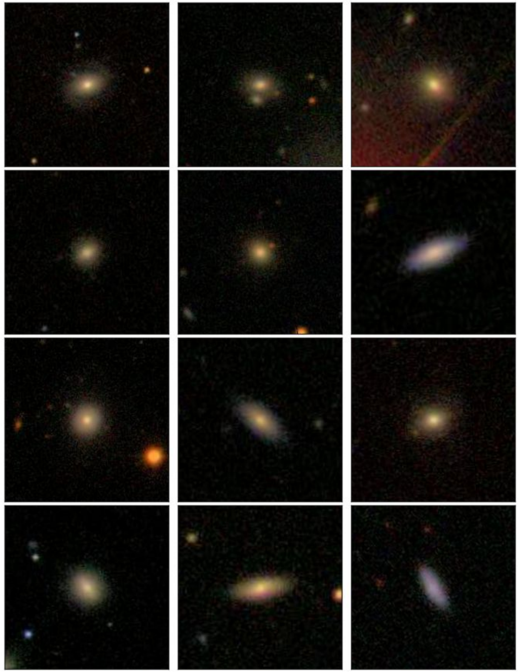
…but they give different answers

Mutual Information I
Entropy H
Votes k
Weights w
Training data D

Selected Galaxies for “Smooth?”



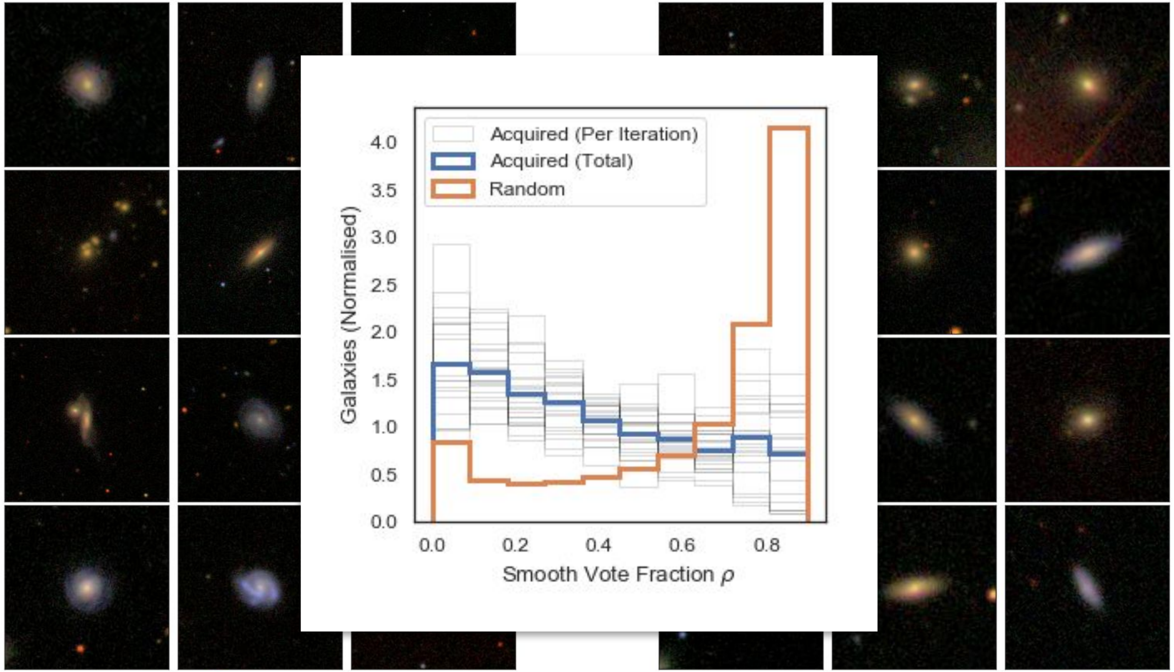
High mutual information



Low mutual information

Mike Walmsley et al

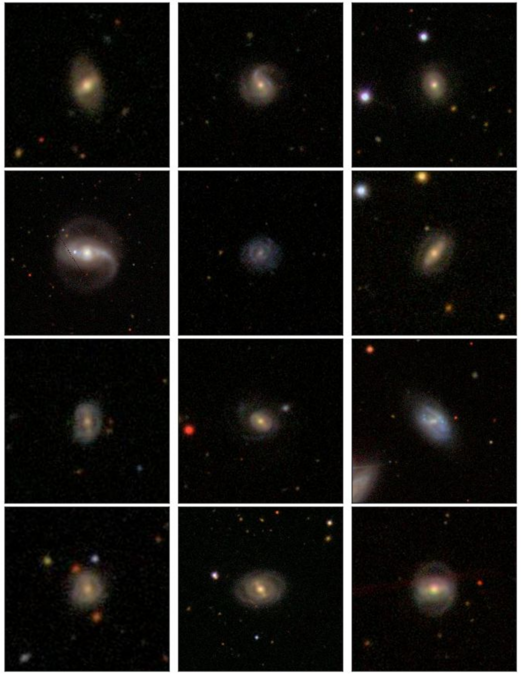
Selected Galaxies for “Smooth?”



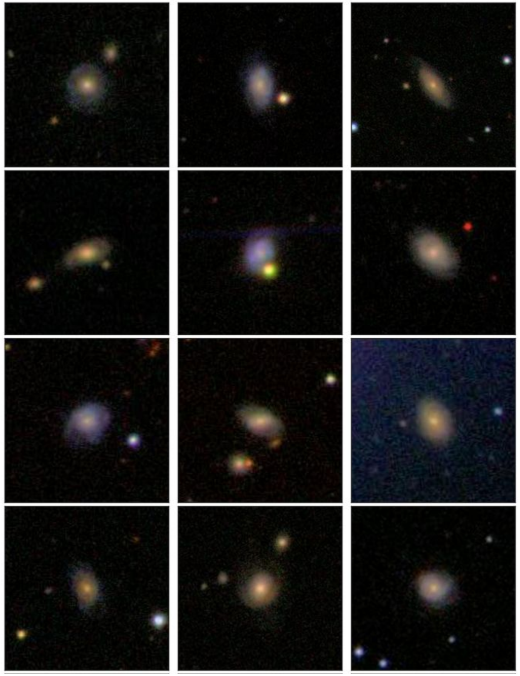
High mutual information

Low mutual information

Selected Galaxies for “Bar?”



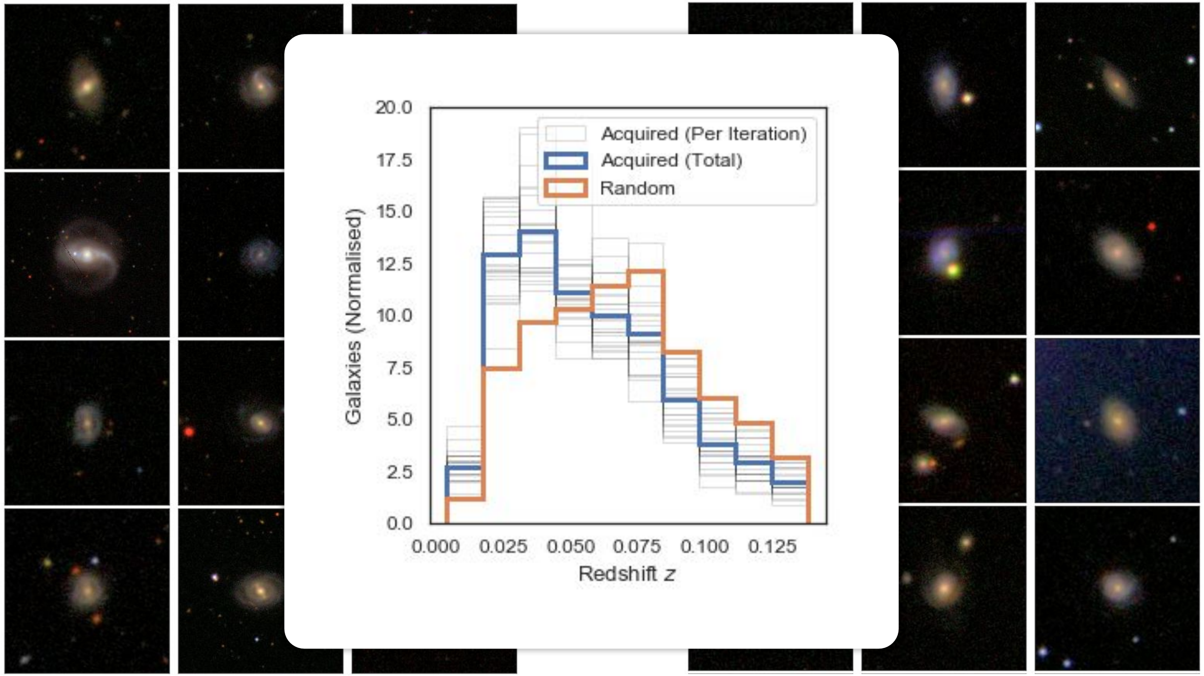
High mutual information



Low mutual information

Mike Walmsley et al

Selected Galaxies for “Bar?”



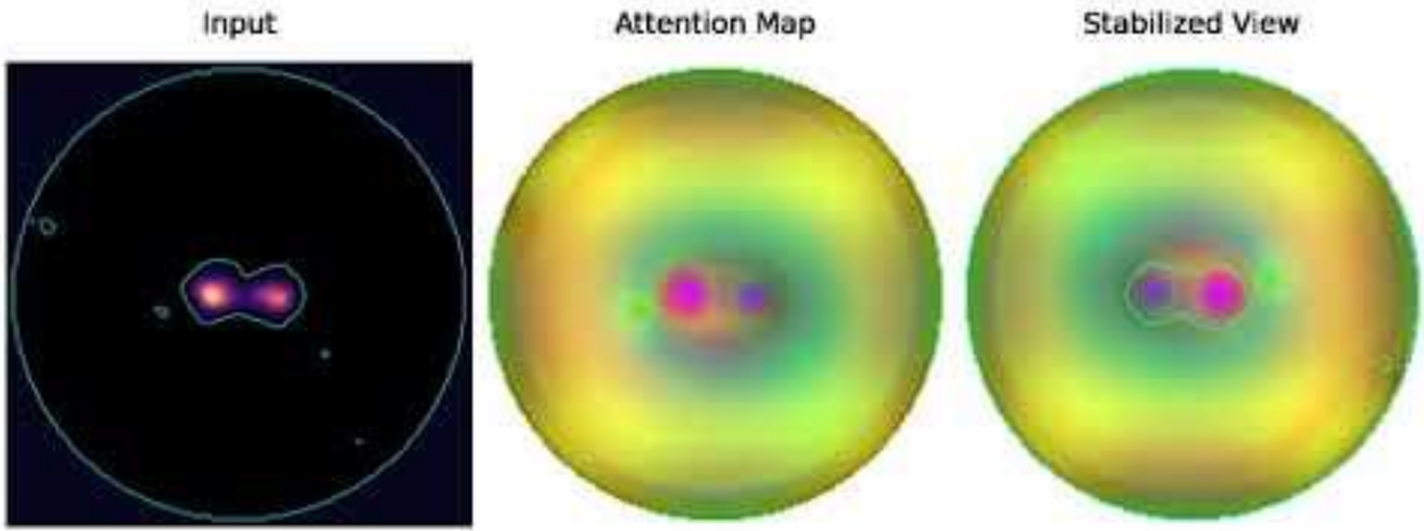
High mutual information

Low mutual information

Mike Walmsley et al

1. Build a Bayesian Galaxy Zoo model

2. Mess around



ZOØNIVERSE

Use Symmetry

Helps constrain model parameters

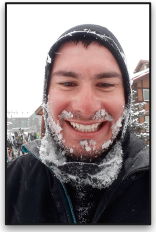
More constraints = less training data needed

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

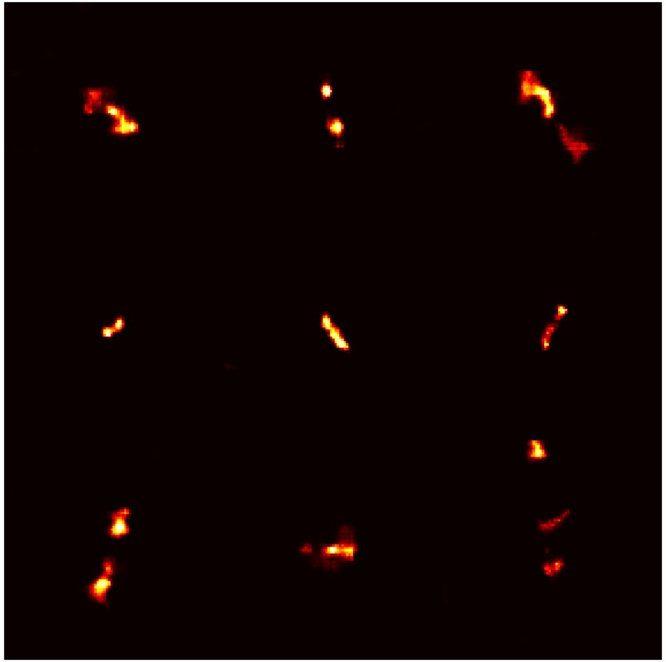
Extra Galaxies with GANs

Synthesise new training data

Train GAN on one class to create more examples

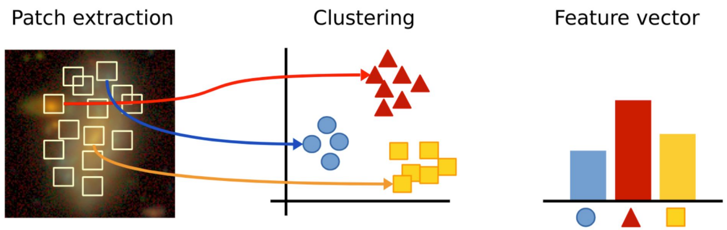


Inigo Val
inigo.val@postgrad.manchester.ac.uk

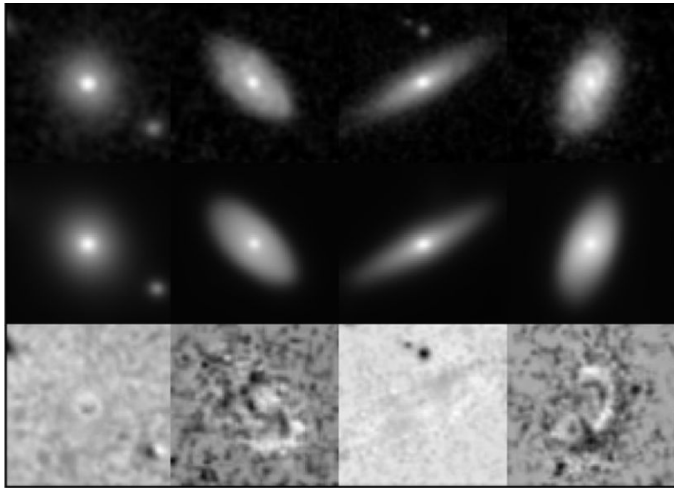


GAN-created ***radio*** *galaxies.*
Not real!

No Labels Needed?



Clustering image patches
Martin (2020)
See also Hocking (2017)



Learning to reconstruct images
Spindler (2020)

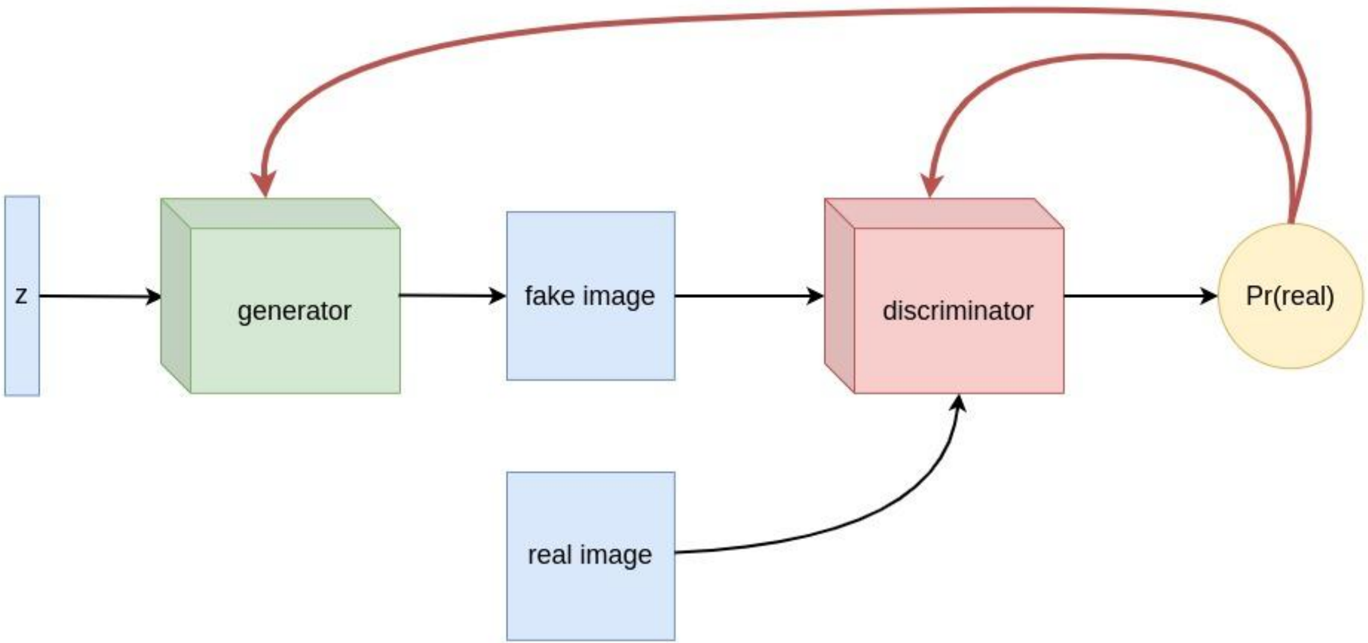


Diagram of a generative adversarial network (GAN)

- 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.