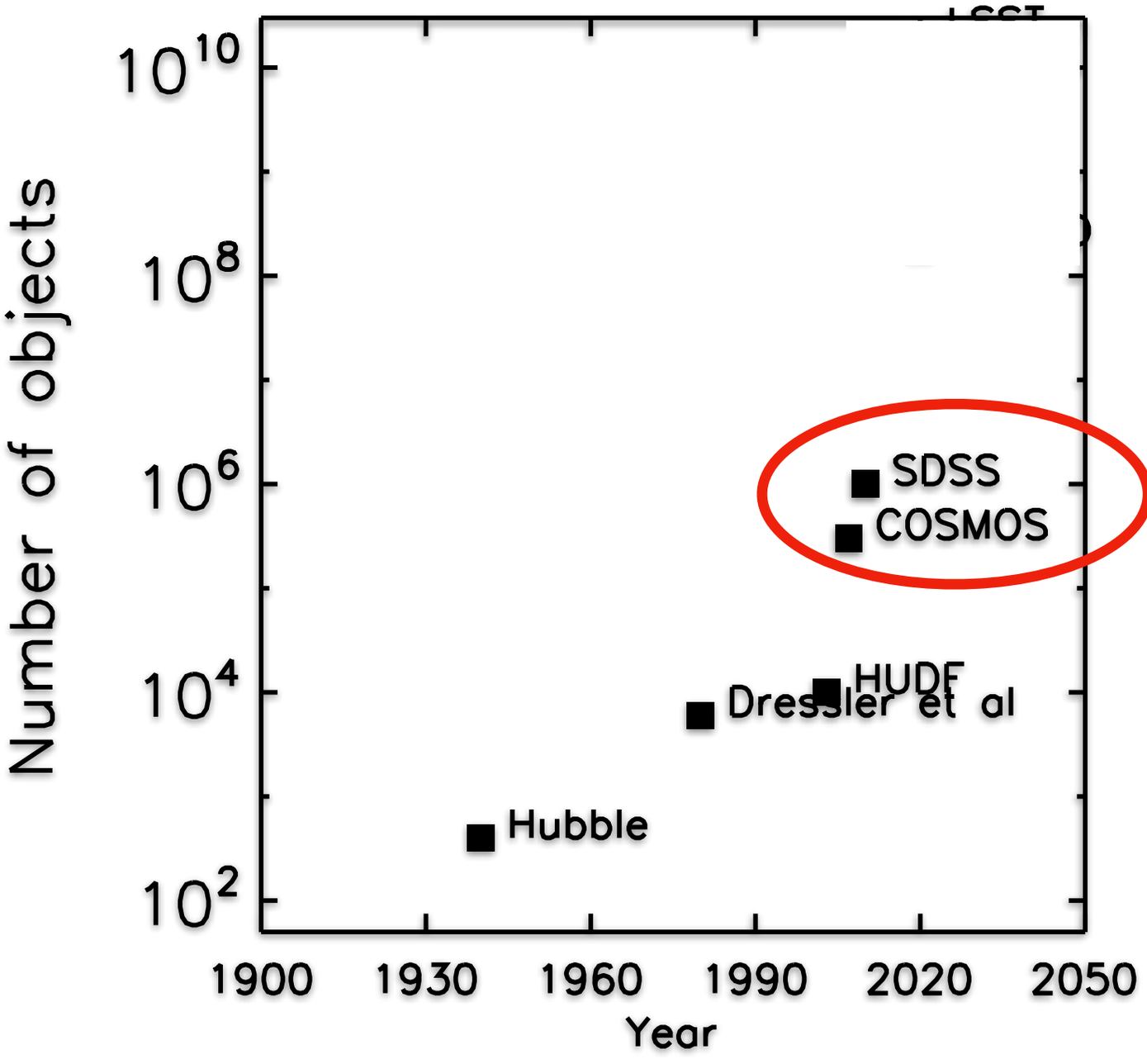
Galaxy classification in surveys

Helena Domínguez Sánchez

In collaboration with: M. Huertas-Company, J. Vega-Ferrero

Somachine School - IAA - 20/04/2021

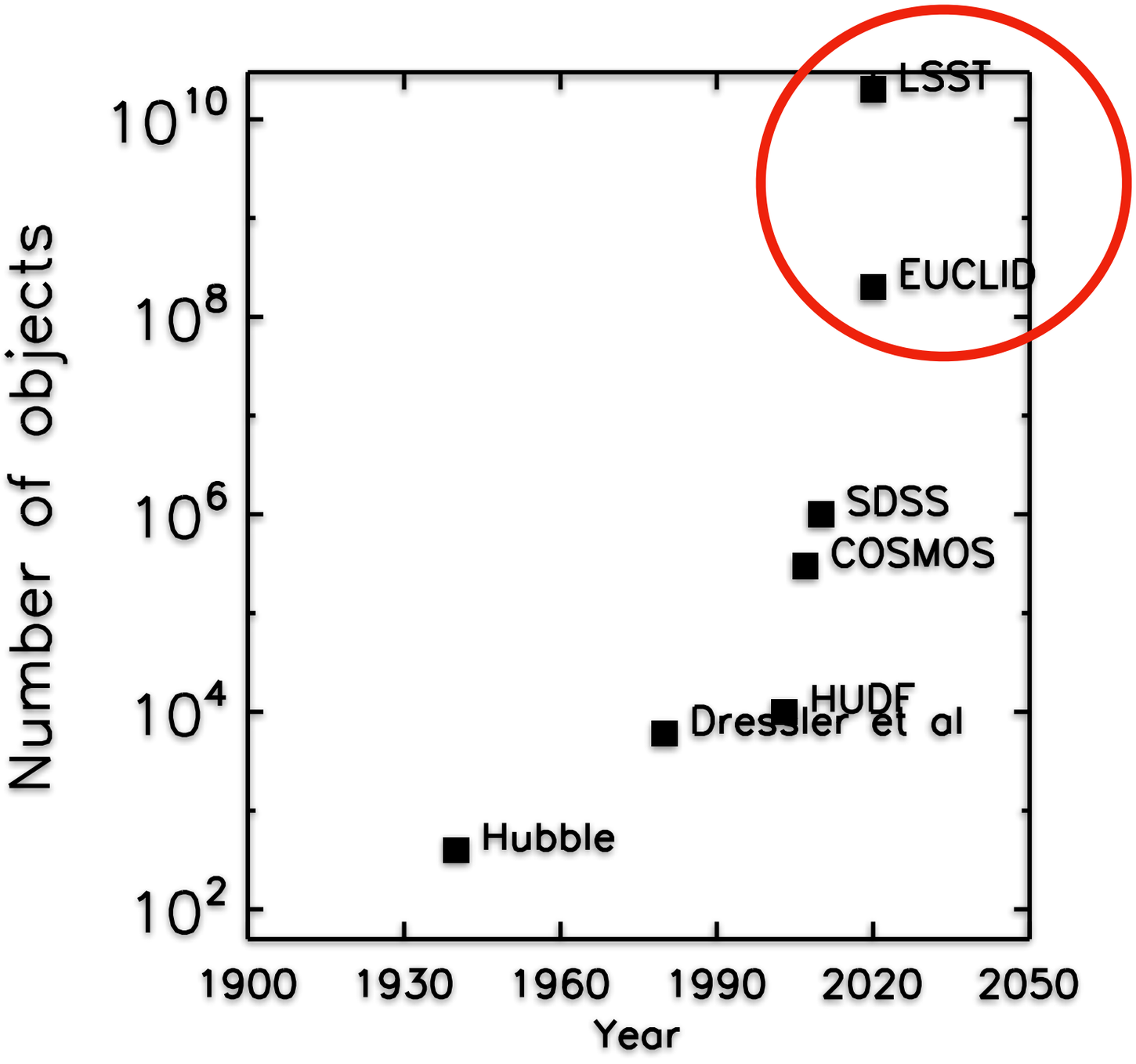
THE ERA OF STATISTICS



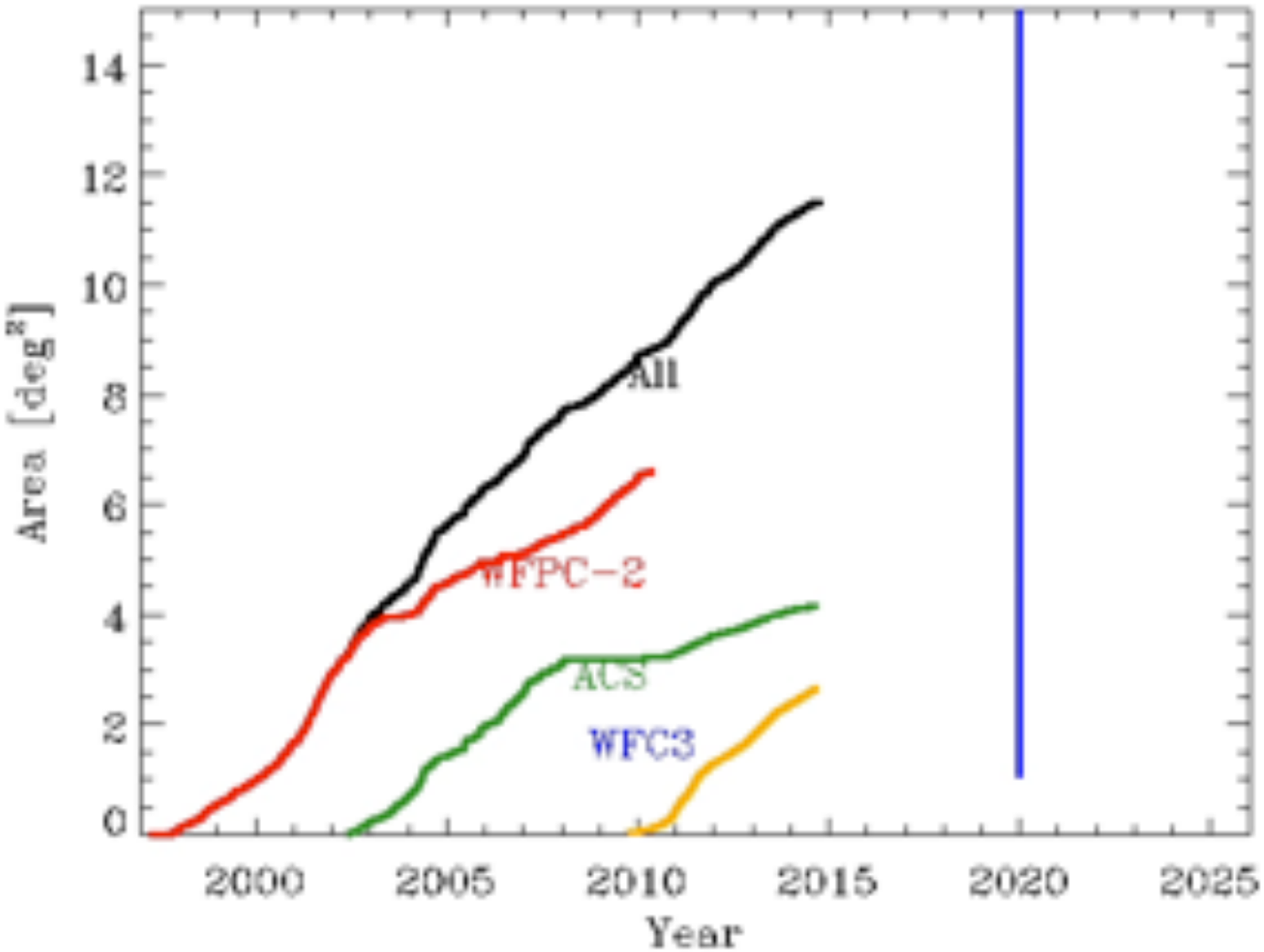
IN THE LAST ~20 YEARS ….

THE FIELD OF OBSERVATIONAL GALAXY EVOLUTION HAS EVOLVED INTO A STATISTICAL SCIENCE

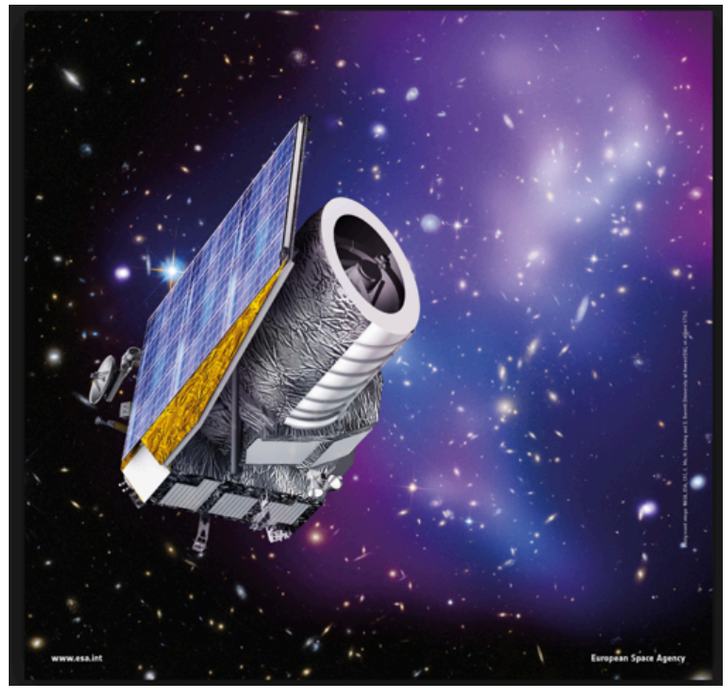
… AND NOW, AS MANY OTHER FIELDS, ENTERS THE BIG-DATA ERA



BIG-DATA HAS ALSO ARRIVED TO ASTRONOMY

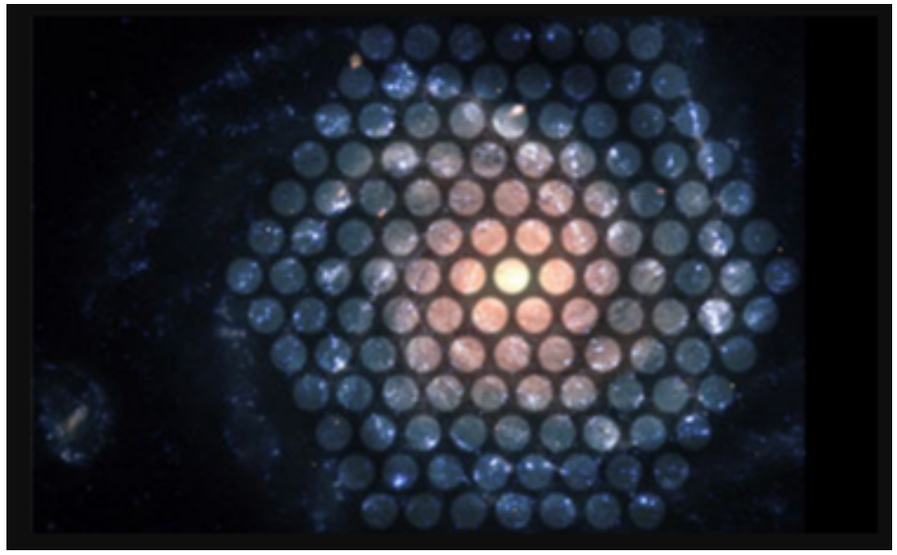


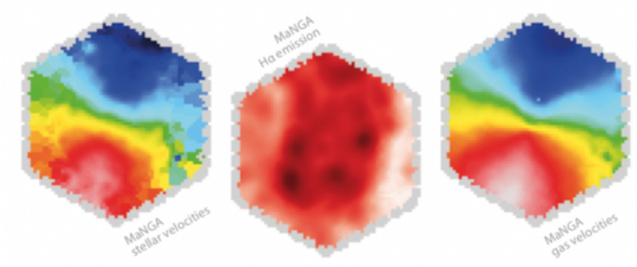
(Thanks to J. Brinchmann)



EUCLID space telescope (2021)

BIG DATA IS NOT ONLY SIZE…ALSO COMPLEXITY



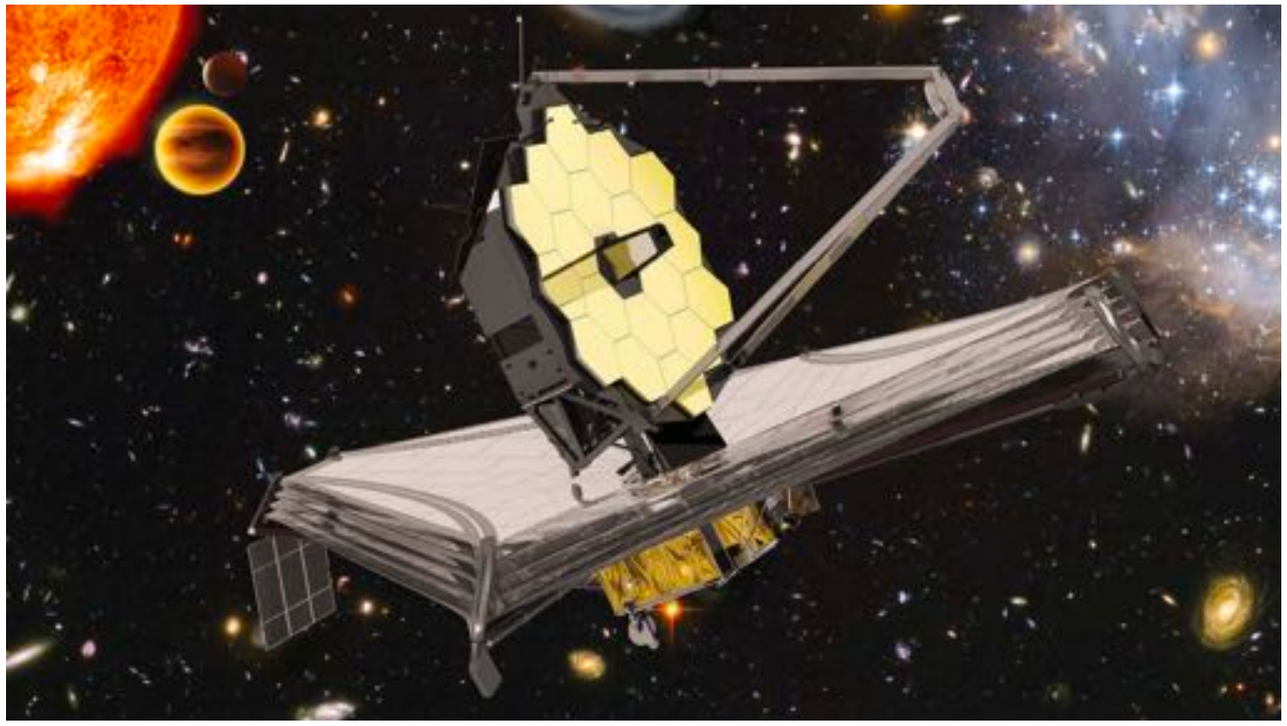


MANGA Survey

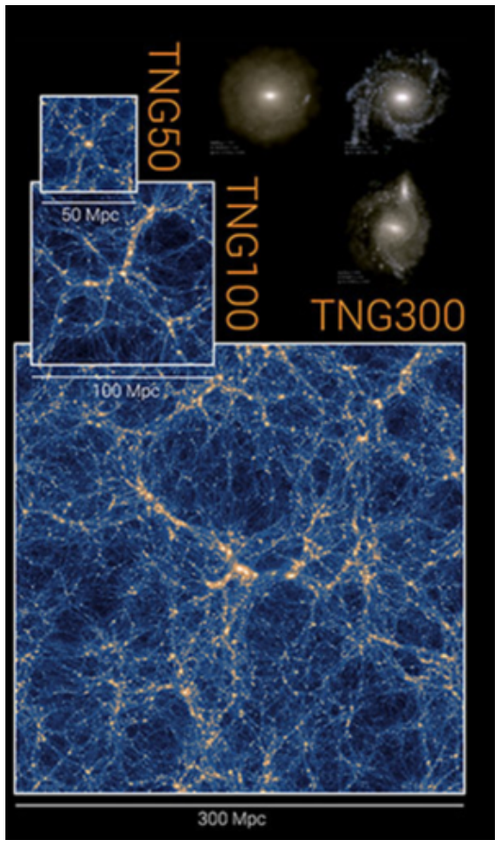
== IMAGES WITH HUNDREDS OF CHANNELS

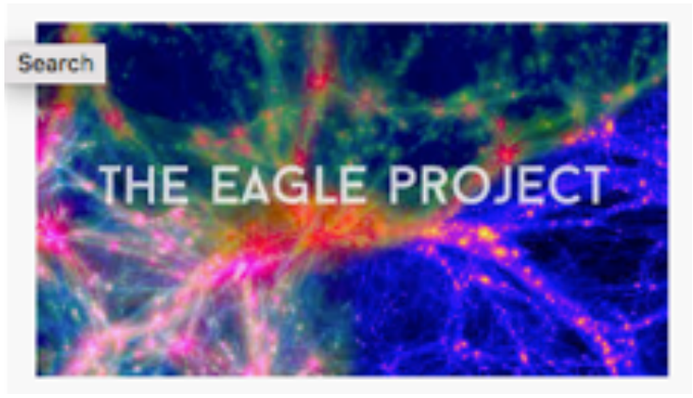
GALAXY STRUCTURE == WITH UNPRECEDENTED DETAIL

JAMES WEBB SPACE TELESCOPE



WE DO NOT HAVE THE CAPACITY TO "LOOK" AT FUTURE ASTRONOMICAL (BIG-DATA) SETS

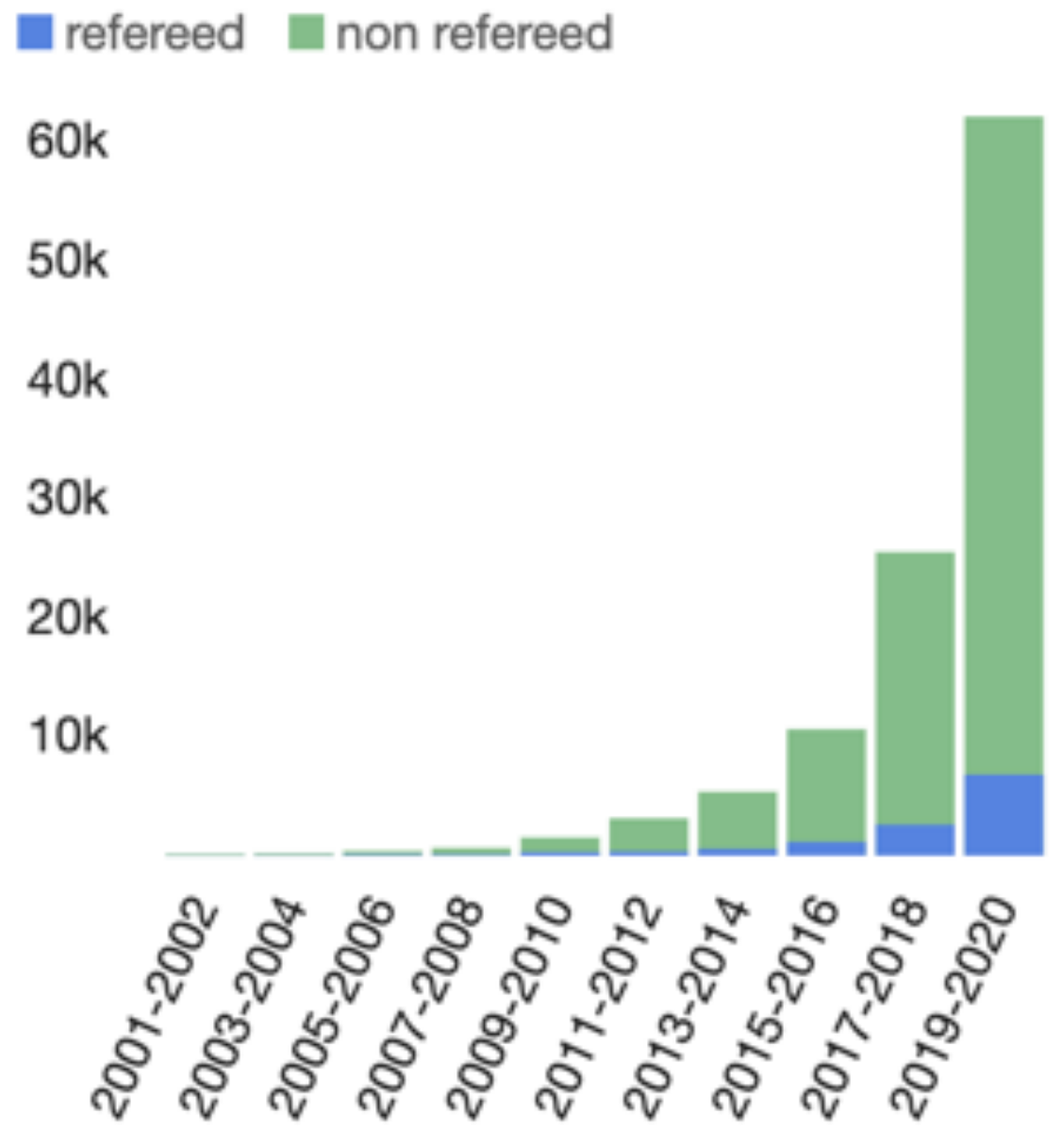




The Horizon Simulation

AI EVERYWHERE!

NUMBER OF PAPERS MENTIONING “MACHINE LEARNING”
IN THE ABSTRACT



BEFORE 2012….

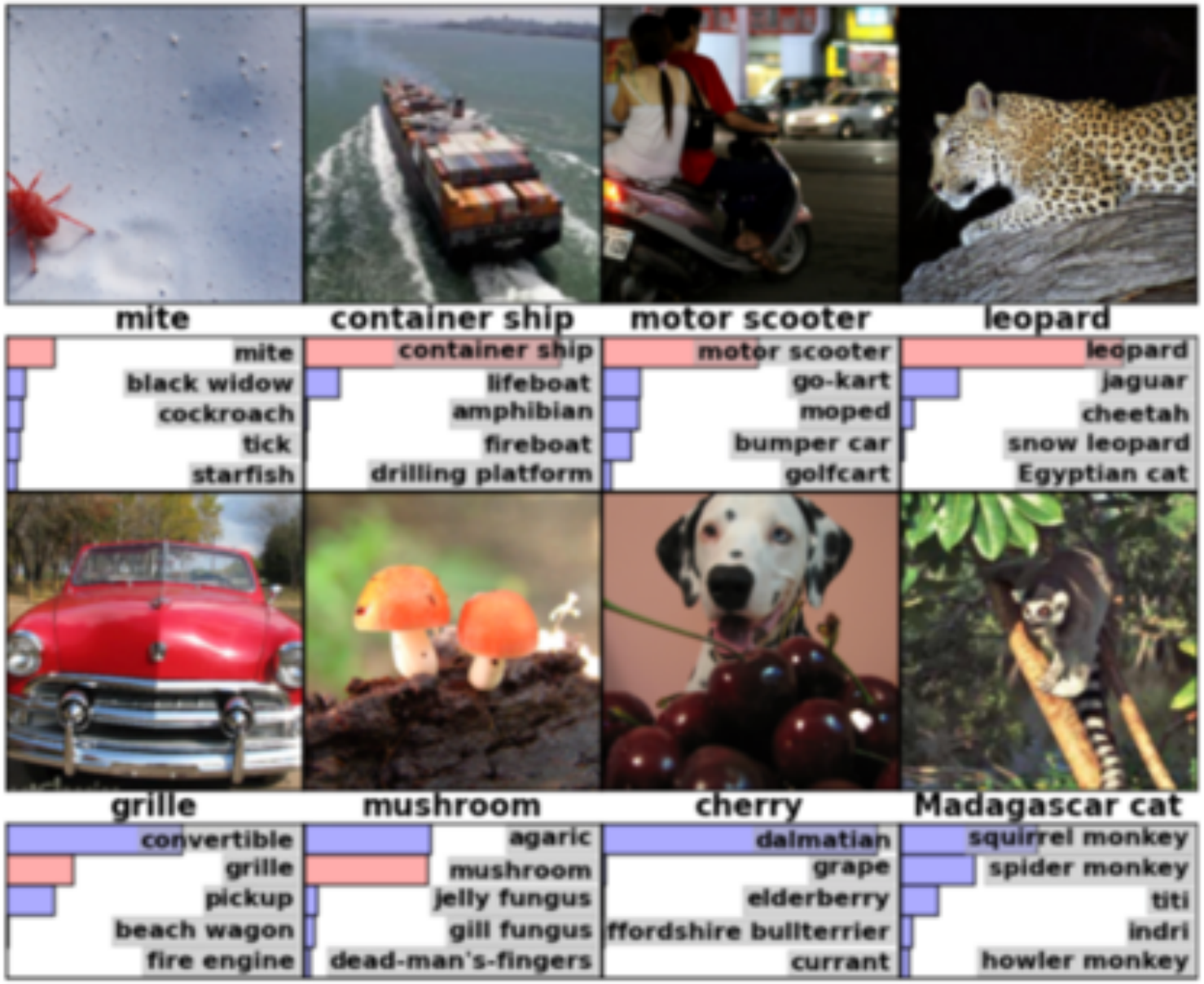
TRIVIAL HUMAN TASKS REMAINED CHALLENGING FOR COMPUTERS

CAT?

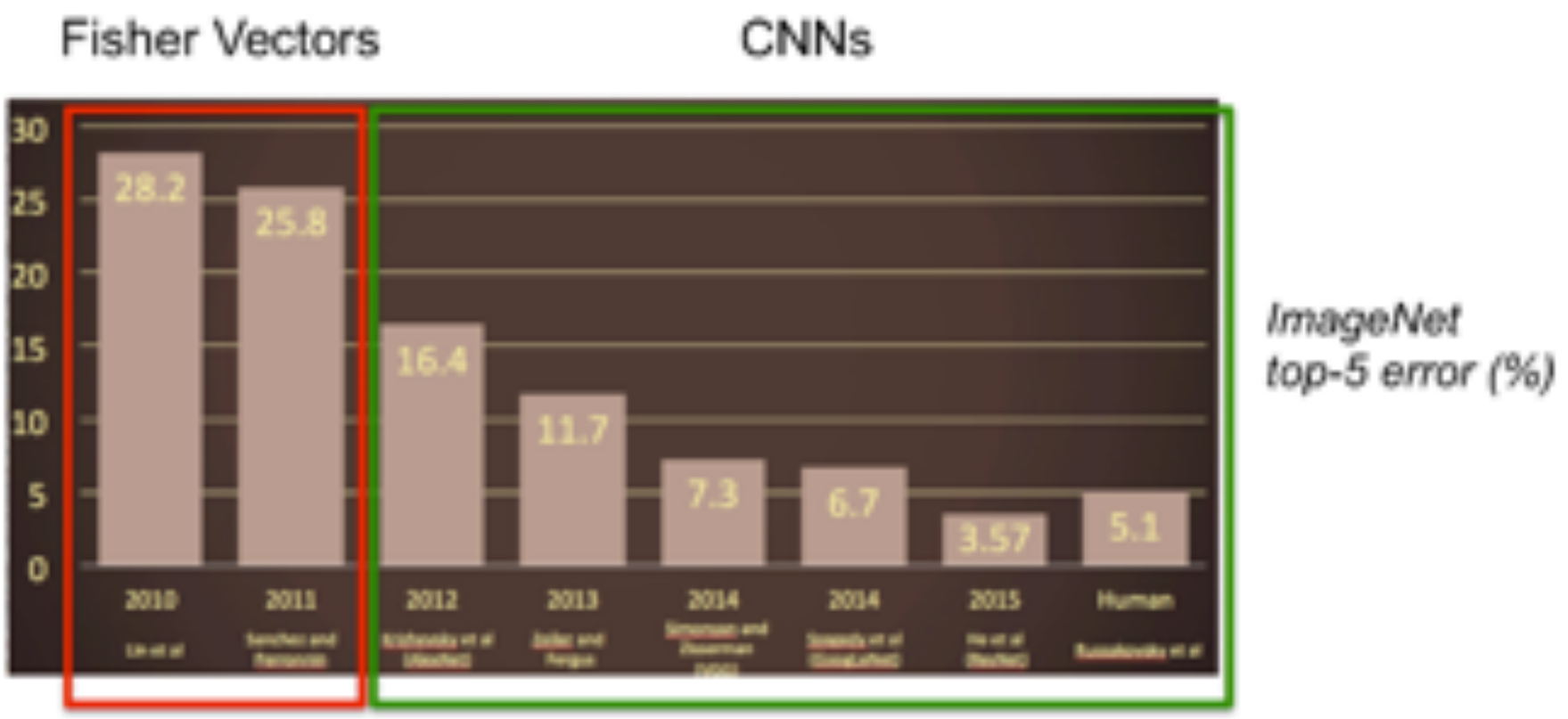
DOG?

AFTER 2012

IT HAS BECOME TRIVIAL...



THIS IS A CHANGE OF PARADIGM!



PART I

SUPERVISED CLASSIFICATION

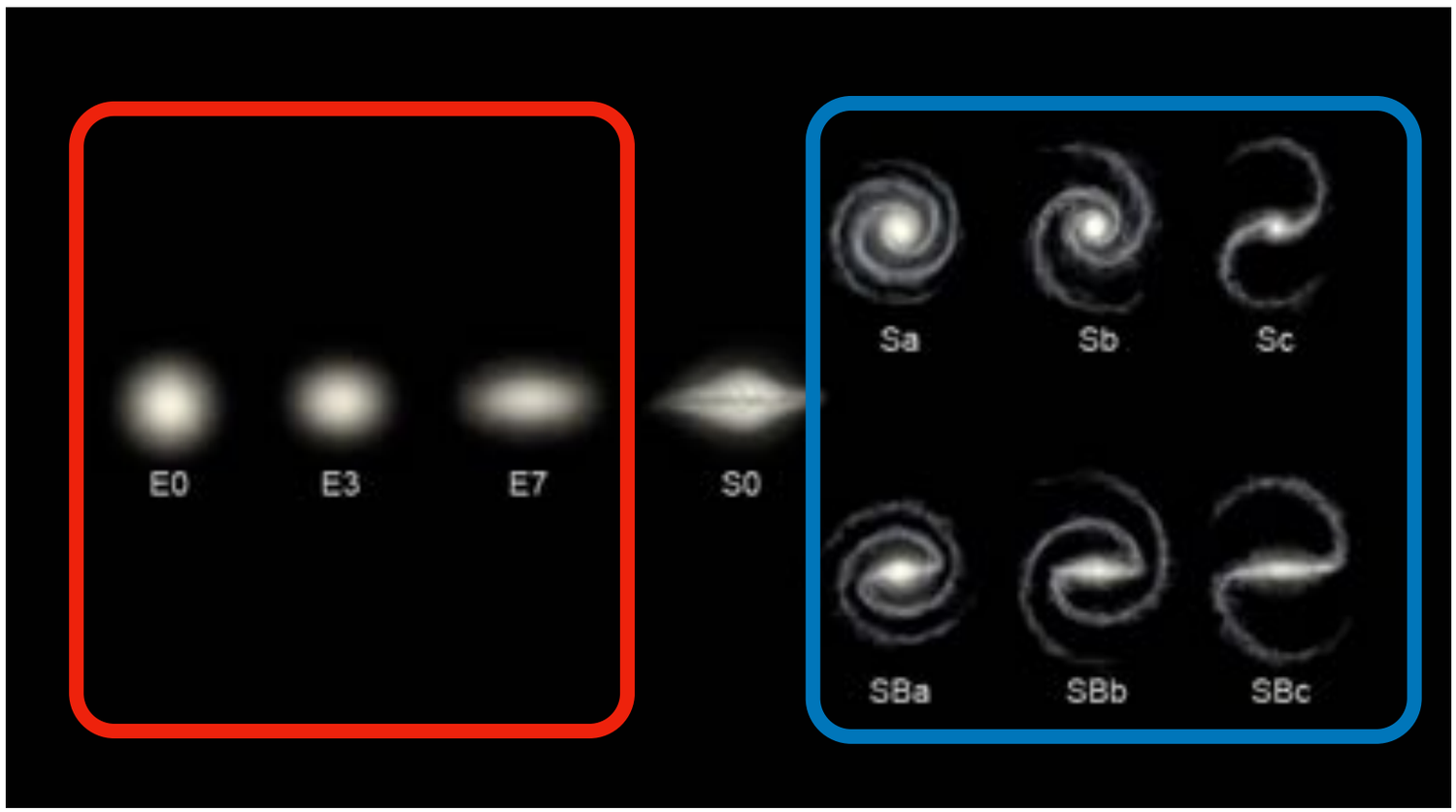
“OUR CATS AND DOGS”: GALAXY MORPHOLOGY

EARLY-TYPE GALAXIES HAVE:
- ELLIPSOIDAL SHAPES,
- OLD STARS,
- LITTLE GAS,
- LITTE ROTATION OF THEIR STARS

LATE-TYPE GALAXIES ARE:
- DISKY,
- GAS-RICH,
- COMPOSED BY YOUNG STARS,
- SUPPORTED BY ROTATION

"OBJECTS IN THE SAME BOX HAVE SIMILAR FORMATION HISTORIES"

The Hubble Sequence



ML WINTER: BEFORE THE DEEP LEARNING BOOM

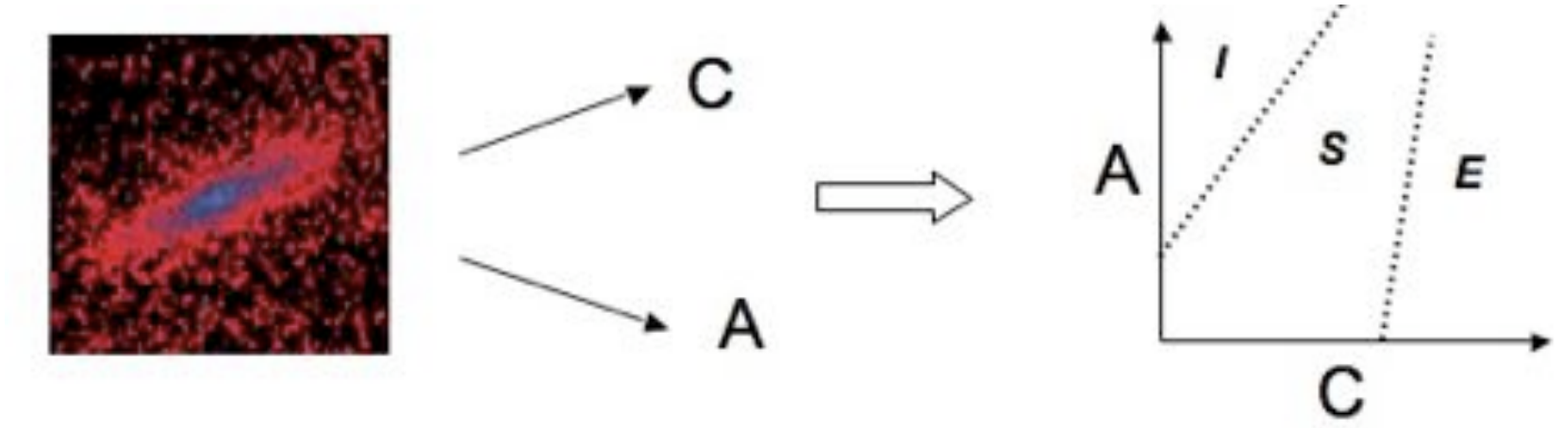


IMAGE **FEATURES** **CLASSIFICATION**

MHC+08,09,11

ML WINTER: BEFORE THE DEEP LEARNING BOOM

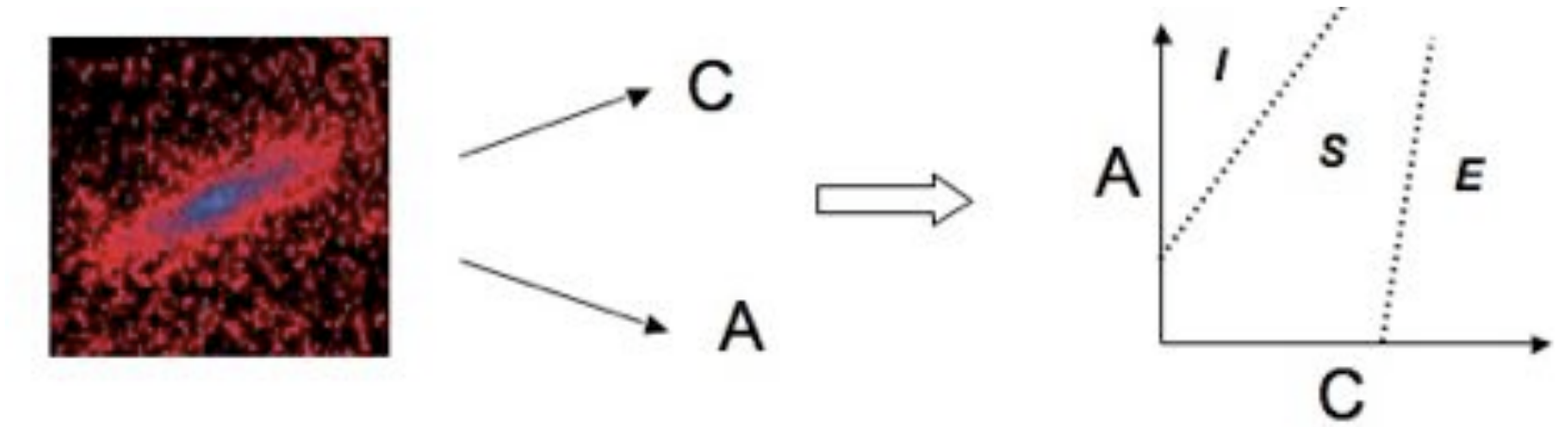


IMAGE **FEATURES** **CLASSIFICATION**

$C = 5log(r_{80}/r_{20})$

$A = \frac{1}{2}\left(\frac{\sum|I(i,j) - I_{180}(i,j)|}{\sum I(i,j)} - \frac{\sum|B(i,j) - B_{180}(i,j)|}{\sum I(i,j)}\right)$

$S = \frac{1}{2}\left(\frac{\sum|I(i,j) - I_{\pi/2}(i,j)|}{\sum I(i,j)} - \frac{\sum|B(i,j) - B_{\pi/2}(i,j)|}{\sum I(i,j)}\right)$ *Ellipticity*

Surface brightness $G = \frac{1}{\bar{X}n(n-1)}\sum_{i=1}^{n}(2i - n - 1)X_i \quad (n > 2)$

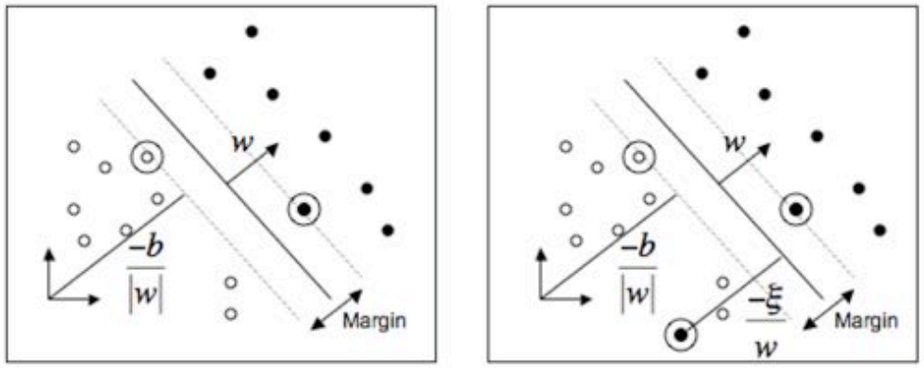
$M_{tot} = \sum_i^n M_i = \sum_i^n f_i[(x_i - x_c)^2 + (y_i - y_c)^2]$

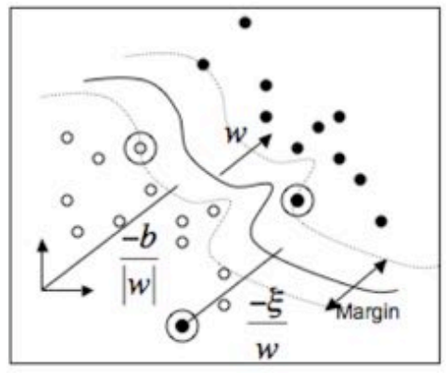
z

Magnitude

10-D space

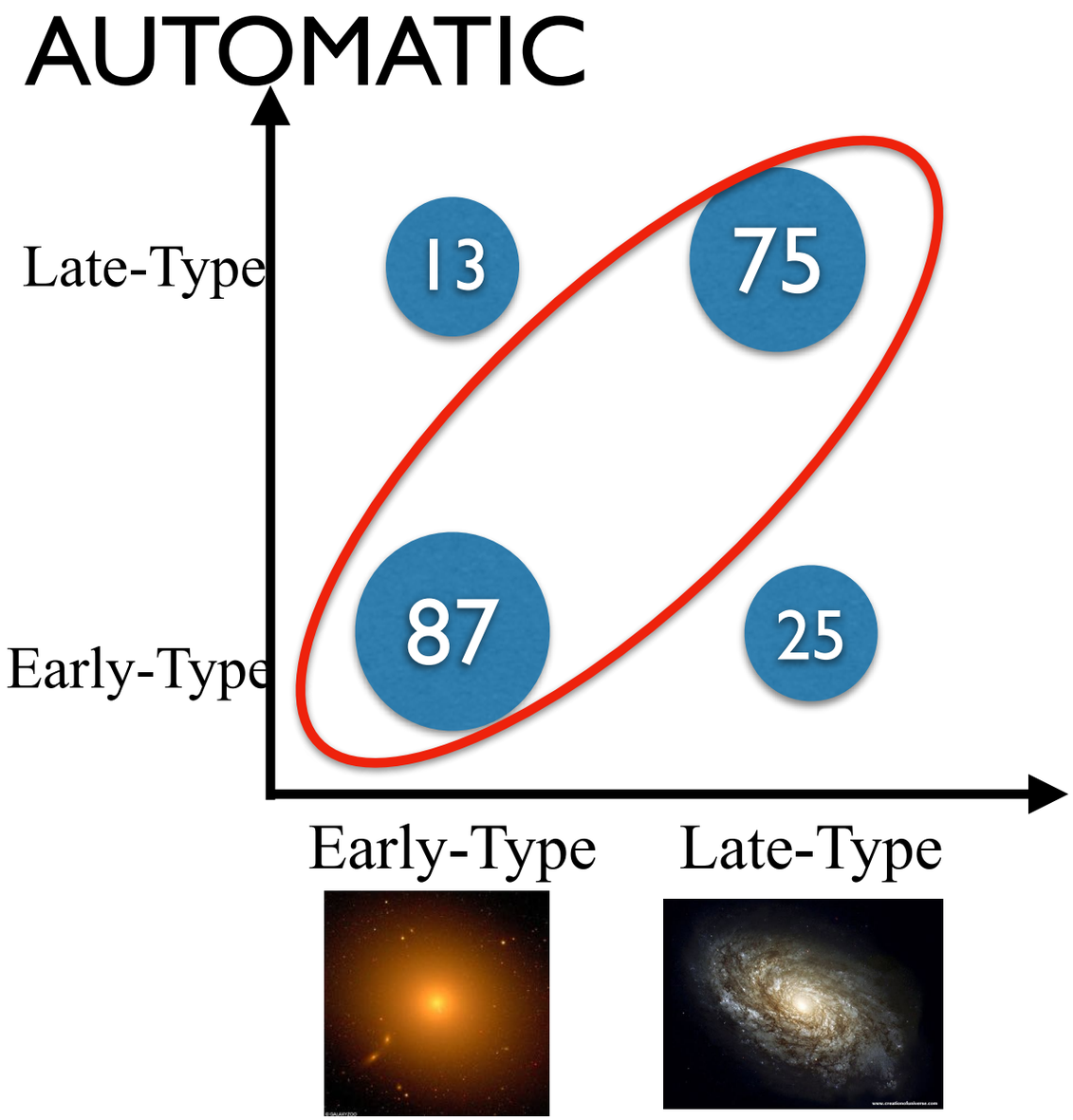
Huertas-Company+08,09,11





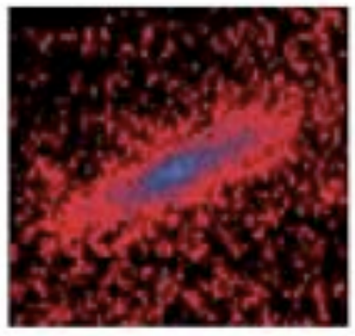
ML WINTER: BEFORE THE DEEP LEARNING BOOM

SVMs APPLIED IN THE DISTANT UNIVERSE



MHC+14

DEEP LEARNING REVOLUTION



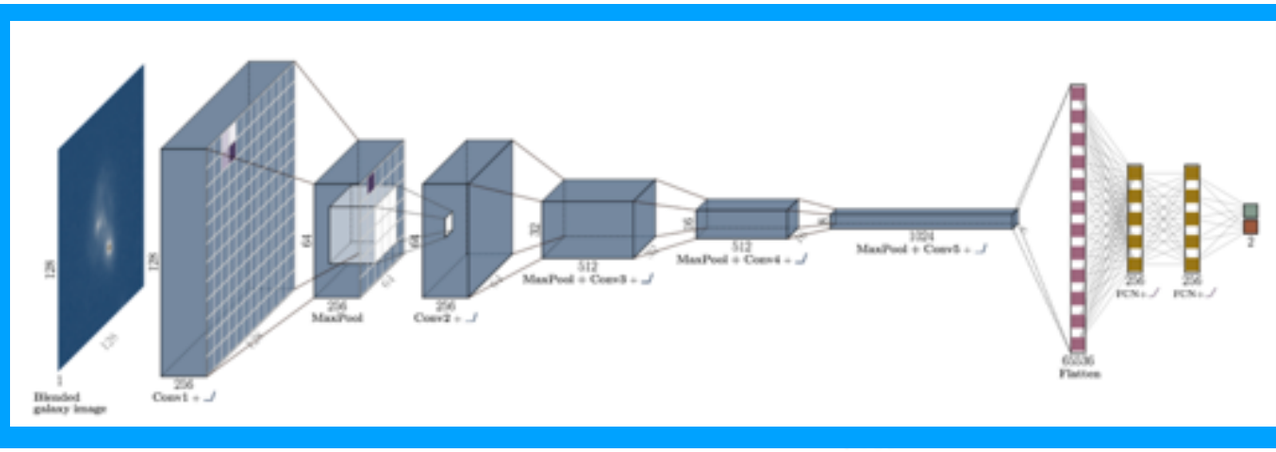
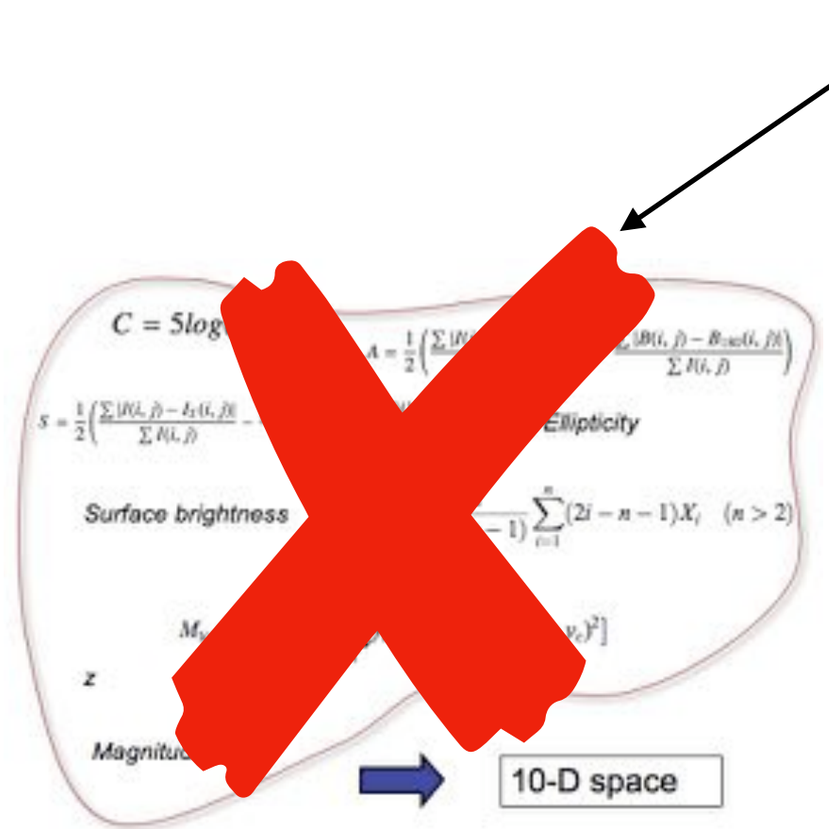
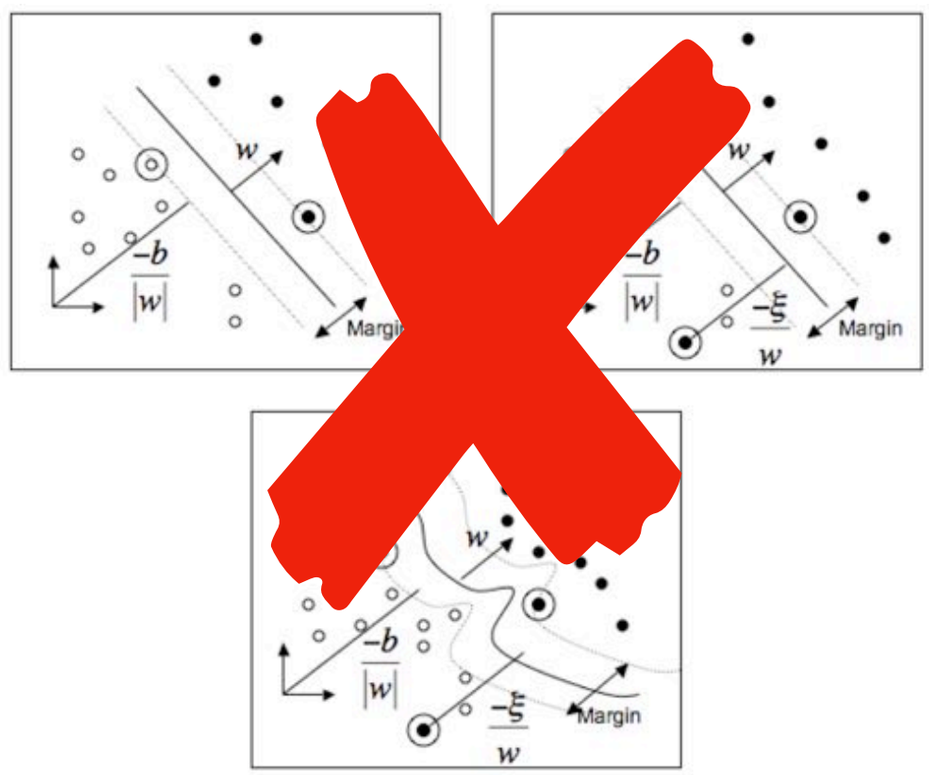


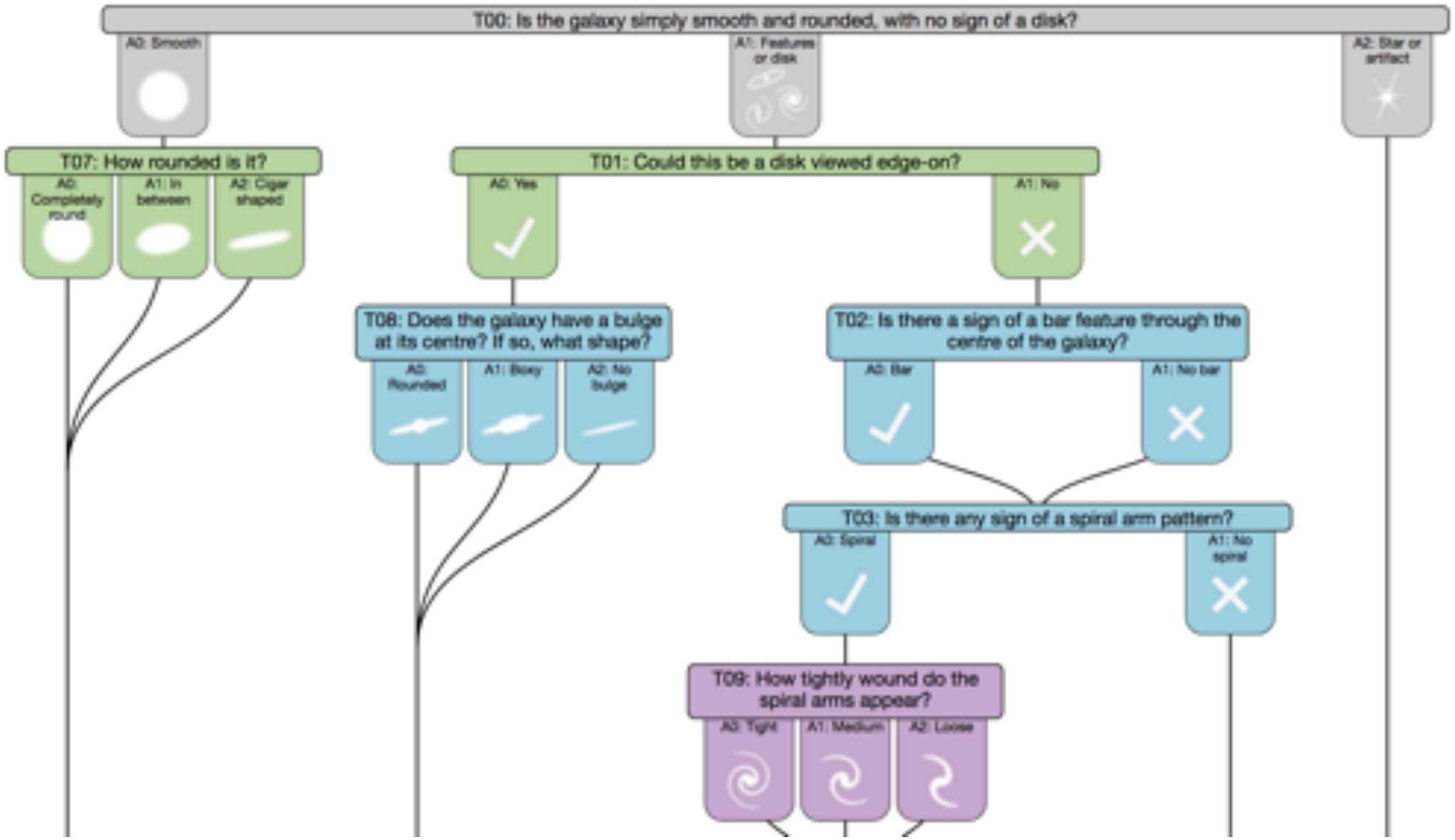
IMAGE FEATURES CLASSIFICATION





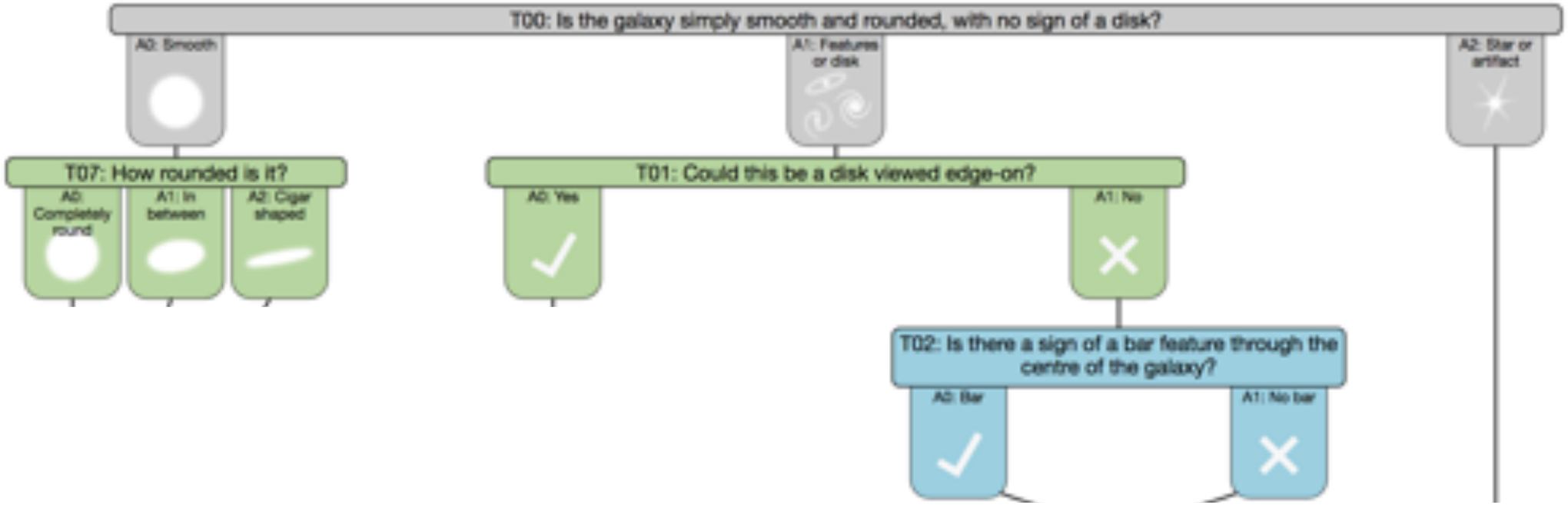
Training sample:

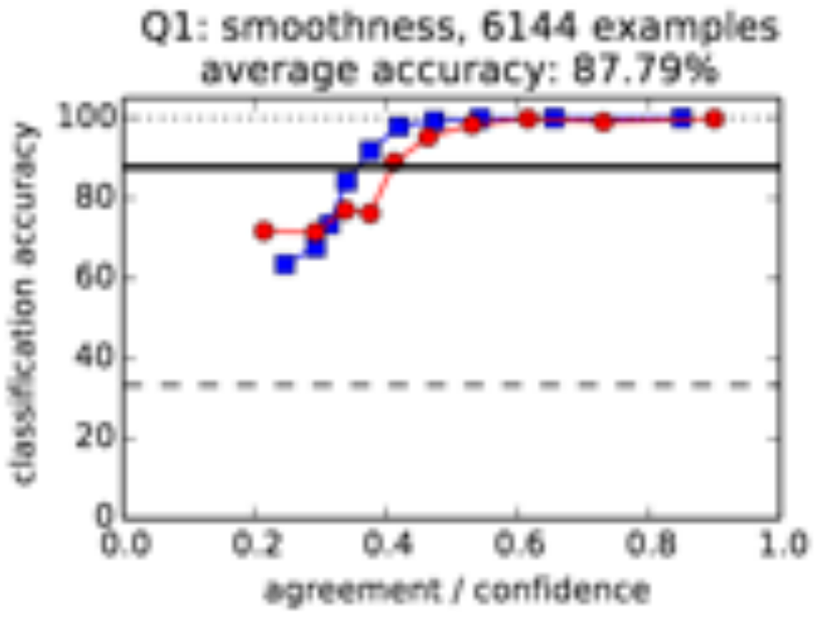
THOUSANDS OF ANNOTATED IMAGES THROUGH CITIZEN SCIENCE

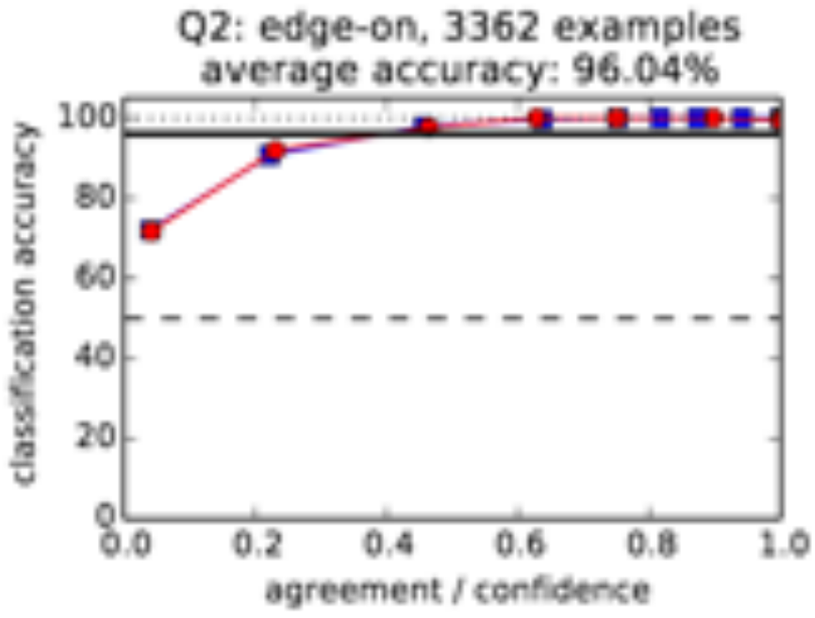


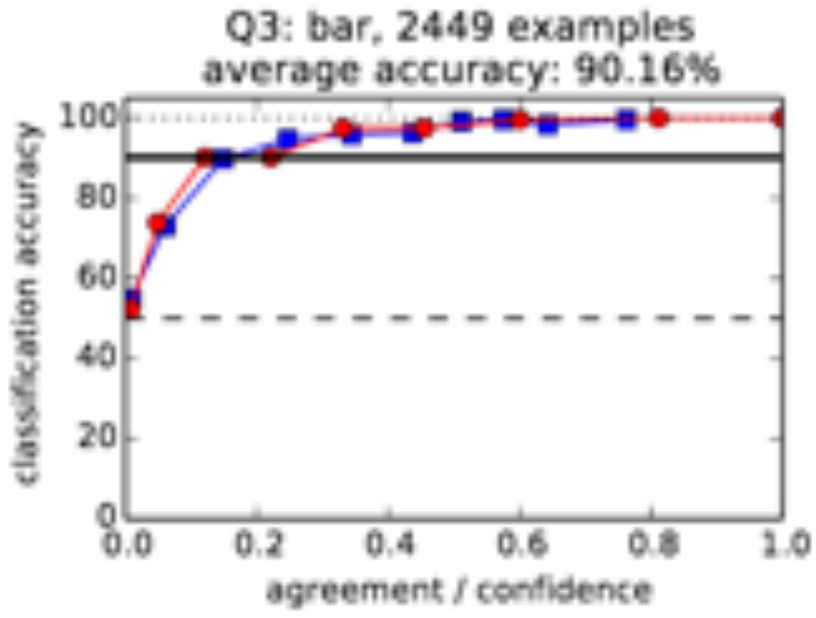
Lintott+11

DEEP LEARNING ACHIEVED UNPRECEDENTED ACCURACY



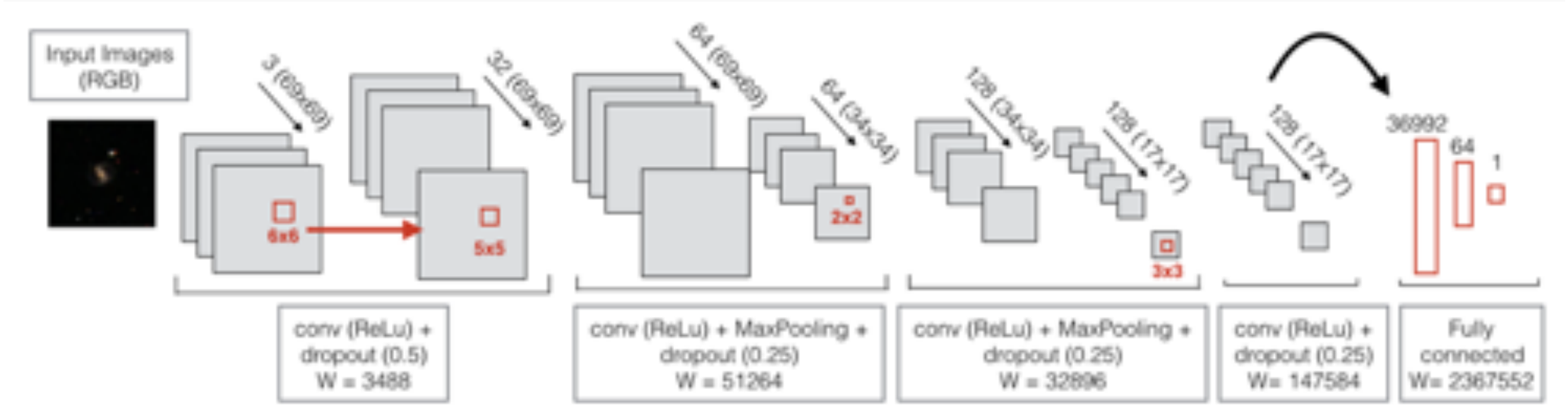






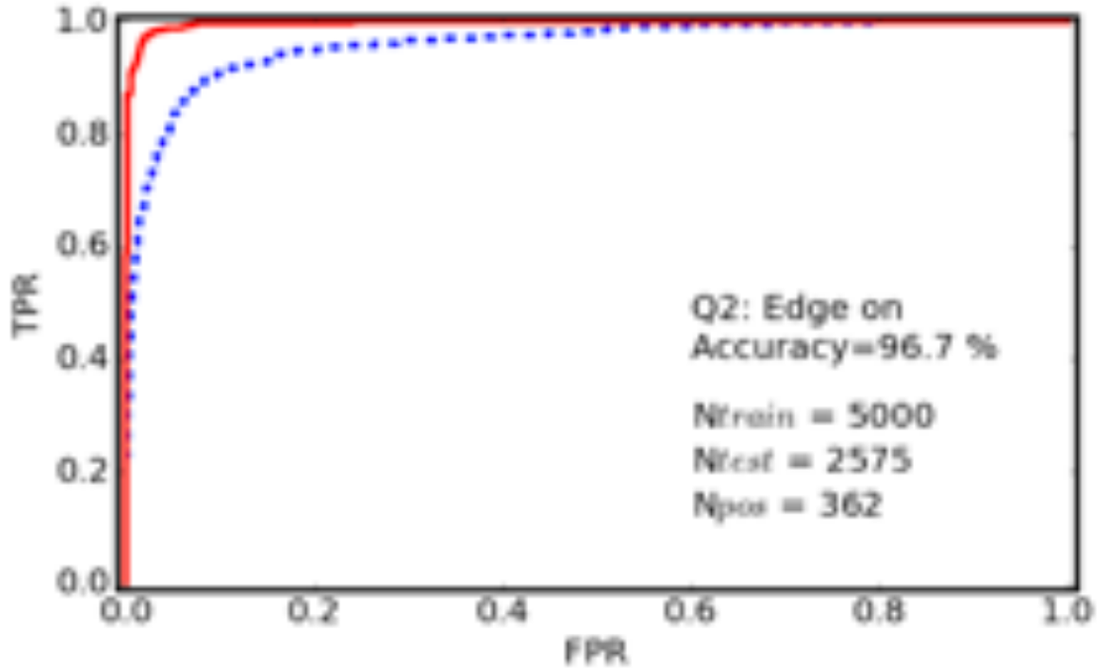
Dieleman+15

VERY SIMPLE VANILLA CNNs ARE USUALLY ENOUGH

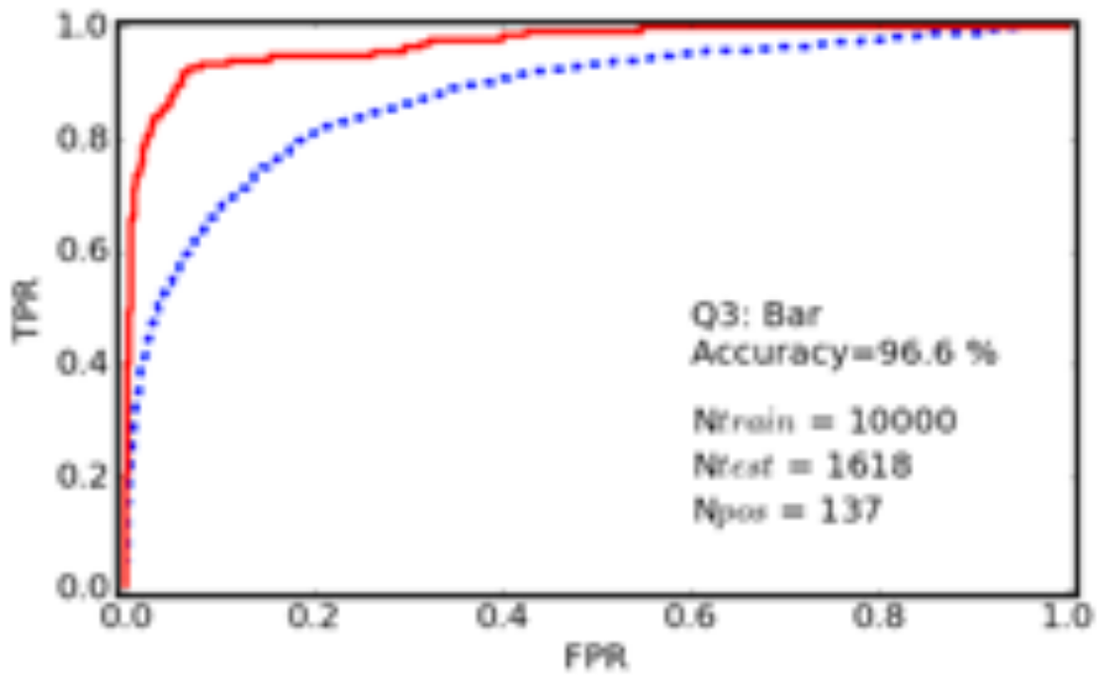


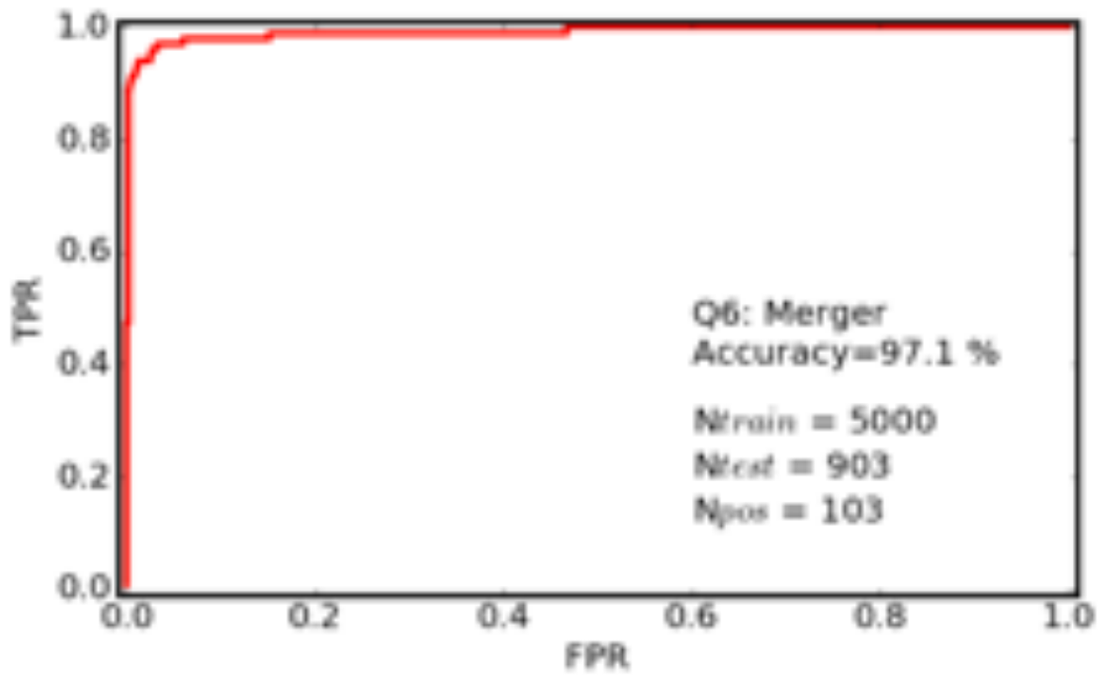
Domínguez Sánchez, MHC+18

Testing the models: ROC curve



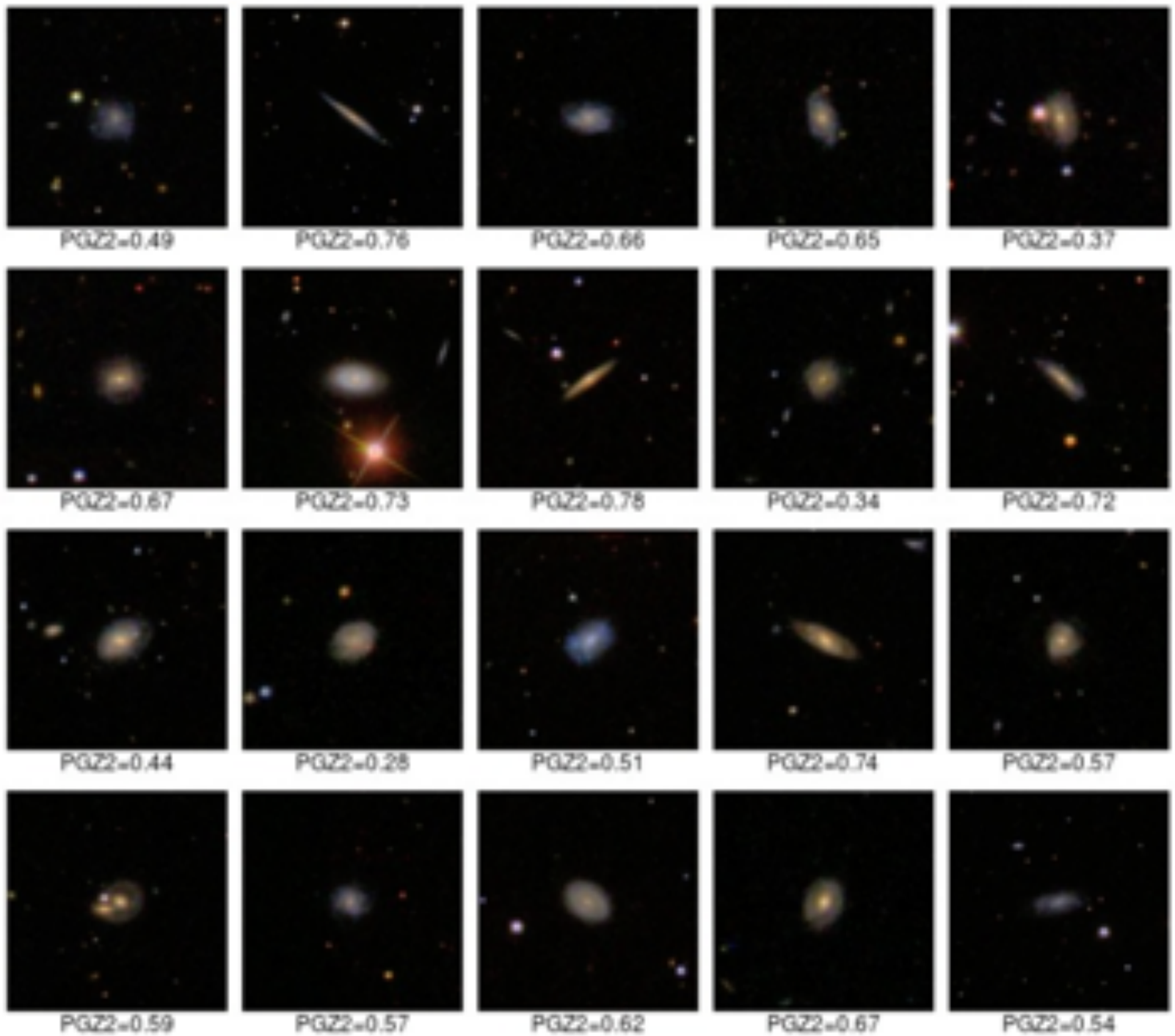
Question	Meaning	P_{thr}	TPR	Prec.	Acc.
Q1	Disk/Features	0.2	0.97	0.91	0.98
		0.5	0.95	0.96	
		0.8	0.90	0.99	
Q2	Edge-on	0.2	1.00	0.67	0.97
		0.5	0.99	0.83	
		0.8	0.92	0.95	
Q3	Bar sign	0.2	0.93	0.48	0.97
		0.5	0.79	0.80	
		0.8	0.58	0.92	
Q6	Merger signature	0.2	0.98	0.54	0.97
		0.5	0.96	0.82	
		0.8	0.90	0.97	





Domínguez Sánchez, MHC+18

SECURE DISK GALAXIES FOR DL - UNCLEAR FOR PEOPLE

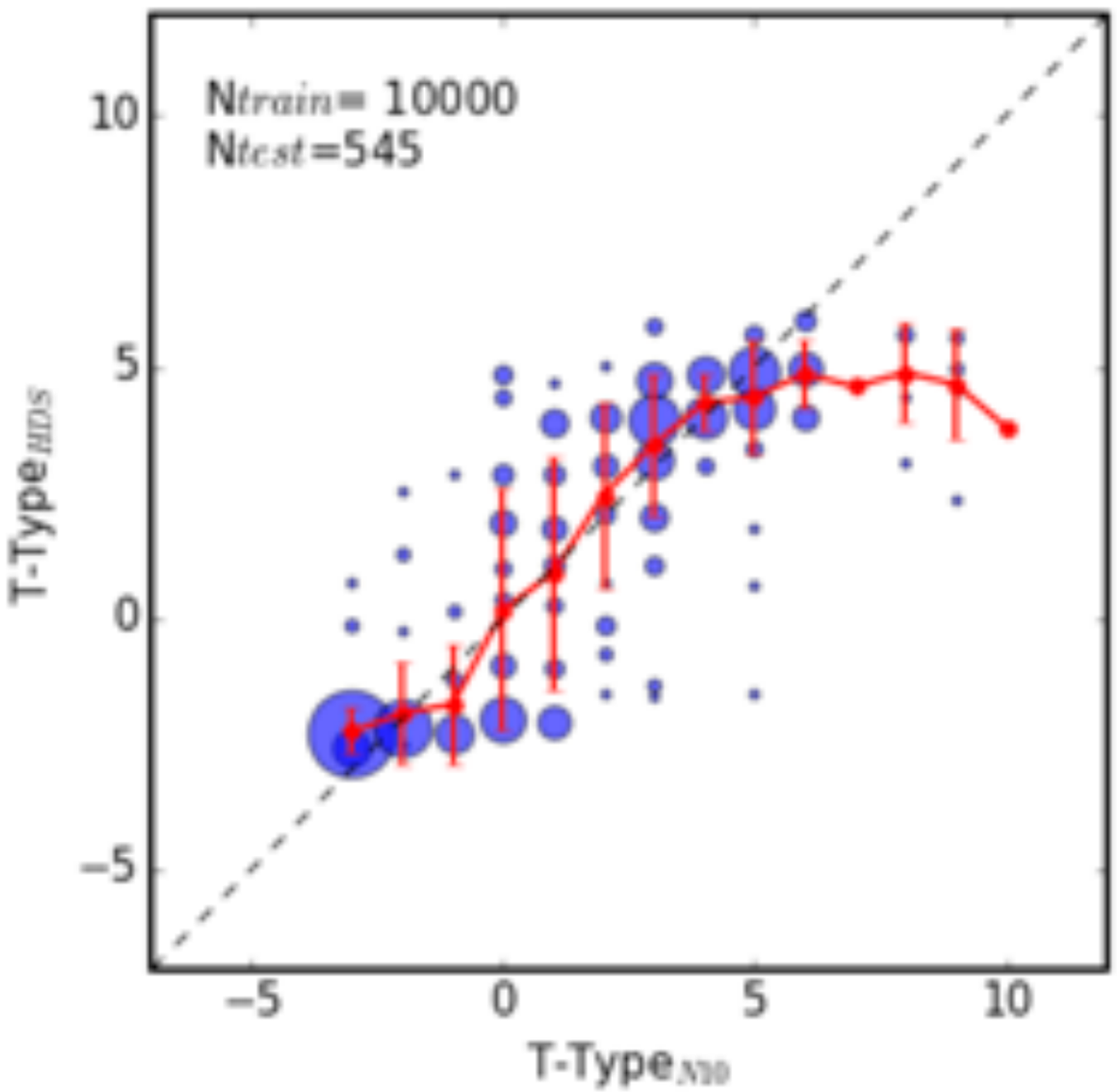


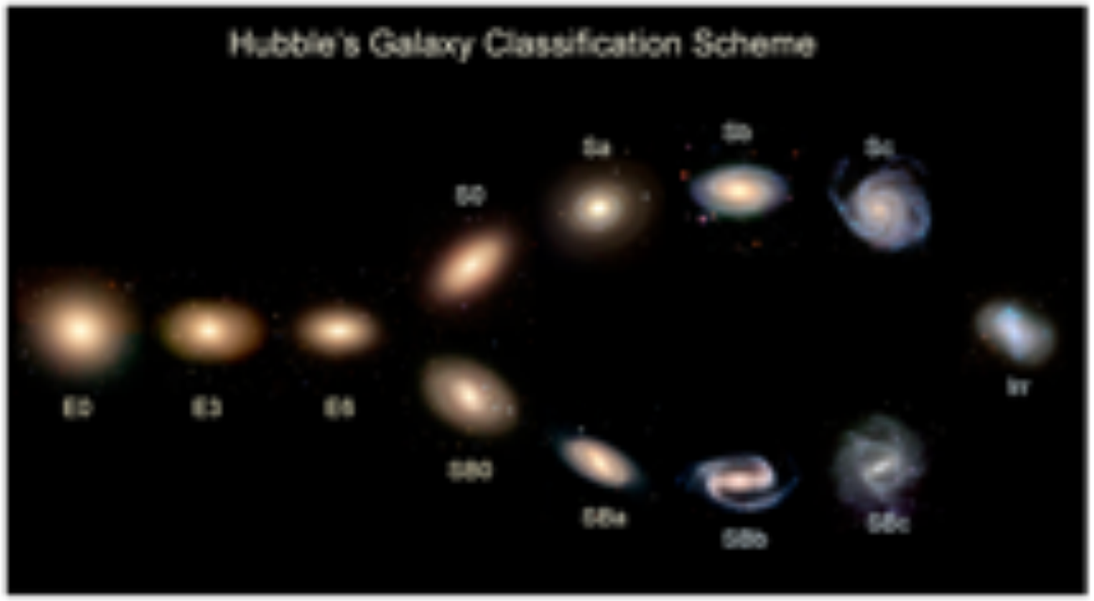
Domínguez Sánchez, MHC+18

T-Type models: Regression

Training catalogue: *Nair+2010*

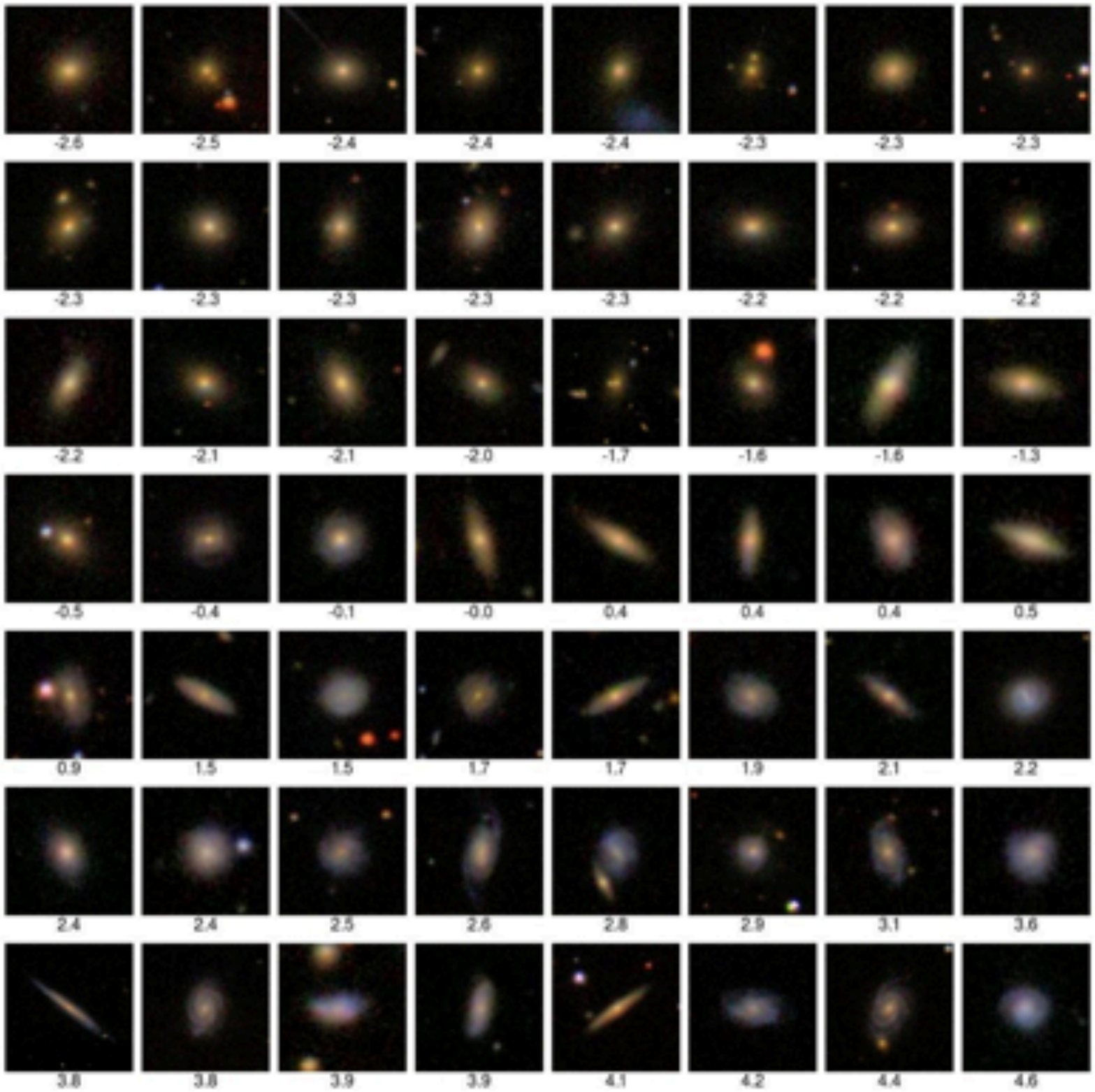
✓Regression mode (-3,10)





Domínguez Sánchez, MHC+18

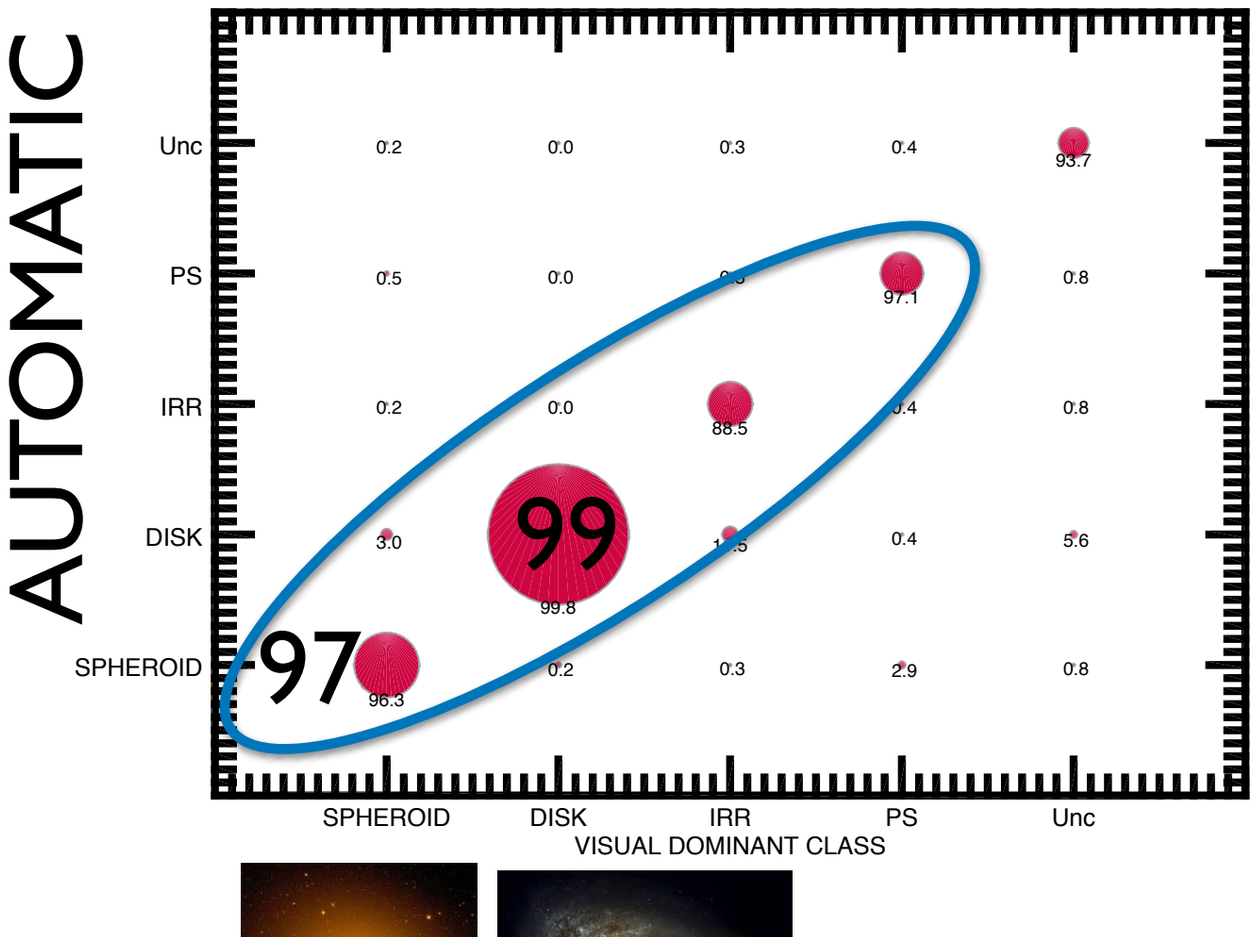
T-Type models: Regression



MORPHOLOGIES OF GALAXIES IN THE DISTANT UNIVERSE

CNNs

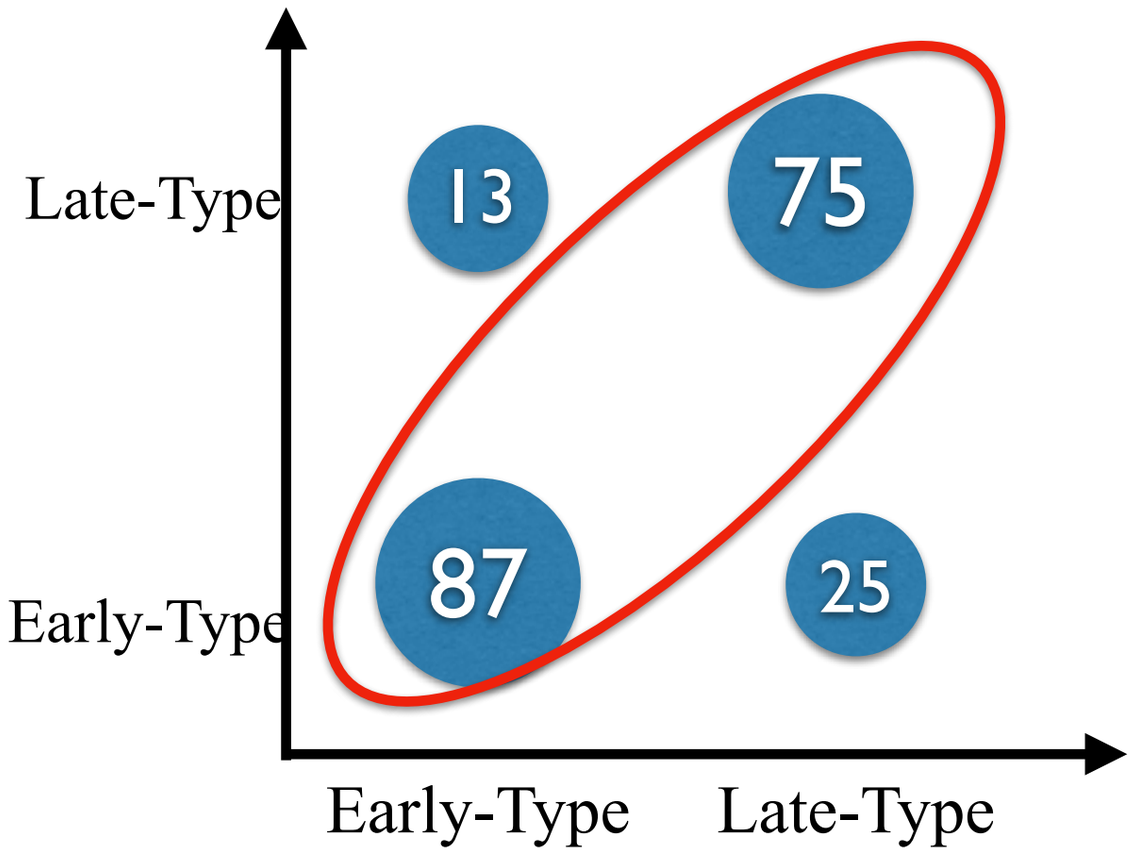
[DEEP LEARNING]



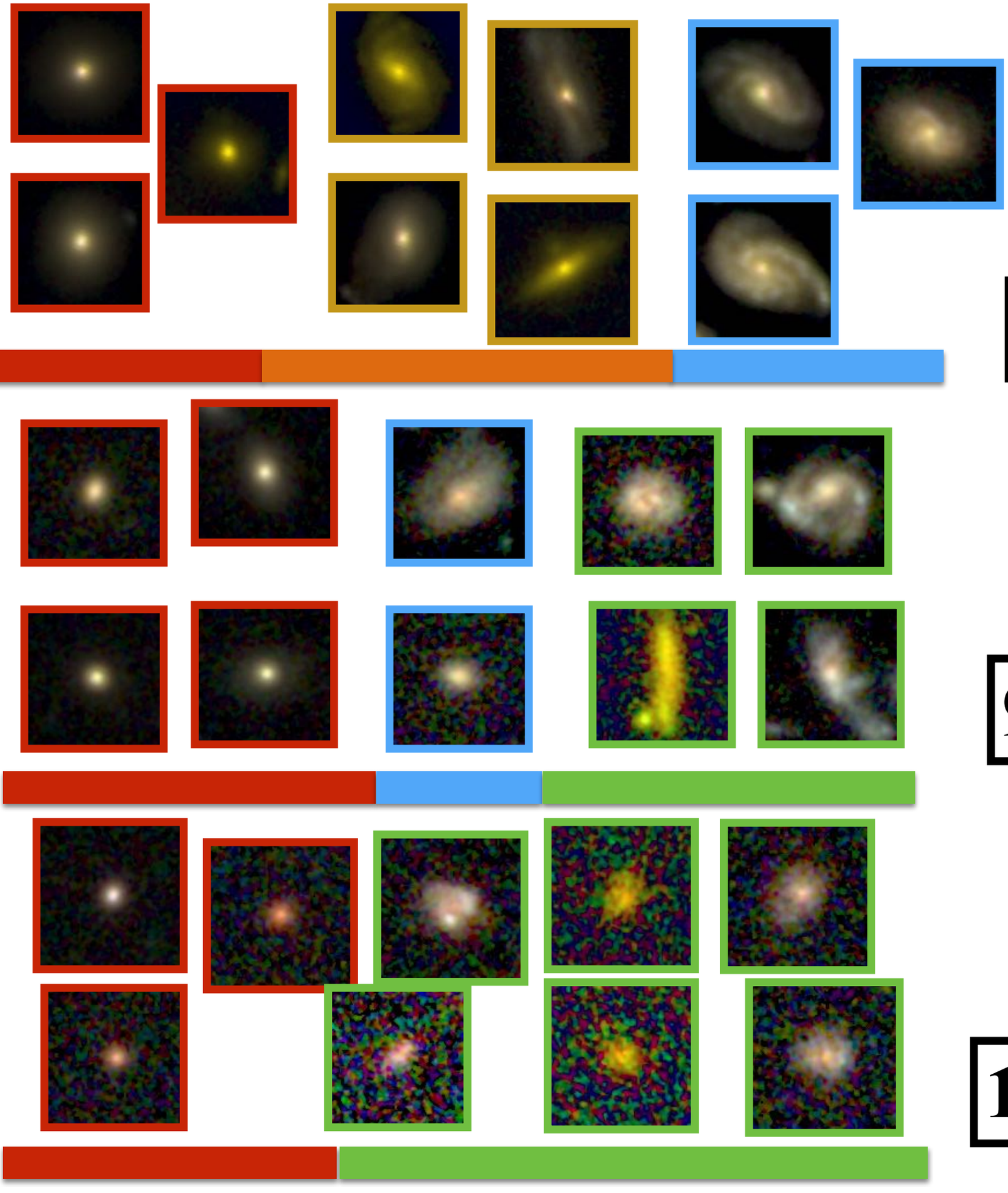
MHC+15b

SVMs

[BEFORE DEEP LEARNING]



MHC+14



~M* galaxies

MHC+16

5 billion years ago

9 billion years ago

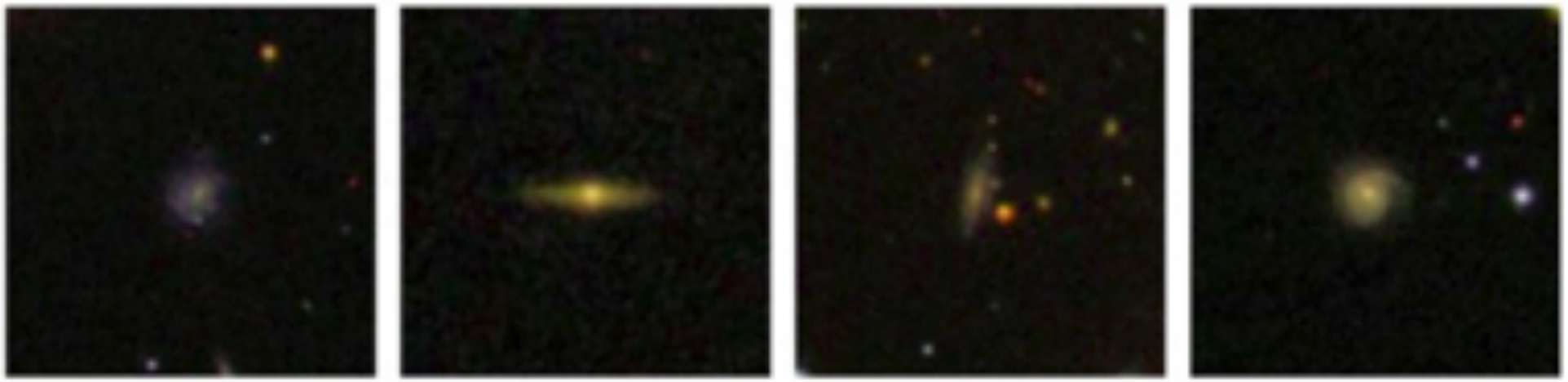
11 billion years ago

DOES DEEP LEARNING SOLVE THE PROBLEM OF GALAXY MORPHOLOGICAL CLASSIFICATION?

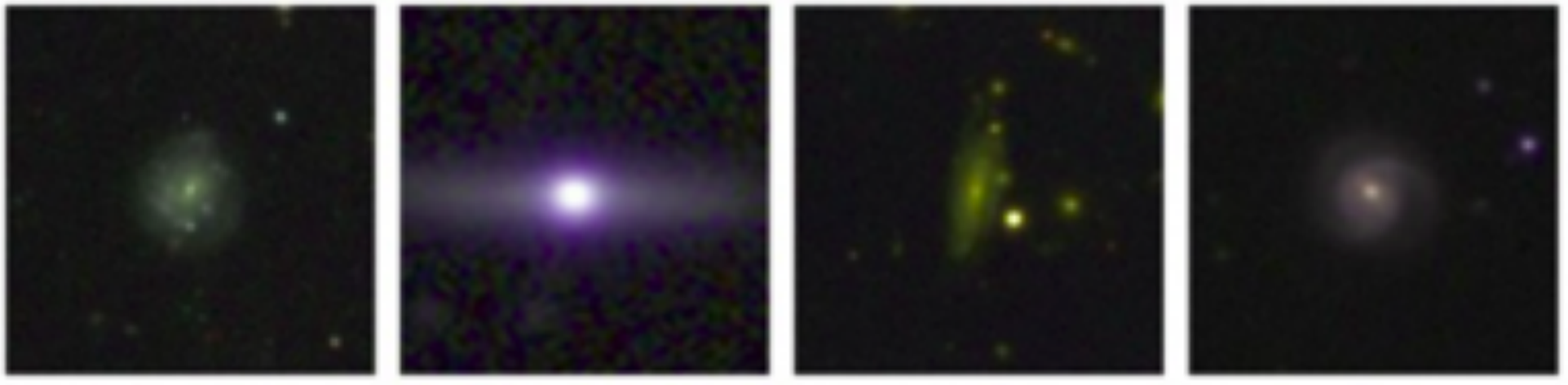
…WE STILL NEED **TRAINING SETS**!

IMAGES OF THE SAME GALAXIES TAKEN WITH DIFFERENT TELESCOPES / CAMERAS LOOK DIFFERENT!

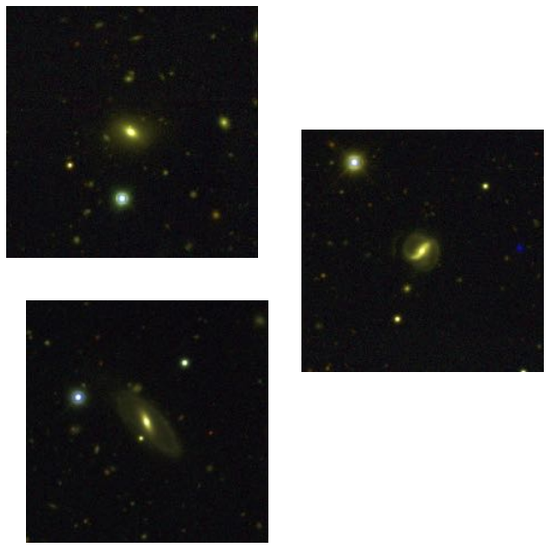
SDSS



DES



Domínguez Sánchez, MHC+19a



DATA FROM
NEW DATASET

TRANSFER
KNOWLEDGE

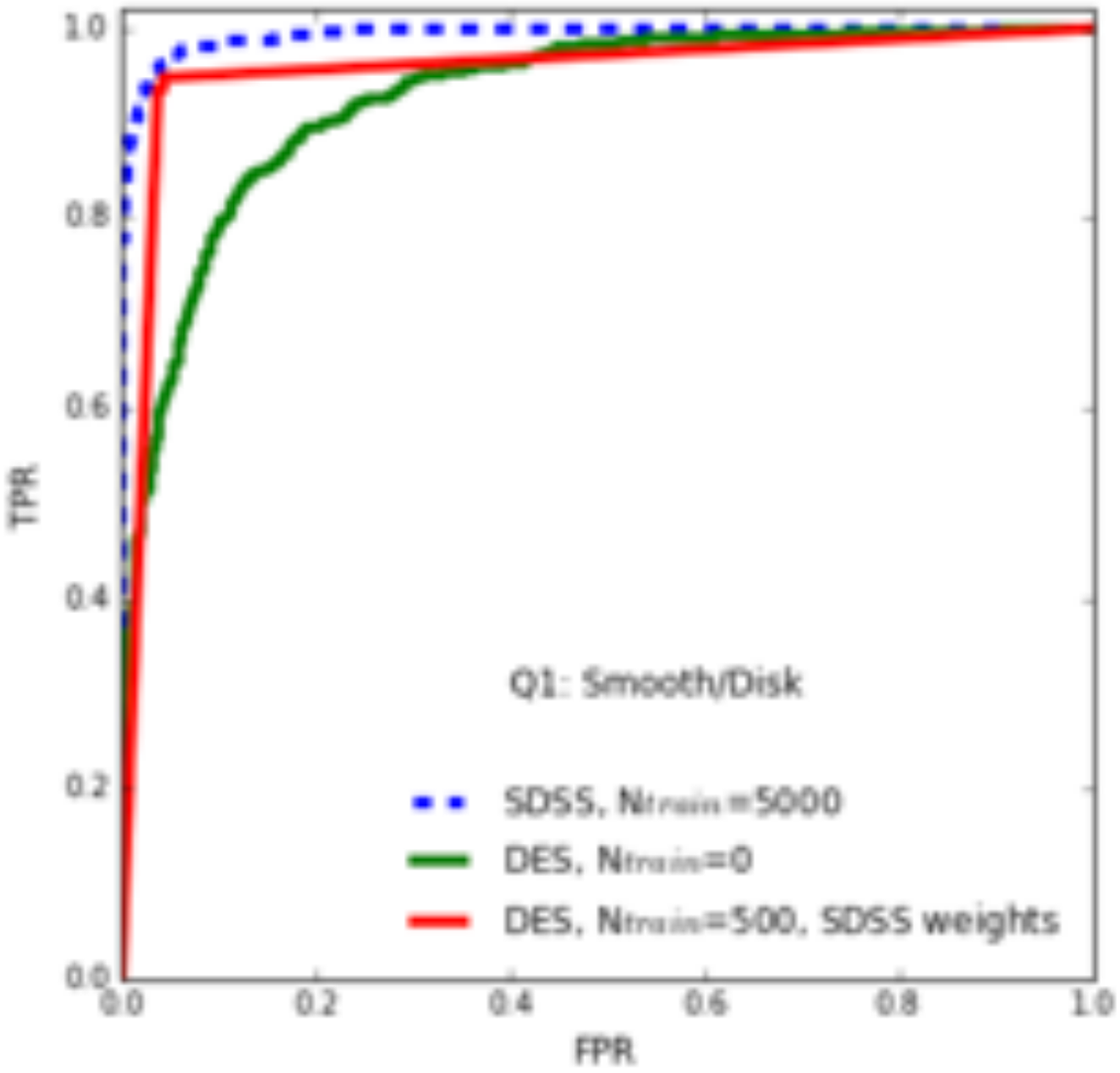
DEEP-LEARNING
BASED
MACHINE

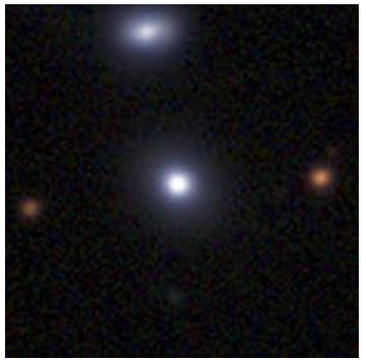
Classifications for
for the entire
sample with
reduced human
labelling

Human classifications
from existing dataset

Domínguez Sánchez, MHC+19a

Transfer Learning for galaxy morphology

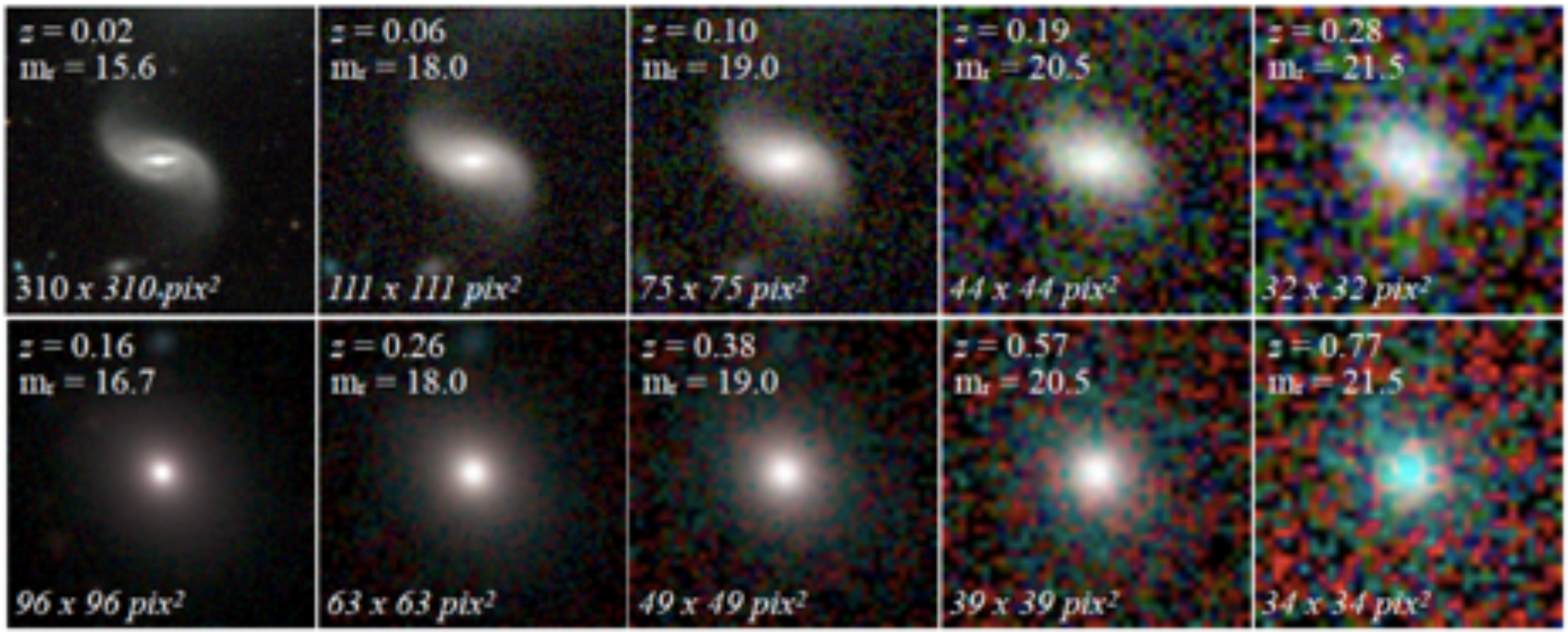




"Recycling" features/ weights learned from a different sample reduces the training sample by one order of magnitude

BUT WHAT HAPPENS IF THERE IS LITTLE OVERLAP BETWEEN SURVEYS AND/OR WE WANT TO GO DEEPER?

Simulating DES galaxies

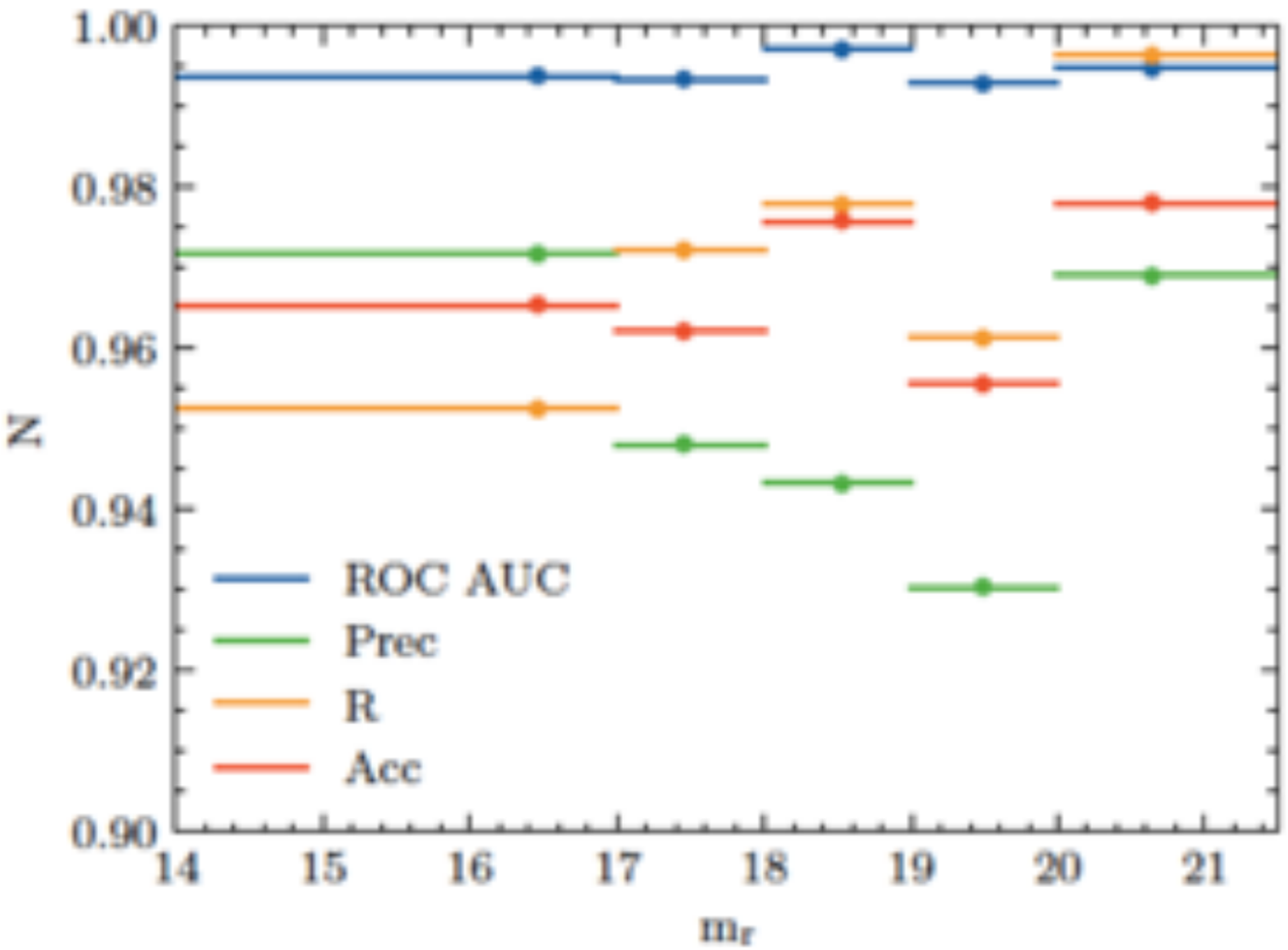


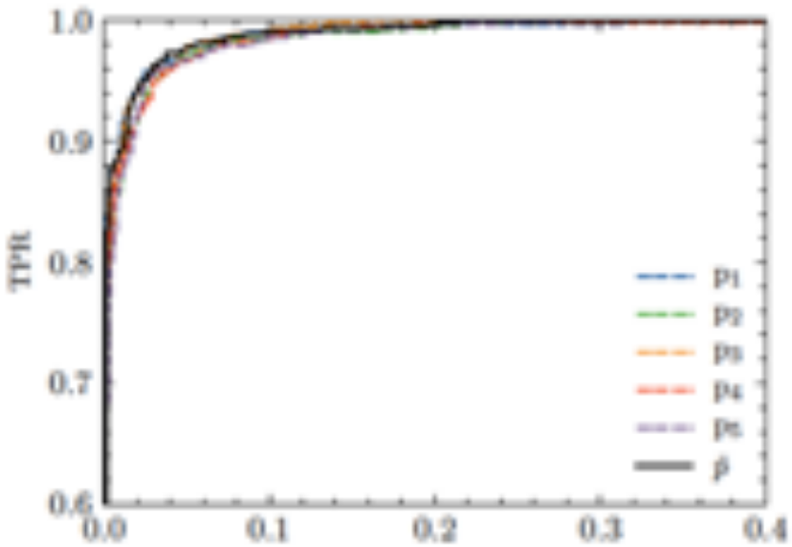
- ✓ Cosmological dimming: flux and size (N pixels)
- ✓ k-correction + evolution
- ✓ PSF +noise

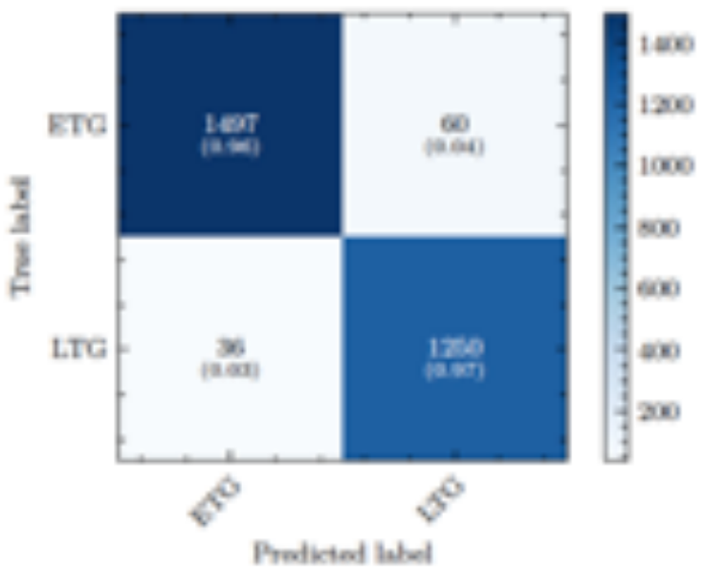
Vega-Ferrero, DS+21

ETGs vs LTGs

Are the results affected
by observed magnitude?





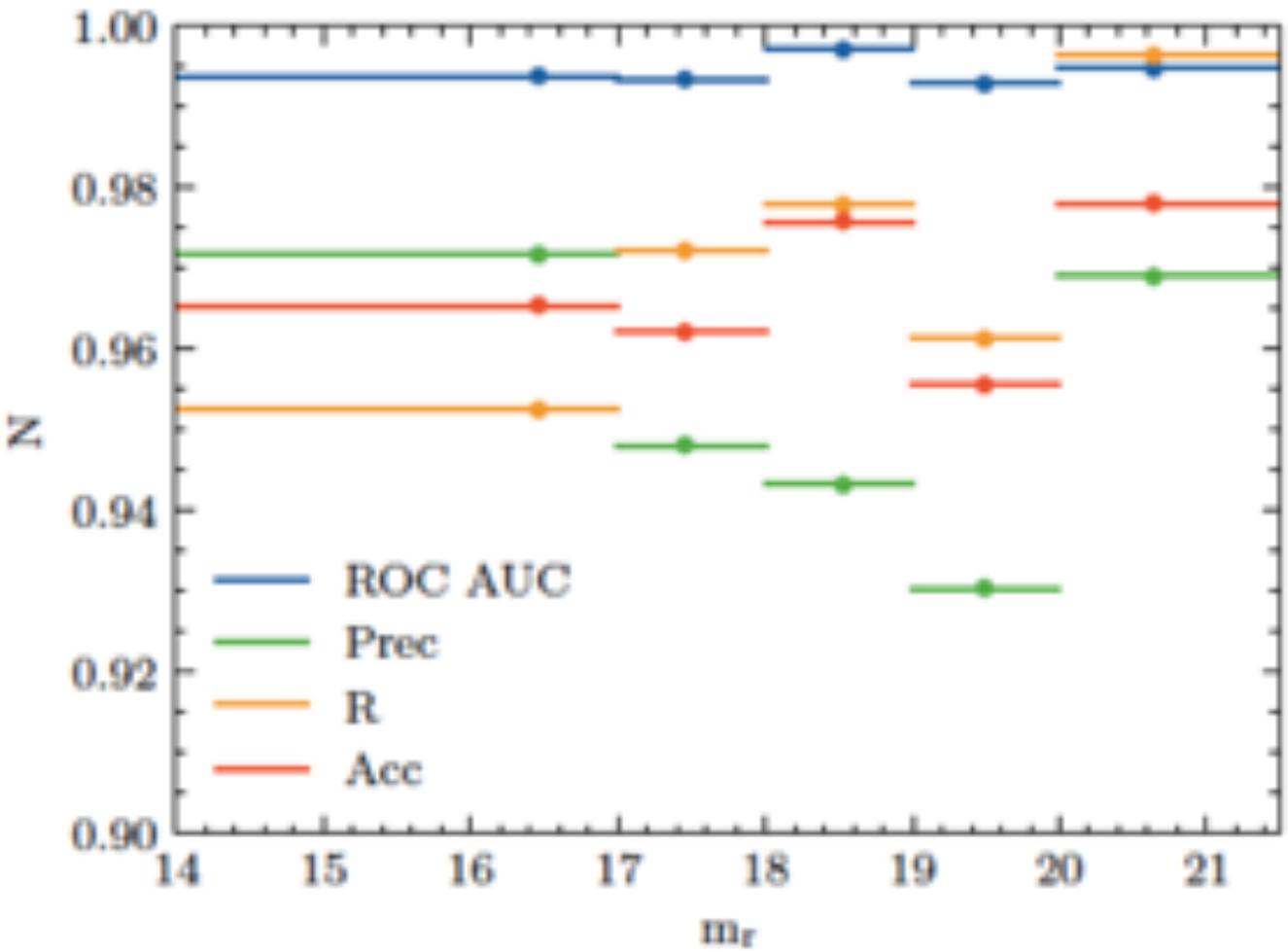


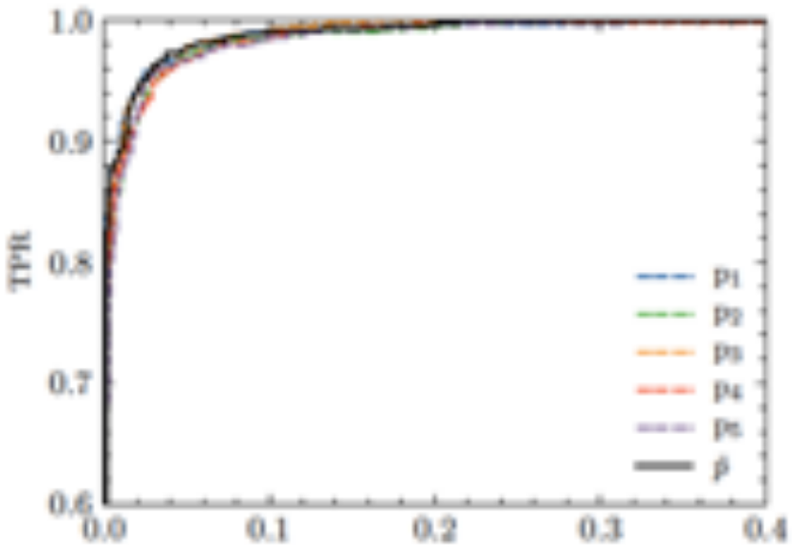
Vega-Ferrero, DS+21

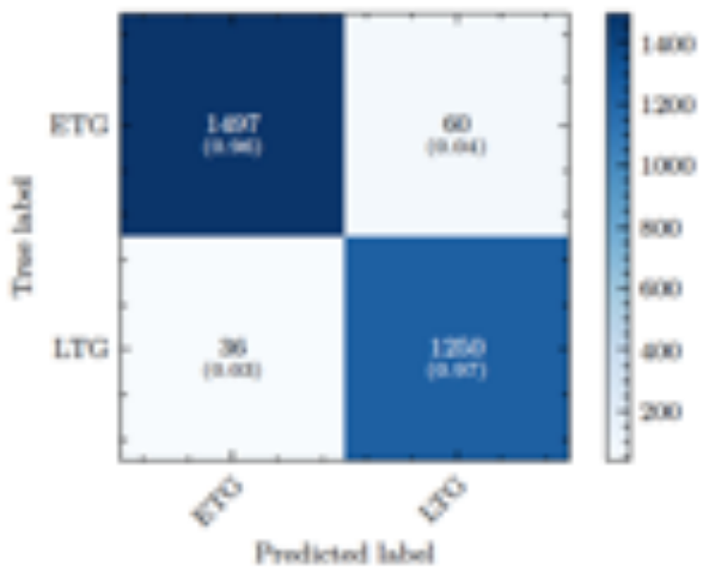
ETGs vs LTGs

ACCURACY BEYOND THE HUMAN EYE! (SEE ALSO MHC+18)

Are the results affected by observed magnitude?



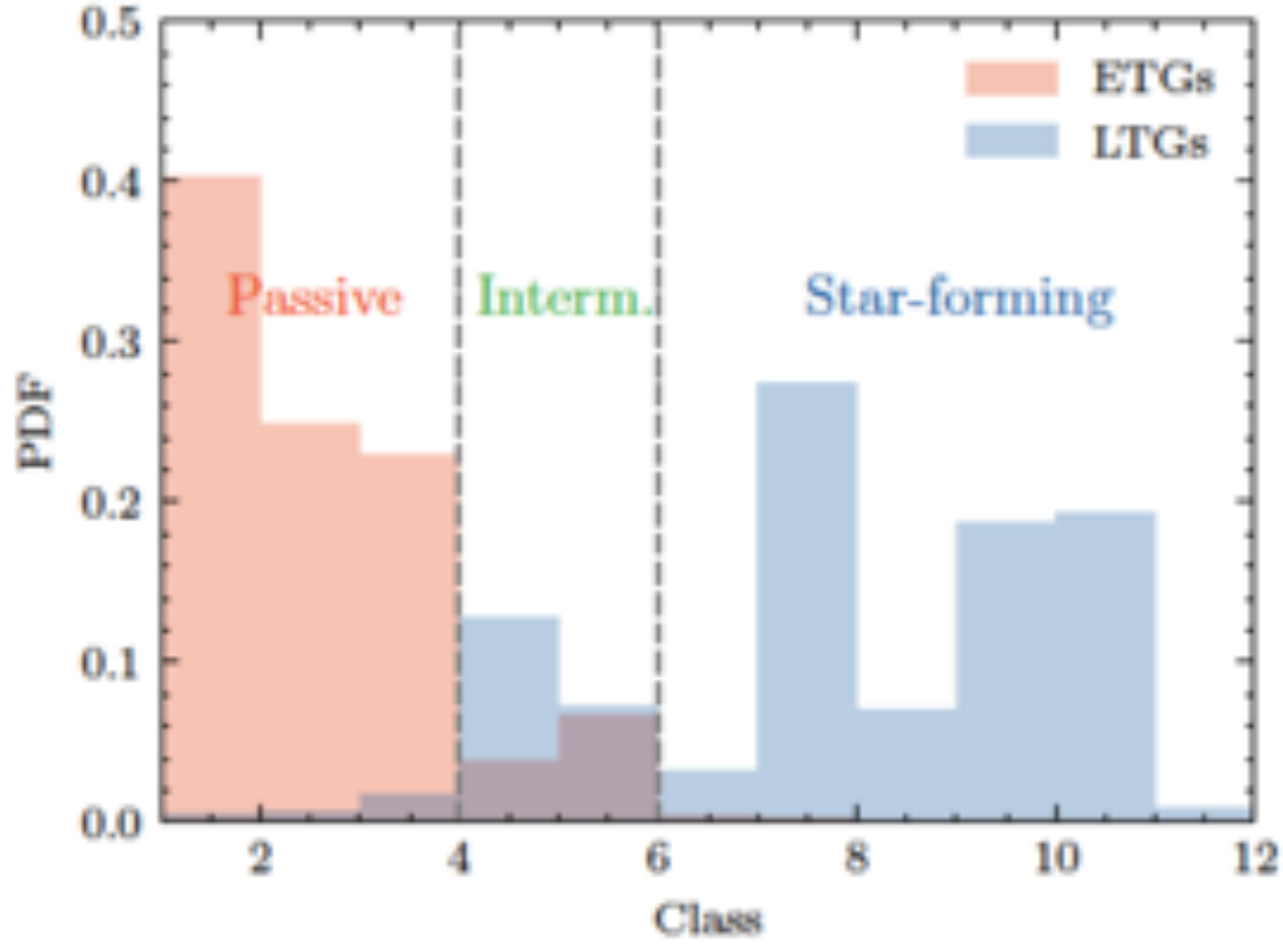




Vega-Ferrero, DS+21

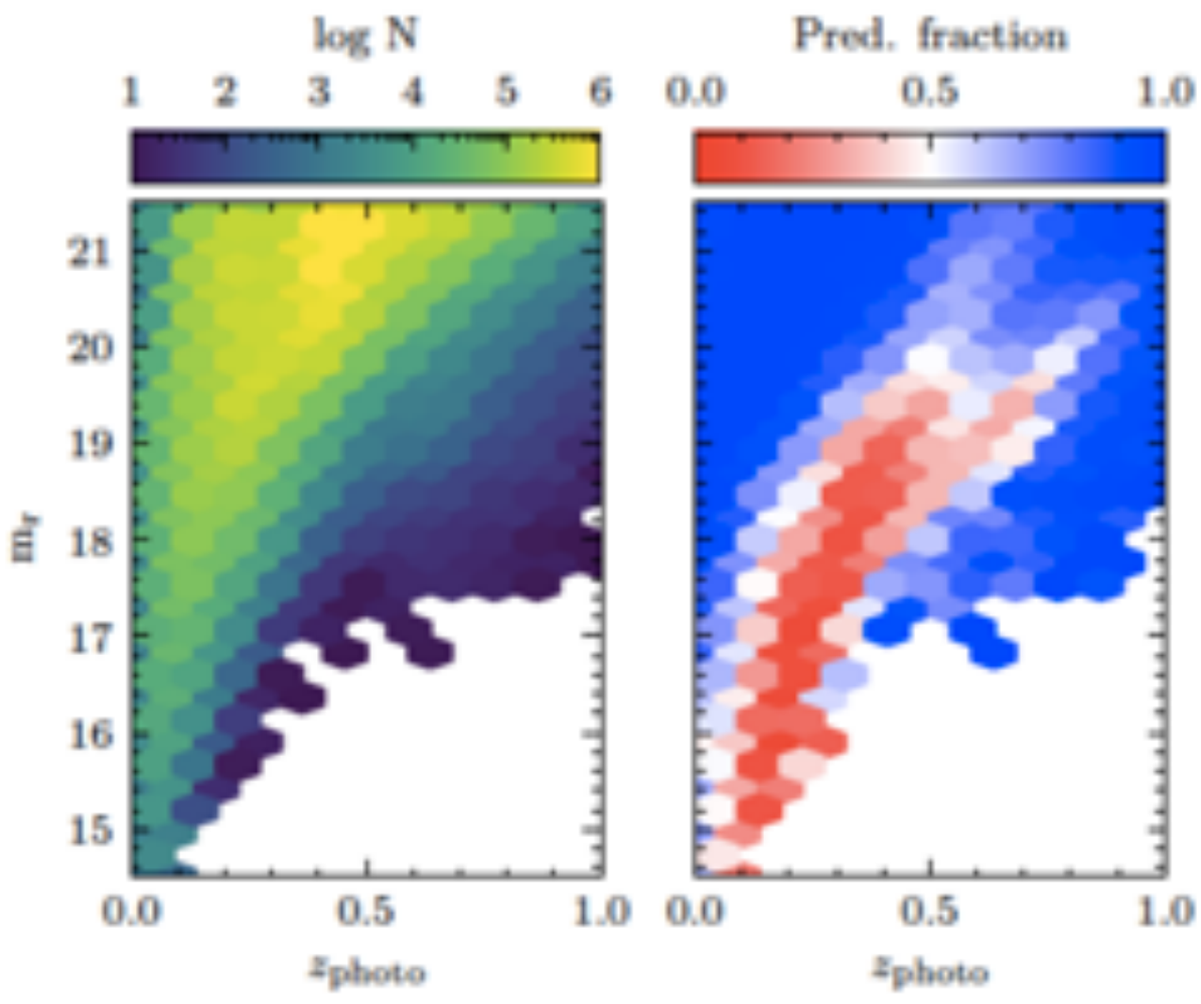
VIPERS spectral class

Vega-Ferrero, DS+21



✓EGT and LTG clearly correlate with spectral class
✓97 % TP (LTG with class > 4) and 89% TN (ETGs with class < 4)

DES morphological catalogue



Largest morphological catalog up to date!

✓27 million galaxies
✓Up to 21.5 mag
✓12% ETGs, 88% LTGs
✓Released with paper

Vega-Ferrero, DS+21

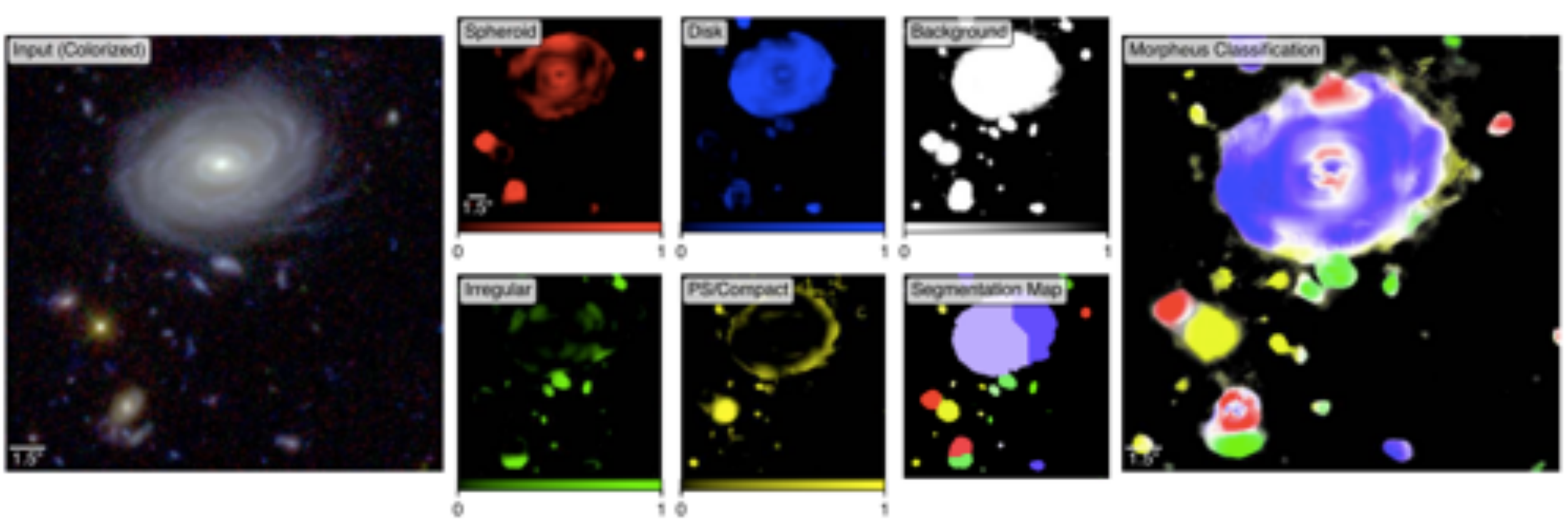
Morphologies at pixel level

Morpheus: A Deep Learning Framework For Pixel-Level Analysis of Astronomical Image Data

RYAN HAUSEN[1] AND BRANT E. ROBERTSON[2, 3]

[1]*Department of Computer Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA*
[2]*Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA*
[3]*Institute for Advanced Study, 1 Einstein Drive, Princeton, NJ 08540 USA*

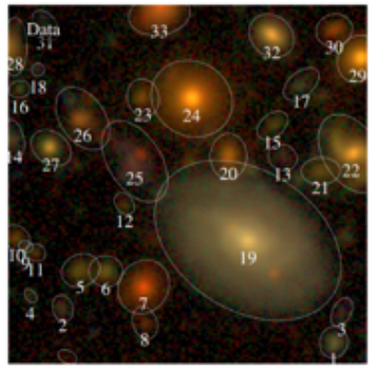
Object detection + segmentation + classification



Beyond Galaxy Morphology

✓Object detection and segmentation

- Deblending (e.g, Bocaud+19, Arcelin+2020)
- Source extraction (e.g., Hausen+19)

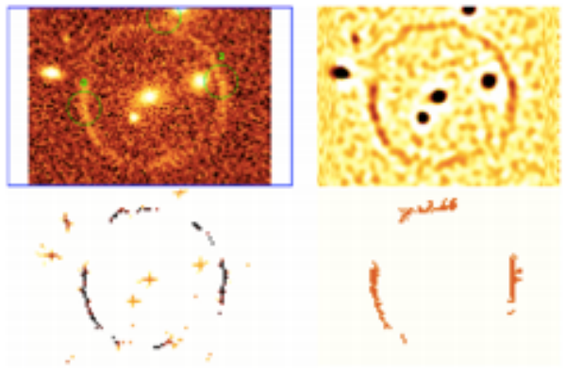


✓Classification

- Tidal streams (e.g., Walmsley+18)
- Mergers (e.g., Bottrell+19, Snyder+19, Pearson+19)
- Gravitational lenses (e.g., Petrillo+19, Metcalf+19, Jacobs+19, Cheng+20)

✓Regression

- Photo-z (e.g., Pasquet+18, Campagne+2020)
- Cluster Masses (e.g., Ho+20, Yan+20, Su+20)
- Galaxy morphometry (e.g., Tuccillo+18)



Extract spectra from images

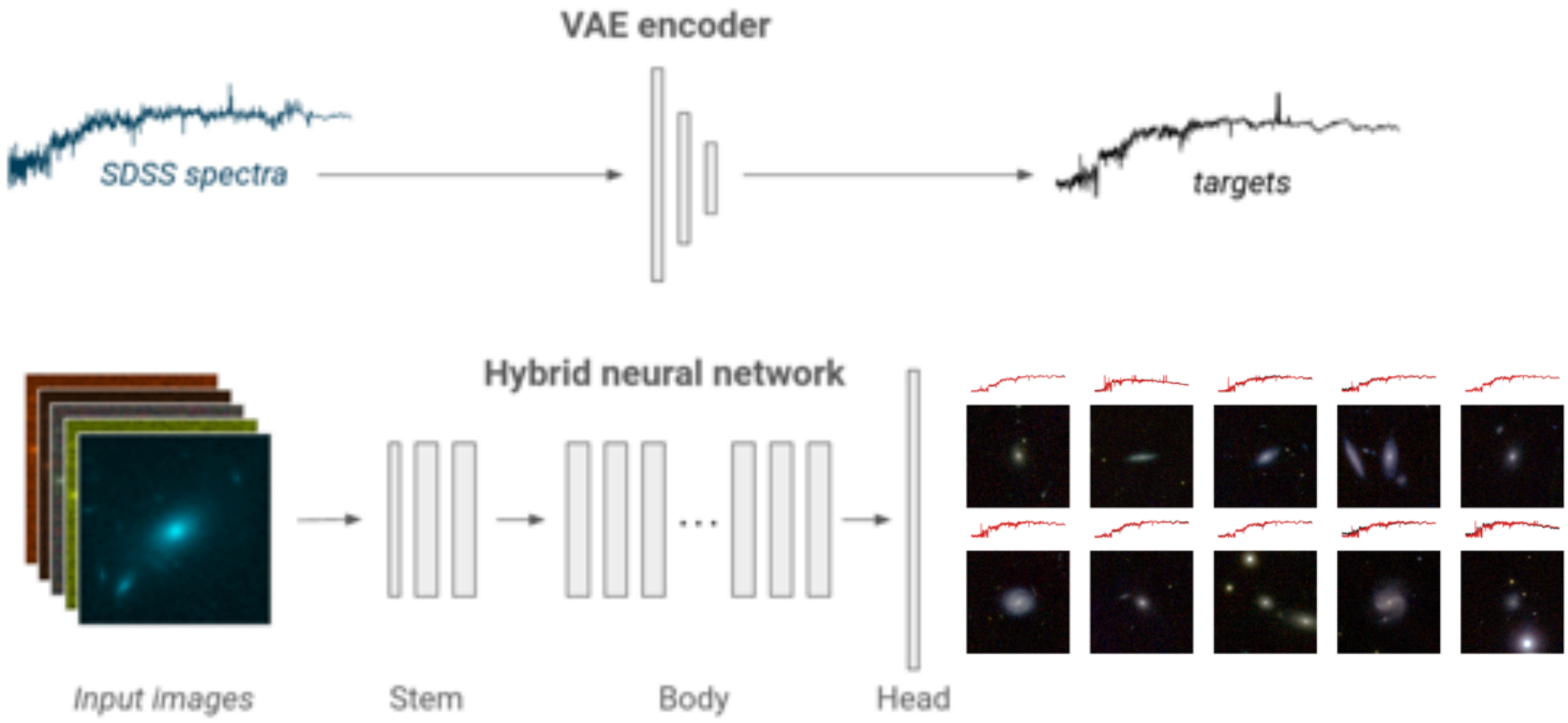
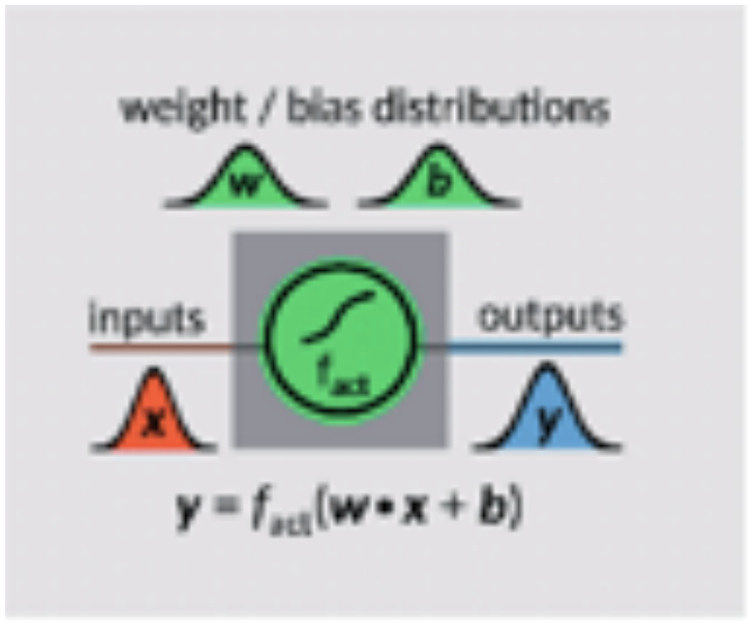


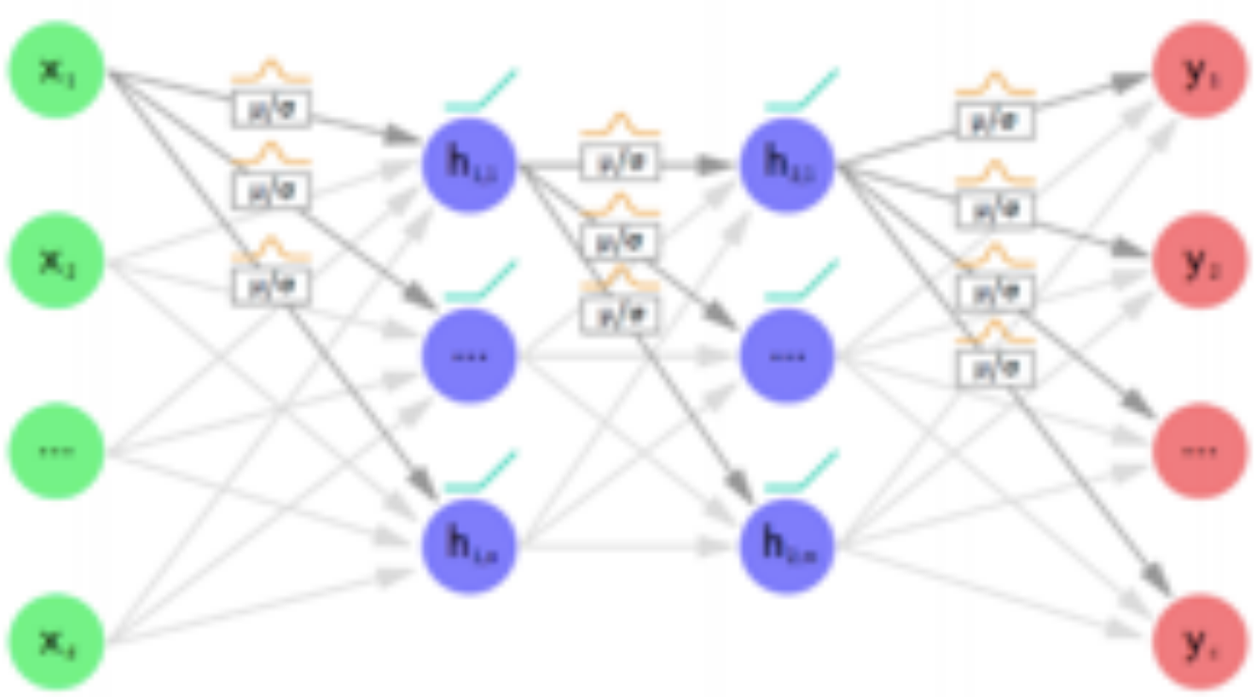
Figure 2: A schematic of our methodology. A pretrained VAE maps SDSS spectra to six latent variables, which we use as training targets (*upper*). We optimize a CNN to estimate the latent variables from *grizy* galaxy images (*lower*). Our best model comprises a deconvolution stem, resnet-like CNN body, and fully-connected layer head. While the loss function compares targets and predictions in the six-dimensional latent space, we show examples of the decoded spectra for visual comparison.

Wu, J., Peek +2020

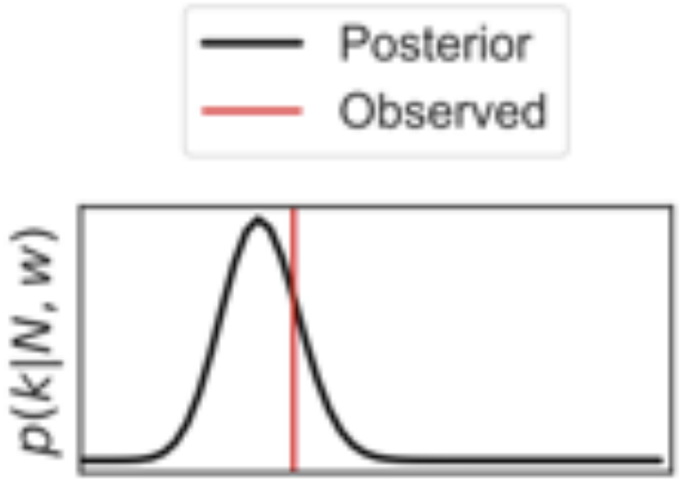
What about errors?

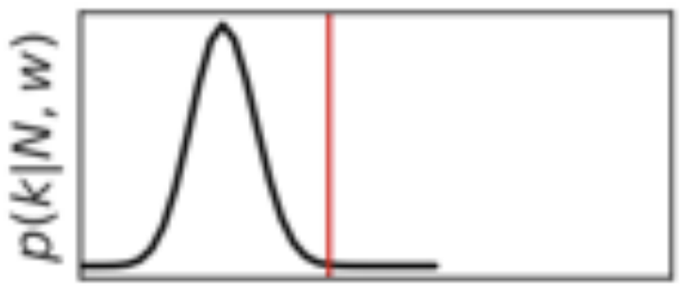
- Reliability and confidence estimates of CNN are important for astronomy.
- **Bayesian Neural Networks** directly model the uncertainty of the estimated network weights (e.g., Perreault Levasseur+17, Lin+20, Ho+20).

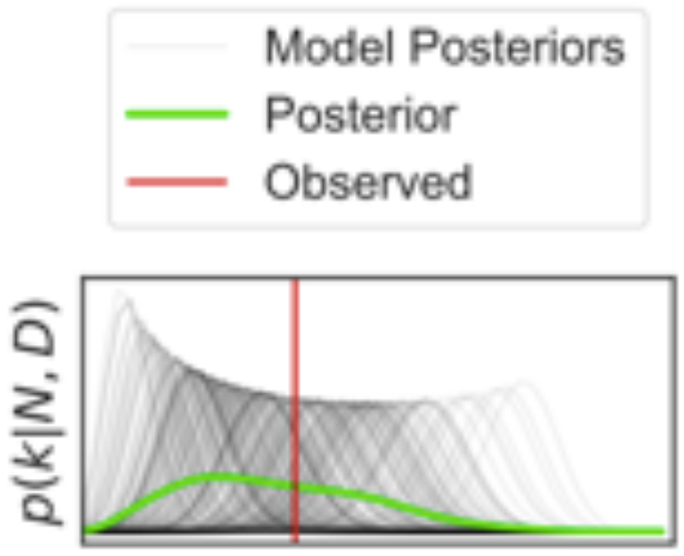


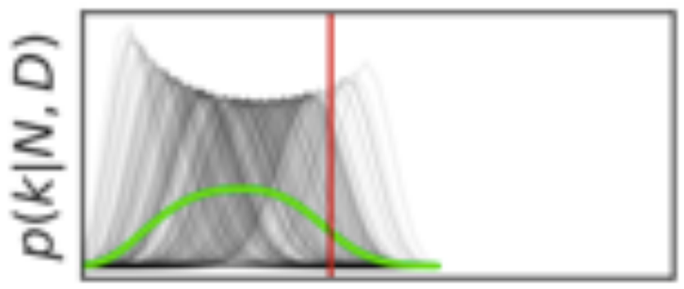


What about errors?









Bayesian Neural Networks + MonteCarlo Dropout (Walsmley+2019)

HOW DO WE MAKE DISCOVERIES?

- FUTURE BIG-DATASETS WILL BE PROCESSED THROUGH AUTOMATED (ML) METHODS - MOST OF THE DATA WILL NEVER BE LOOKED BY HUMANS

- BUT, SUPERVISED LEARNING, BY DEFINITION, LOOKS FOR KNOWN OBJECTS

  - WHAT IF WE MISS SOMETHING?

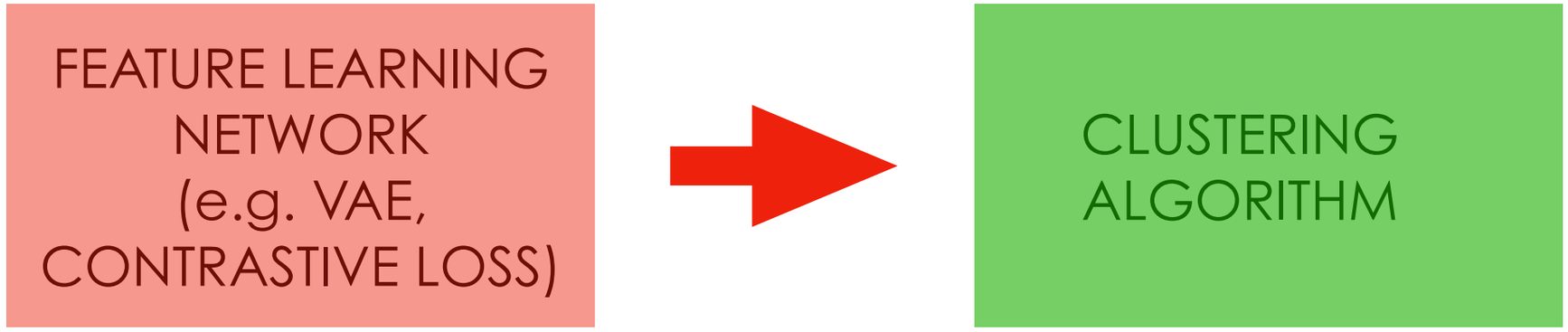
  - UNKNOWN UNKNOWNS IS WHERE INTERESTING (NEW) SCIENCE WILL BE FOUND

PART II

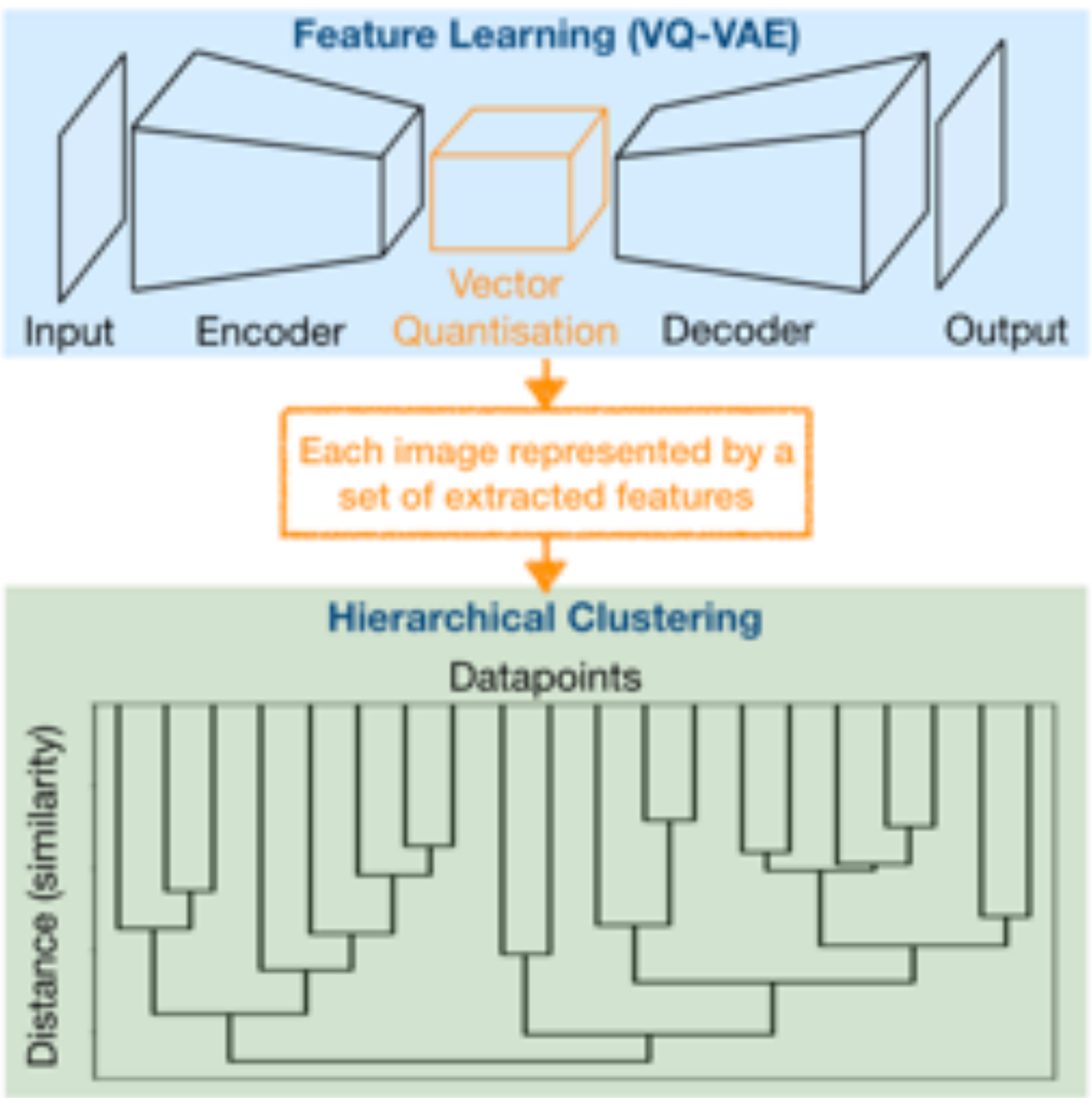
UNSUPERVISED OR SELF-SUPERVISED CLASSIFICATION

Let the data speak!!

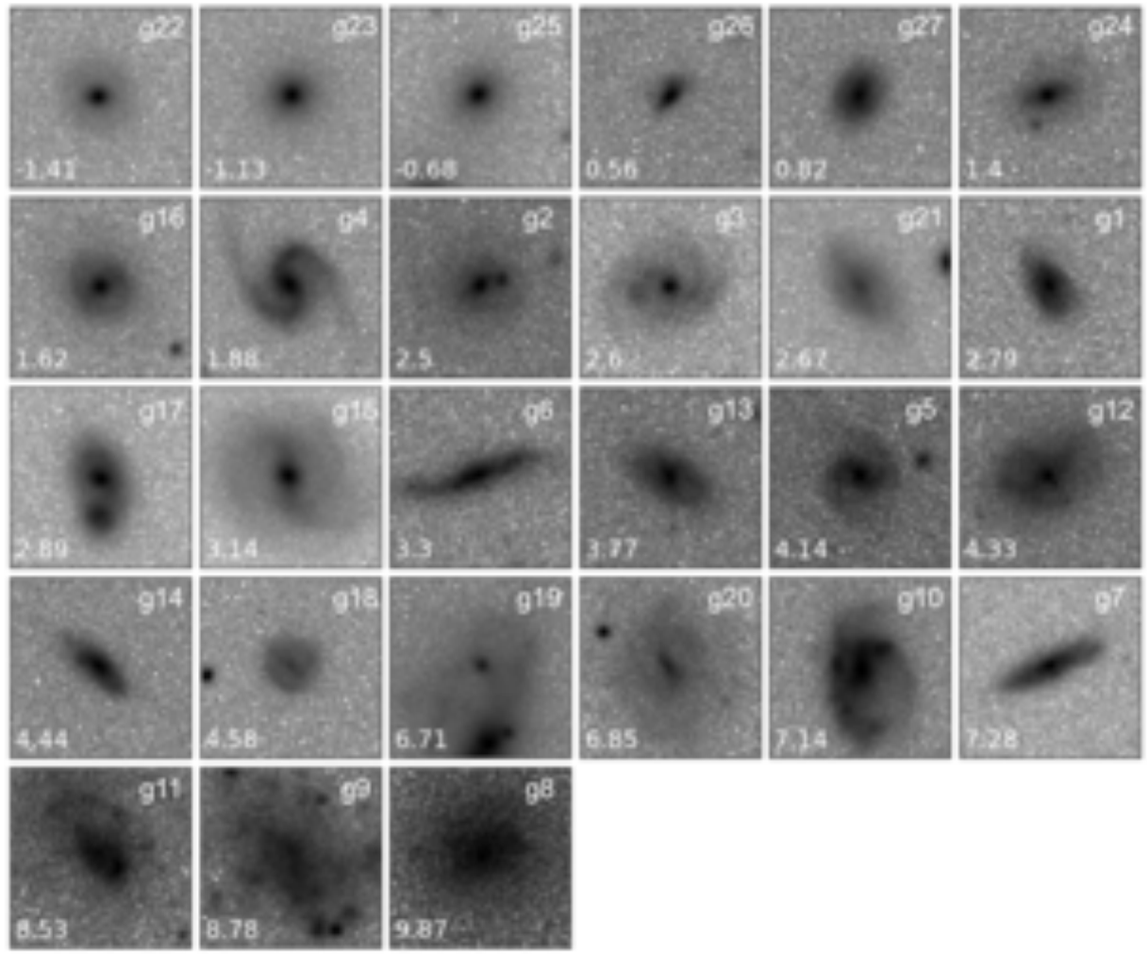
CNNs ARE POWERFUL FEATURE EXTRACTORS - THIS PROPERTY CAN BE COMBINED WITH CLUSTERING



Morphology Clustering



27 clusters based on galaxy structure/shape only



Interpretation is difficult!

Chen+20

Self-supervised learning

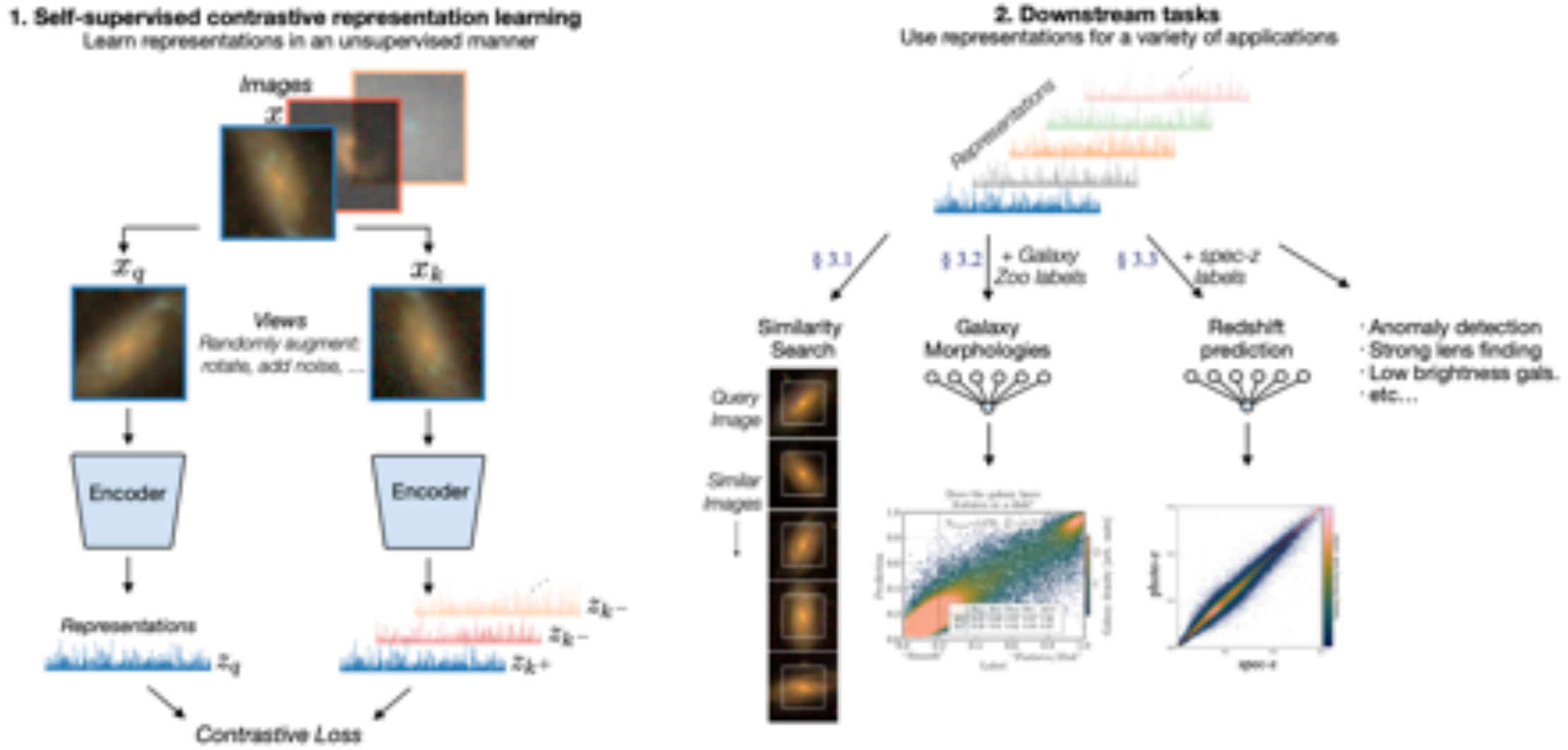


Figure 1. (Left) A schematic of the contrastive self-supervised framework. (Right) Examples of downstream tasks that can be learned on the learned representations.

Self-supervised learning

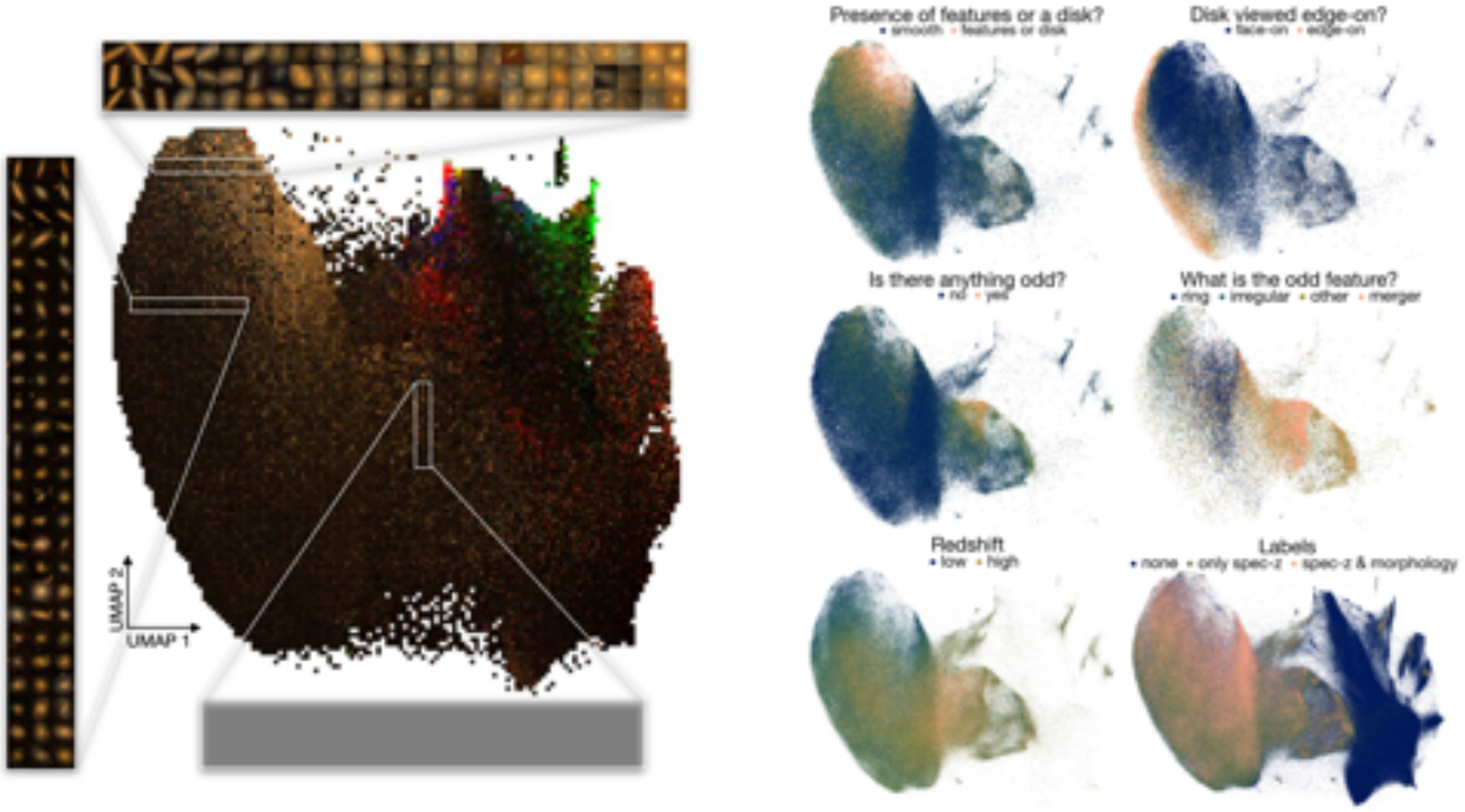


Figure 2. Visualizing the two dimensional UMAP projection of the self-supervised representations. The left panel shows randomly sampled representative images at each point in the space, while the right colors the space using answers to morphological classification questions from Galaxy Zoo 2, SDSS spectroscopic redshifts, or by labels.

BUT HOW DO WE MAKE DISCOVERIES?

WHAT IS AN ANOMALY OR OUTLIER?

$$p(X)$$

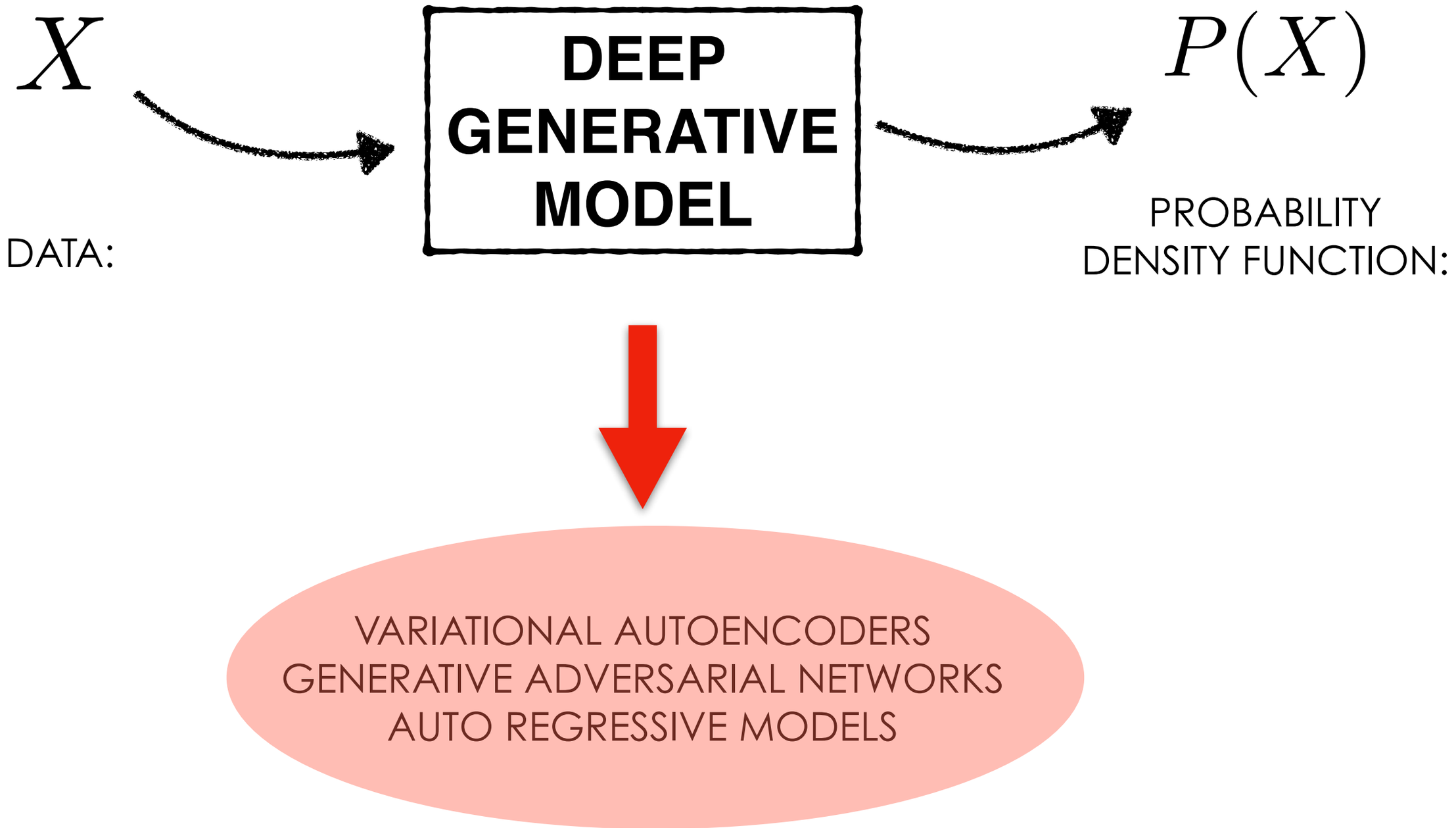
you data follows some probability distribution p

Then an object will be anomalous if:

$$p(x_i) < \epsilon$$

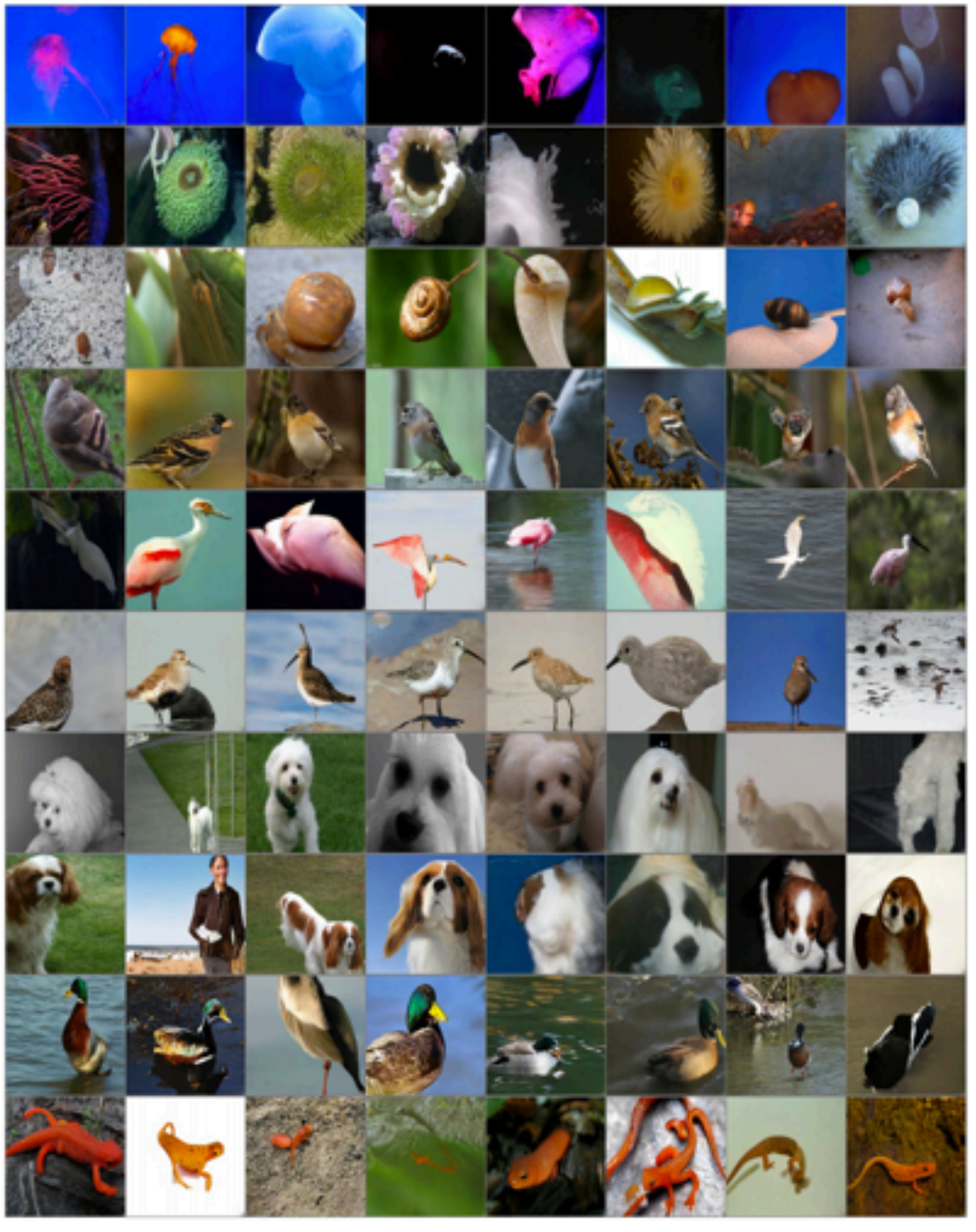
HOW DO WE COMPUTE THE PROBABILITY DISTRIBUTION p?

GENERATIVE MODELS DO PRECISELY THAT:



Razavi+19

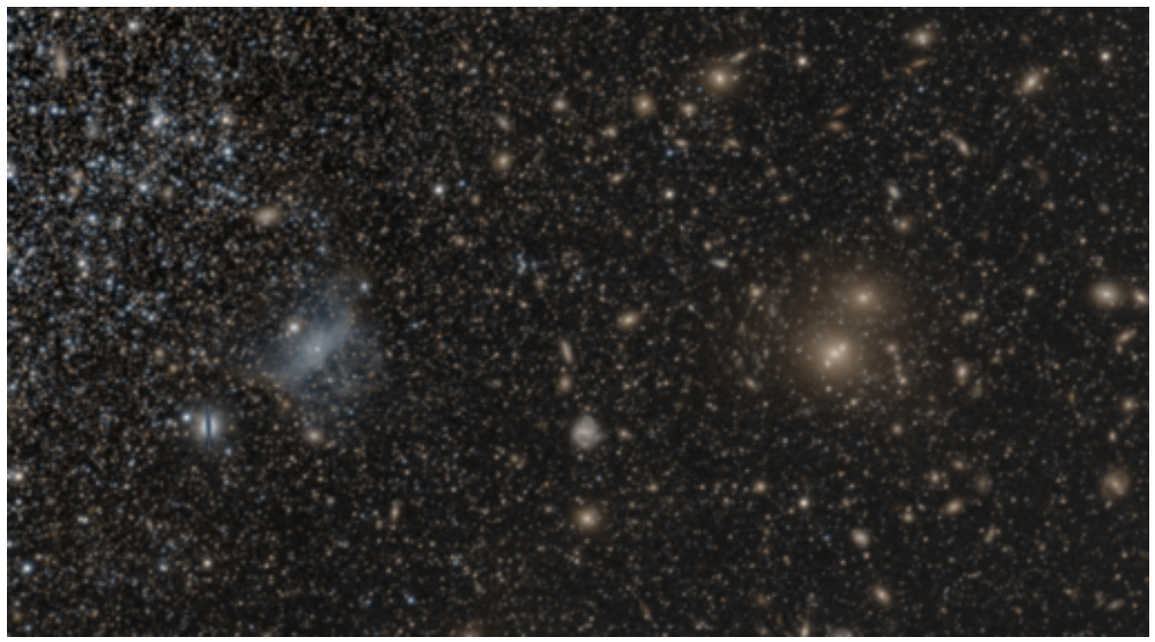
GENERATIVE MODELS DO PRECISELY THAT:



Anomaly Detection

Hyper Suprime Cam Survey (HSC)
Subaru telescope
1400 sq. deg

(20 < i < 20.5; 1 million)



WGAN → $P_{HSC}(X)$

x_i → $P_{HSC}(x_i)$ → identify most "unlikely" objects

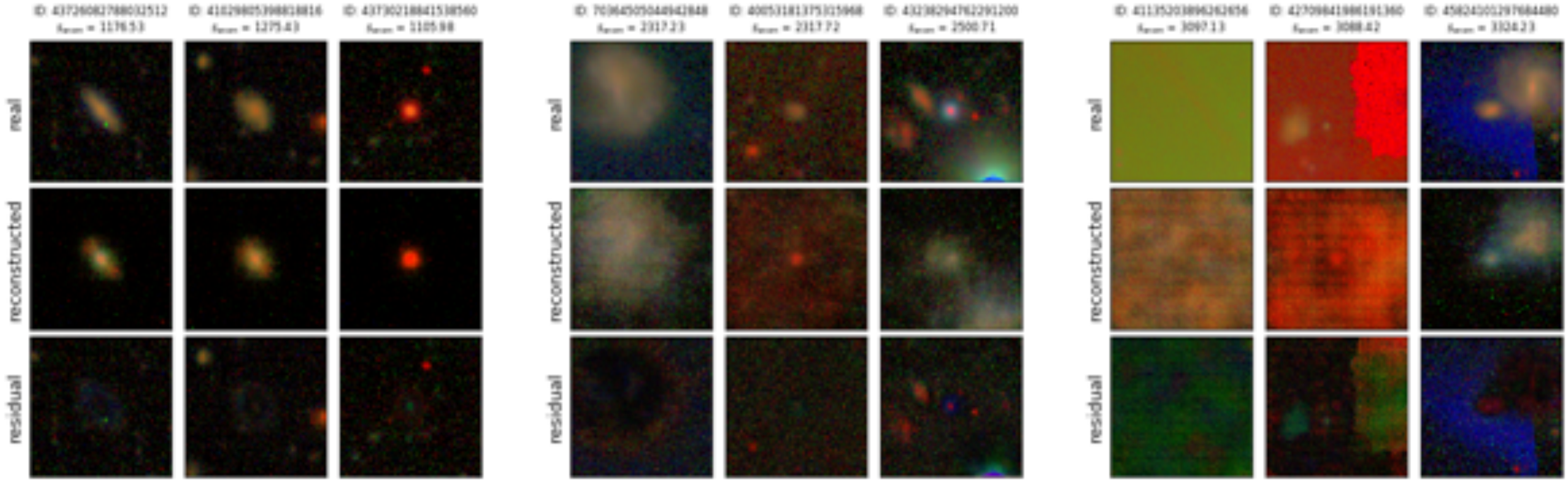
individual detection

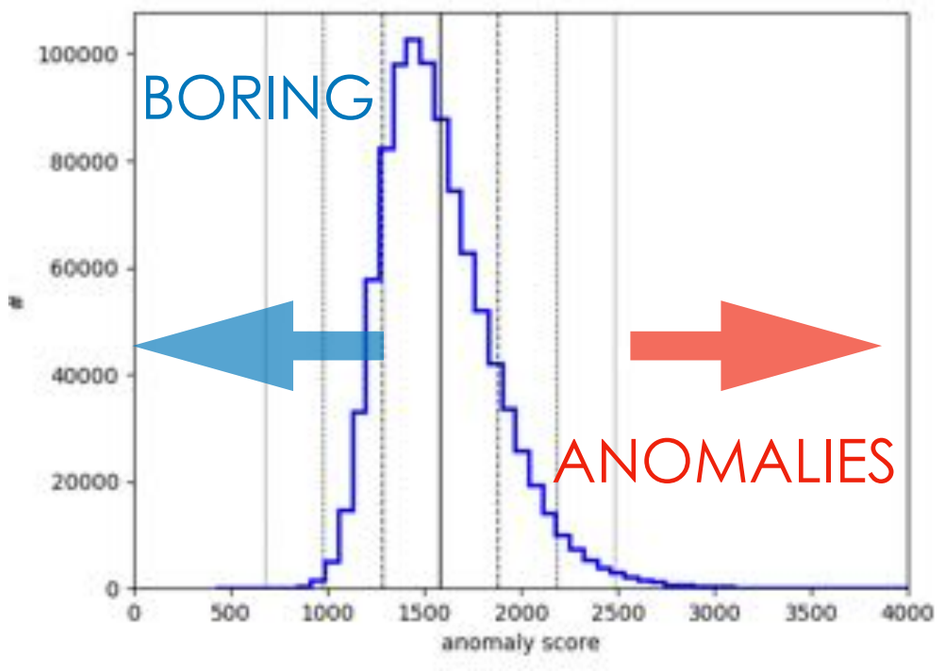
Storey-Fisher, MHC + 2021

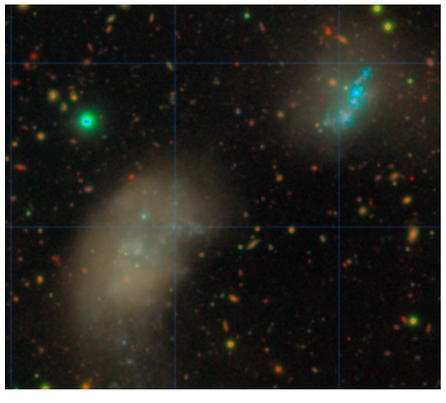
AVERAGE

1-SIGMA

>3-SIGMA

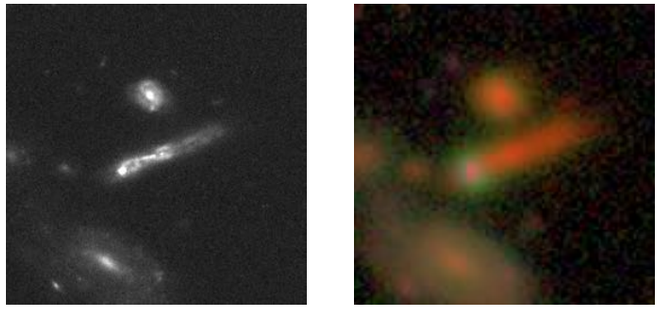






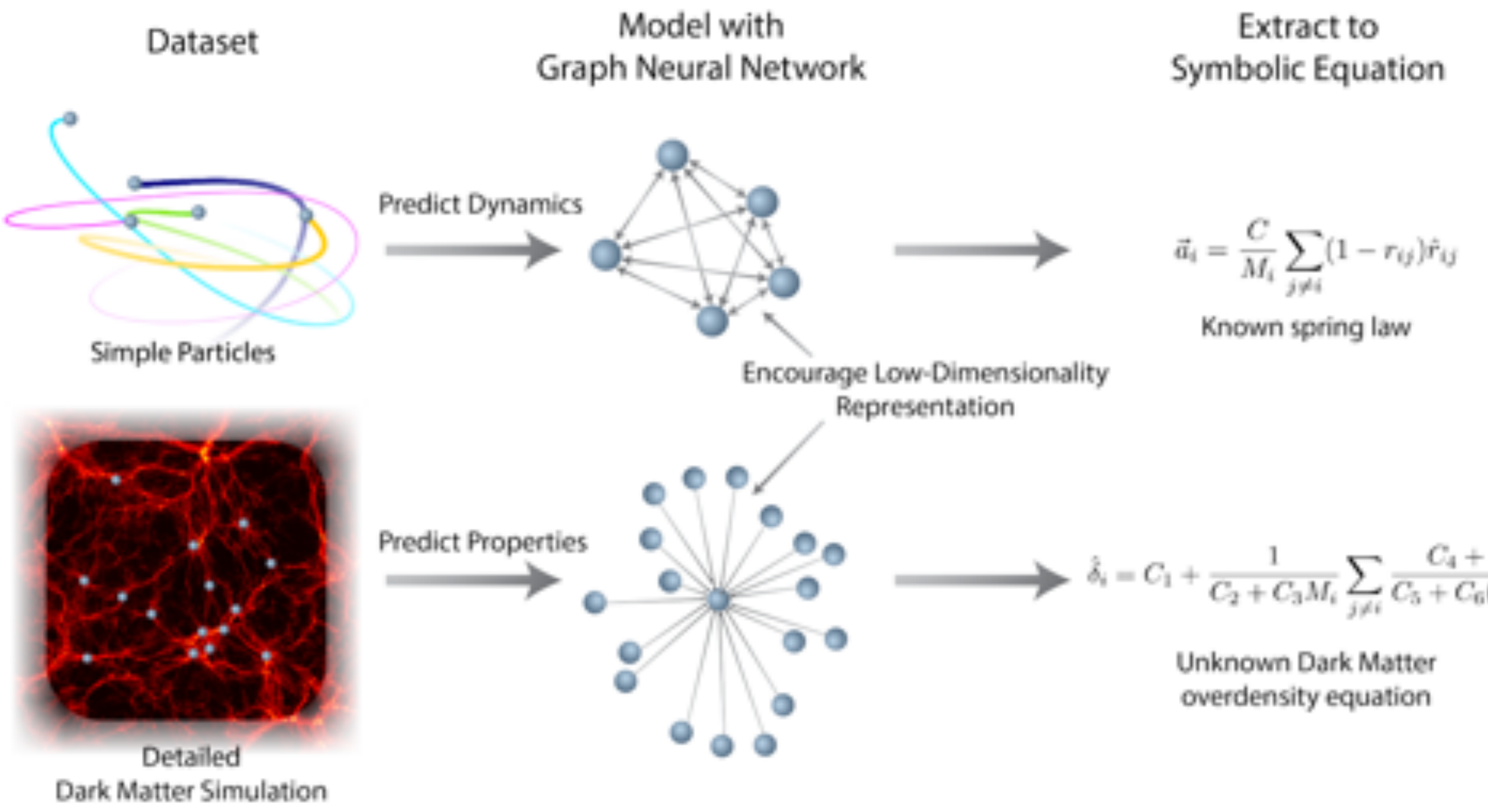
Extreme SF dwarf galaxy

HST HSC



Wandering Black Hole?

How do we learn new physics?



TAKE HOME MESSAGES

IN RECENT YEARS, MACHINE LEARNING AND ESPECIALLY DEEP LEARNING HAS REVOLUTIONIZED THE FIELD OF GALAXY CLASSIFICATION

DEEP LEARNING FOR CLASSIFICATION WILL BE A PART OF ANALYSIS PIPELINES OF FUTURE BIG DATA SURVEYS

THE BOTTLENECK FOR SUPERVISED LEARNING ARE TRAINING SETS; SIMULATIONS AND/OR TRANSFER LEARNING TECHNIQUES CAN BE USED

UNSUPERVISED CLASSIFICATIONS ARE GROWING AS AN ALTERNATIVE ALTHOUGH THIS IS LESS DEVELOPED. CAN BE USEFUL FOR AUTOMATED DISCOVERIES, AMONG OTHERS.