Investigating 'failed' galaxy classifications with machine learning

— The limits of visual classifications, new galaxy classes, and a pathway to unsupervised machine learning

Ting-Yun Cheng (Sunny)

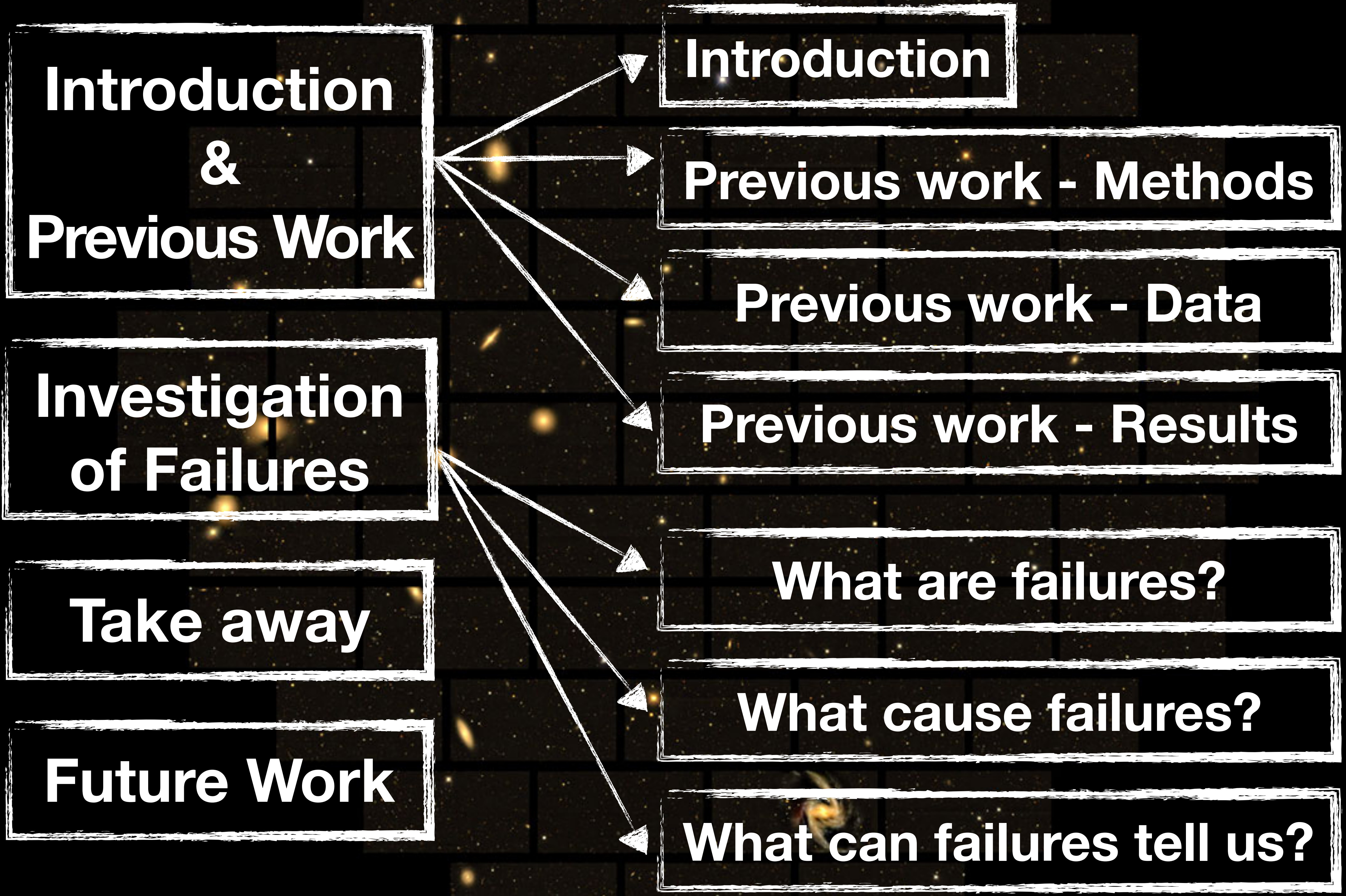
School of Physics & Astronomy
University of Nottingham

Supervisors: Christopher, J. Conselice & Alfonso Aragon-Salamanca

University of Nottingham
UK | CHINA | MALAYSIA

12 Sep 2018 Machine Learning Workshop
@Nottingham

Outline



Introduction & Previous Work

- Introduction
- Previous work - Methods
- Previous work - Data
- Previous work - Results

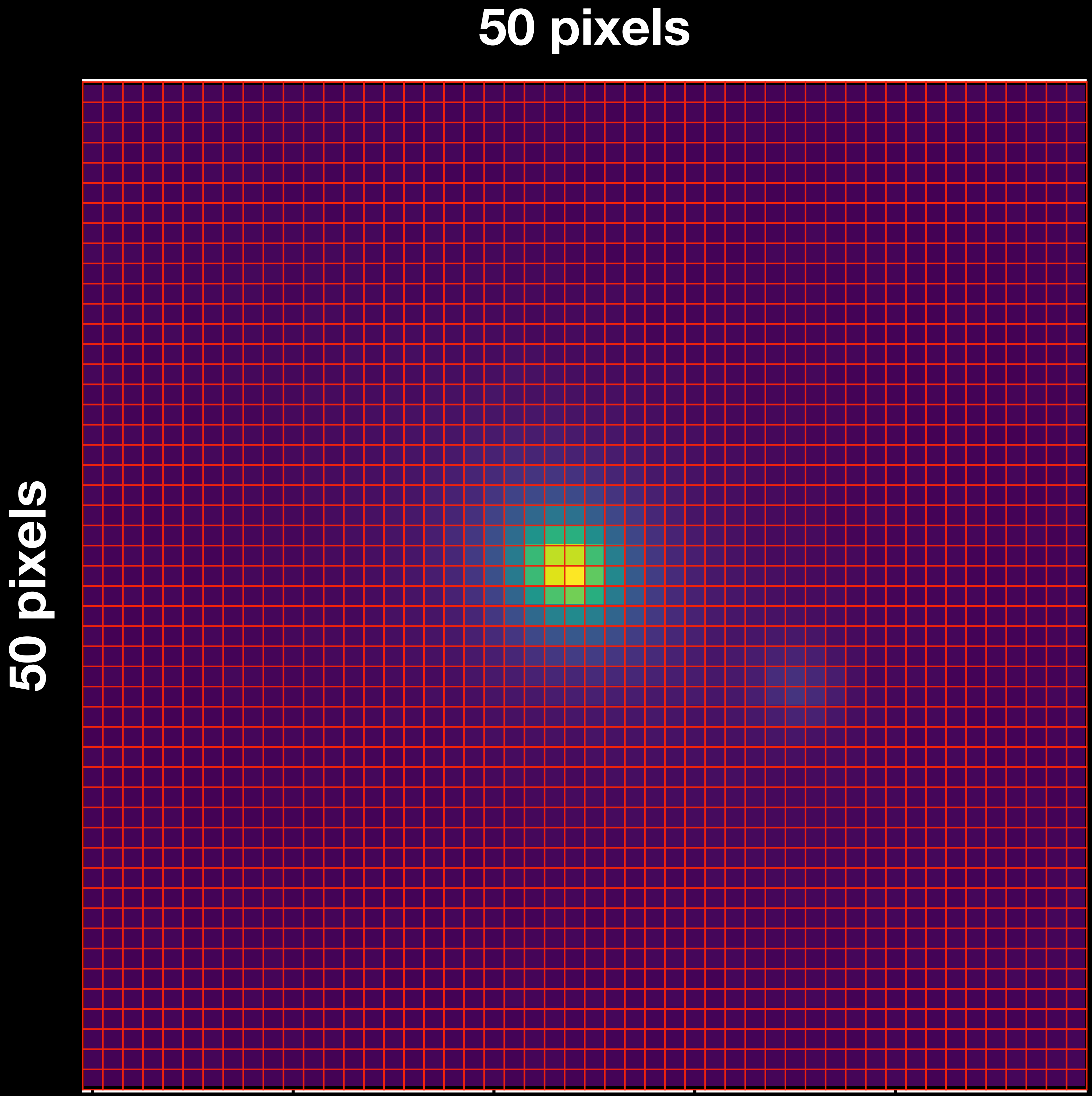
Investigation of Failures

- What are failures?
- What cause failures?
- What can failures tell us?

Take away

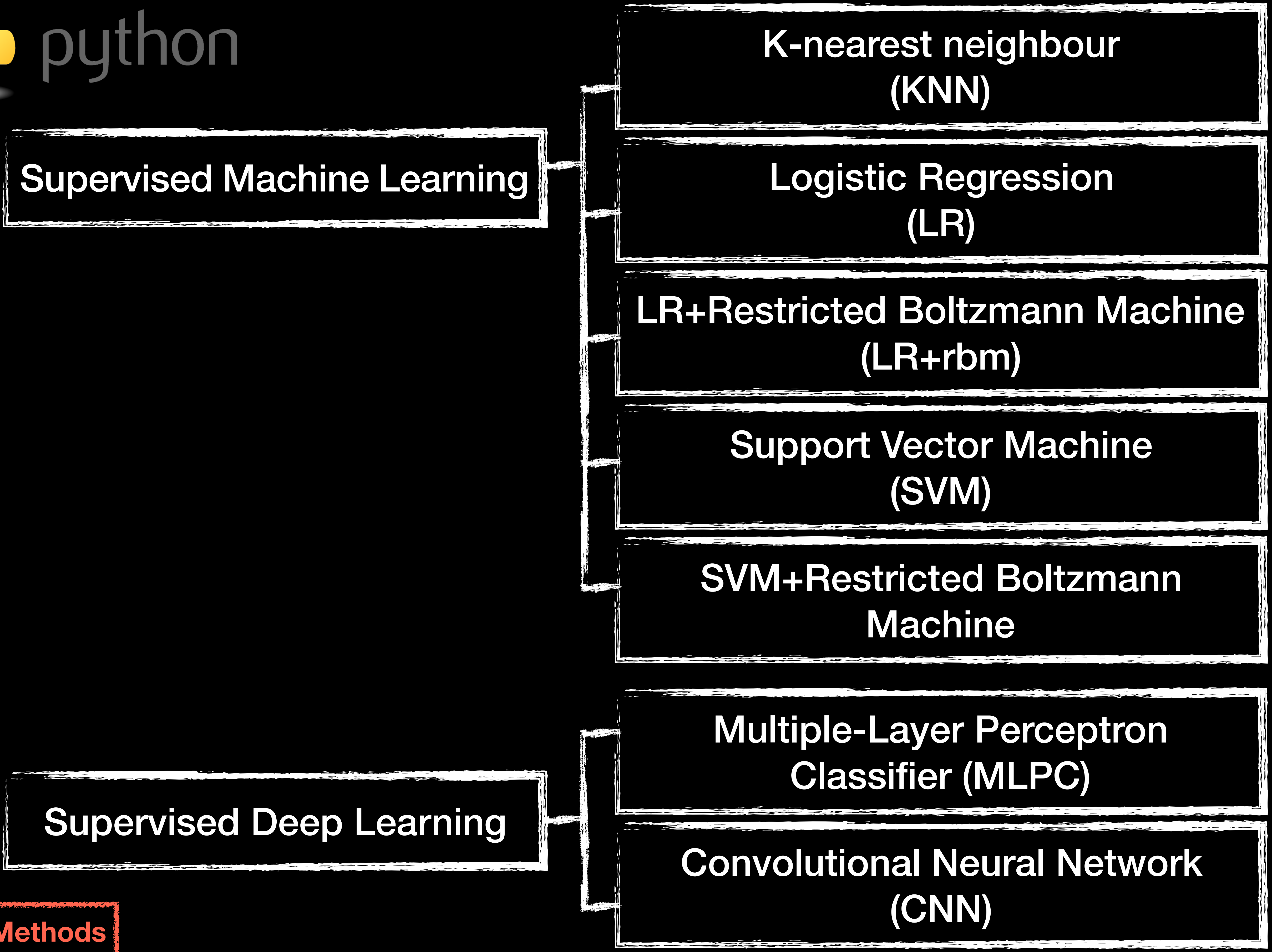
Future Work

Reference	Methods	Input types	
Storrie-Lombardi+92	Neural Network	Parameter input	Surface brightness, colour, etc.
Naim et al.+95	Neural Network	Parameter input	Surface brightness, diameter of ellipses fit, etc.
Lahav et al.+96	Neural Network	Parameter input	Surface brightness, diameter of ellipses fit, etc.
de la Calleja & Fuentes+04	Neural Network	Pixel input	
Ball et al.+04	Neural Network	Parameter input	Surface brightness profile, colour, etc.
Huertas-Company et al.+08	Support Vector Machine	Parameter input	C-A-S systems
Banerji et al.+10	Neural Network	Parameter input	de Vaucouleurs, exponential profile, colour, etc.
Huertas-Company et al.+11	Support Vector Machine	Parameter input	C-A-S systems
Polsterer et al.+12	Support Vector Machine	Pixel input	
Dieleman et al.+15	**Convolutional Neural Network**	**Pixel input**	
Huertas-Company et al.+15	Convolutional Neural Network	Pixel input	
Domínguez Sánchez et al.+18	Convolutional Neural Network	Pixel input	
Sreejith et al.+18	Support Vector Machine, Neural Network, Classification Trees, CTRF	Parameter input	Stellar mass, mass-to-light ratio, colour, sersic index, etc.



= 2500 features

python



Supervised Machine Learning

K-nearest neighbour (KNN)

Logistic Regression (LR)

LR+Restricted Boltzmann Machine (LR+rbm)

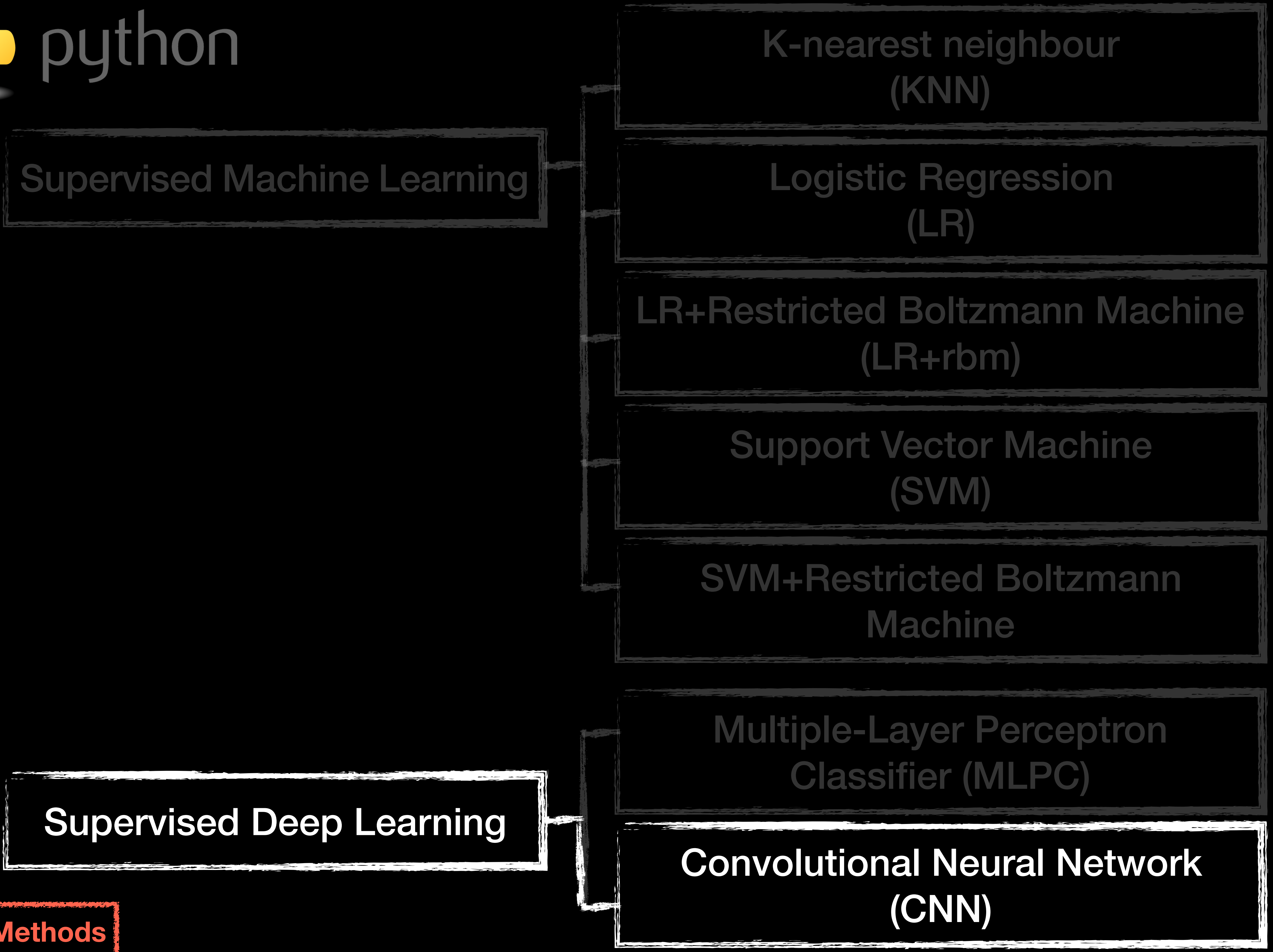
Support Vector Machine (SVM)

SVM+Restricted Boltzmann Machine

Supervised Deep Learning

Multiple-Layer Perceptron Classifier (MLPC)

Convolutional Neural Network (CNN)



Supervised Machine Learning

K-nearest neighbour (KNN)

Logistic Regression (LR)

LR+Restricted Boltzmann Machine (LR+rbm)

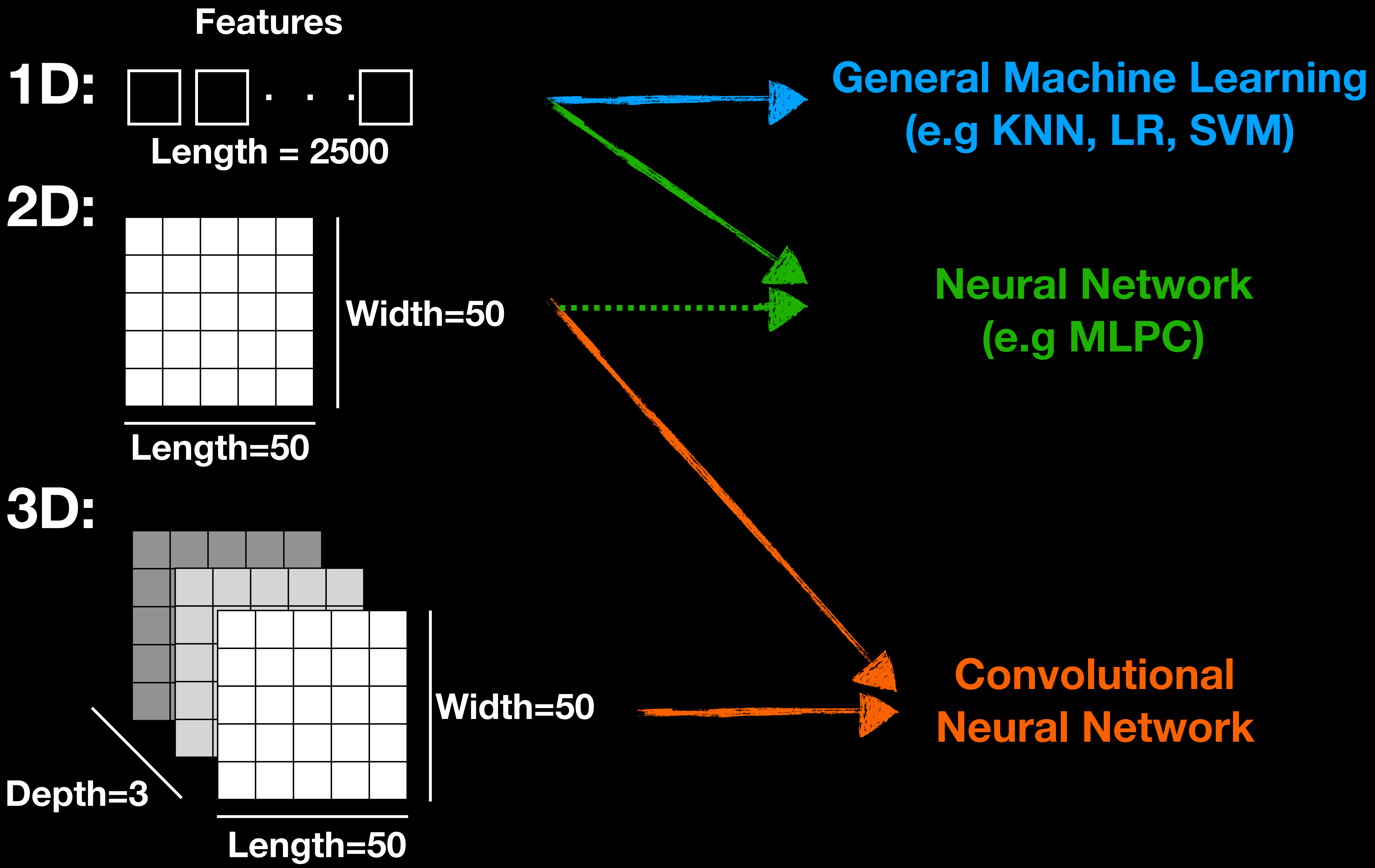
Support Vector Machine (SVM)

SVM+Restricted Boltzmann Machine

Supervised Deep Learning

Multiple-Layer Perceptron Classifier (MLPC)

Convolutional Neural Network (CNN)



Features

1D:

Length = 2500

2D:

Width=50

Length=50

3D:

Width=50

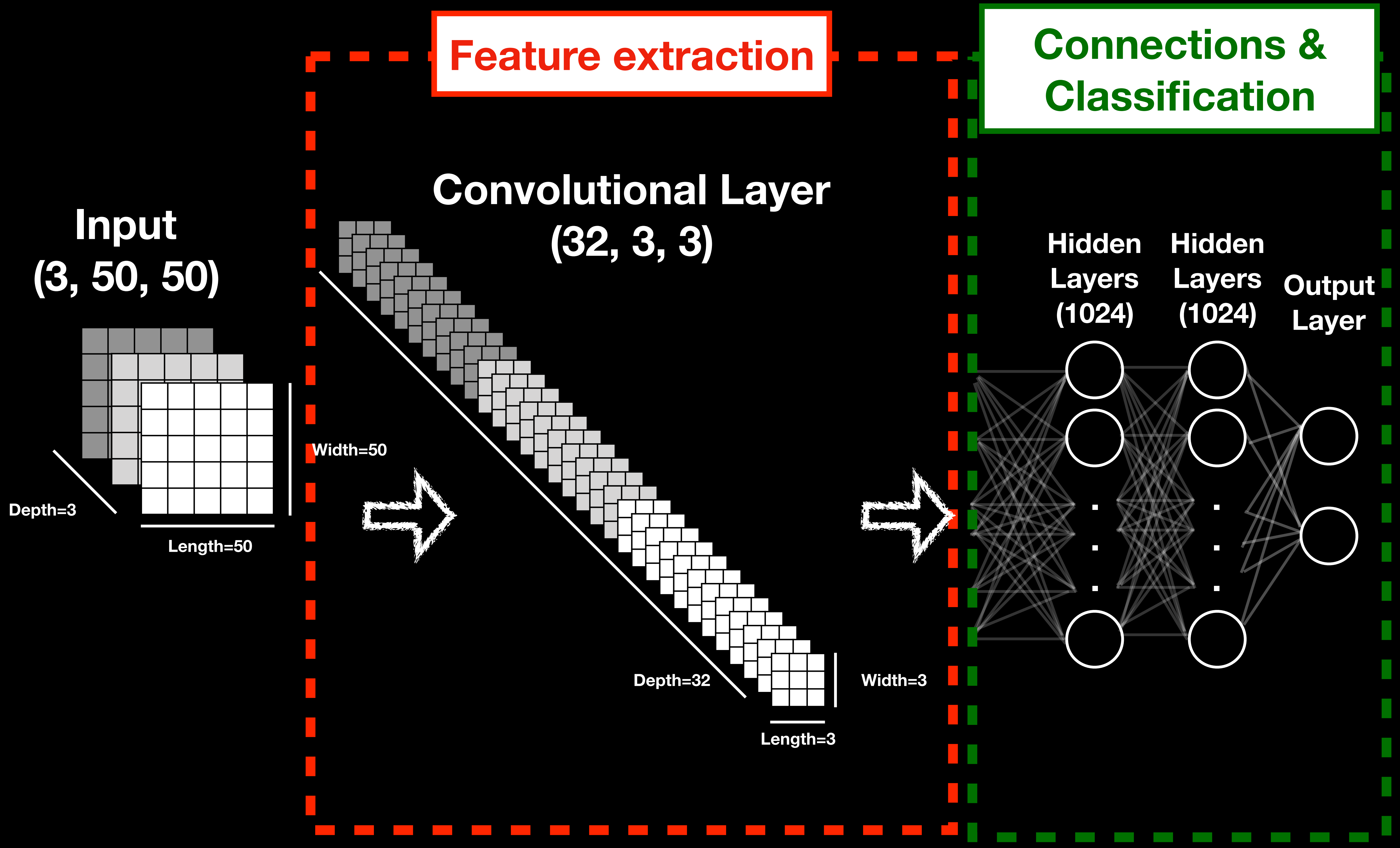
Depth=3

Length=50

General Machine Learning
(e.g KNN, LR, SVM)

Neural Network
(e.g MLPC)

Convolutional
Neural Network



Feature extraction

Connections & Classification

Input (3, 50, 50)

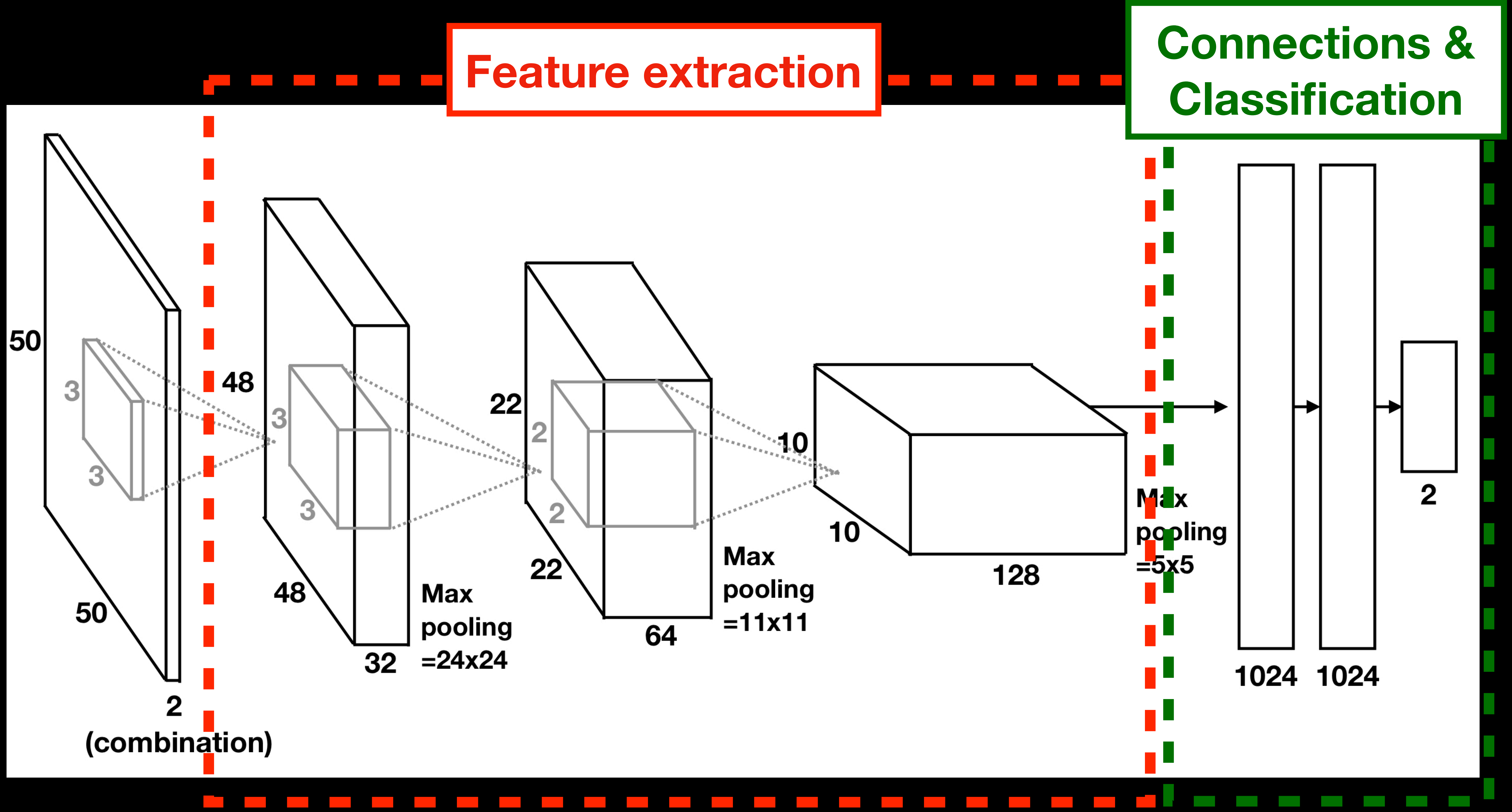
Convolutional Layer (32, 3, 3)

Hidden Layers (1024)

Hidden Layers (1024)

Output Layer

The architecture of our CNN



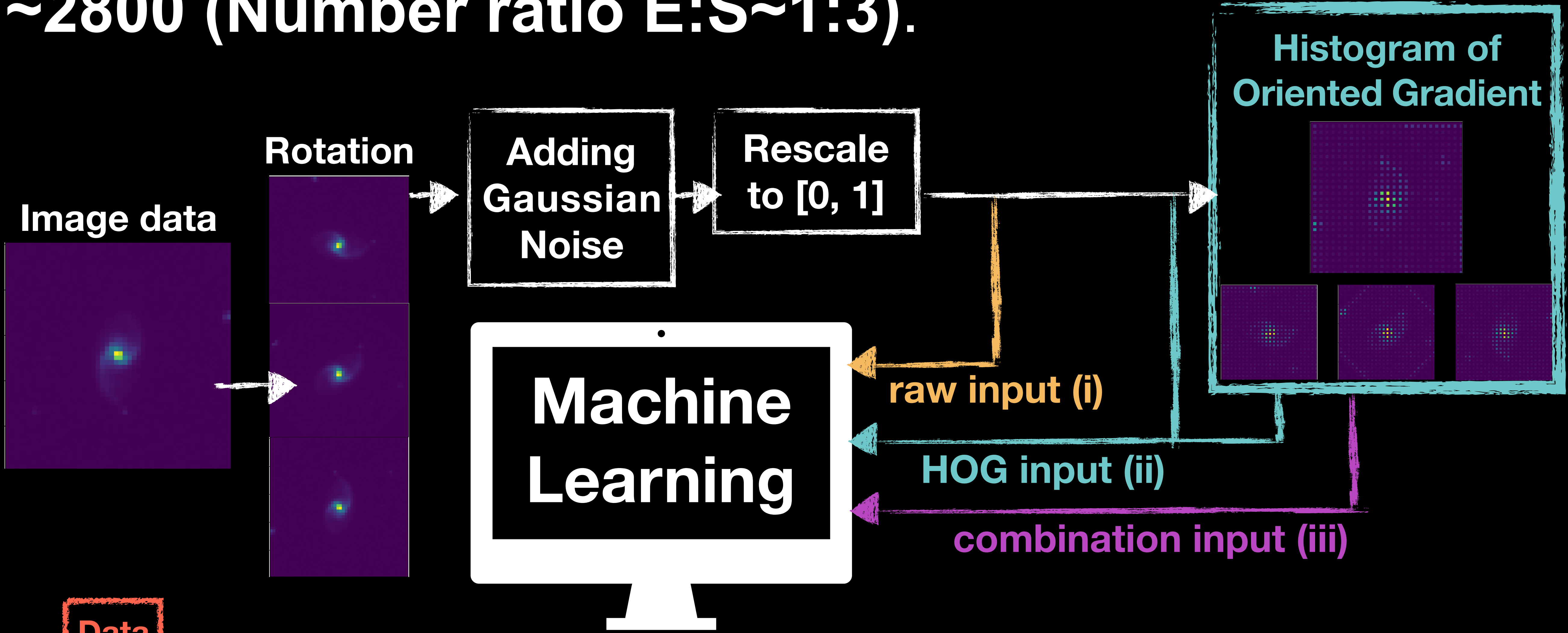
Feature extraction

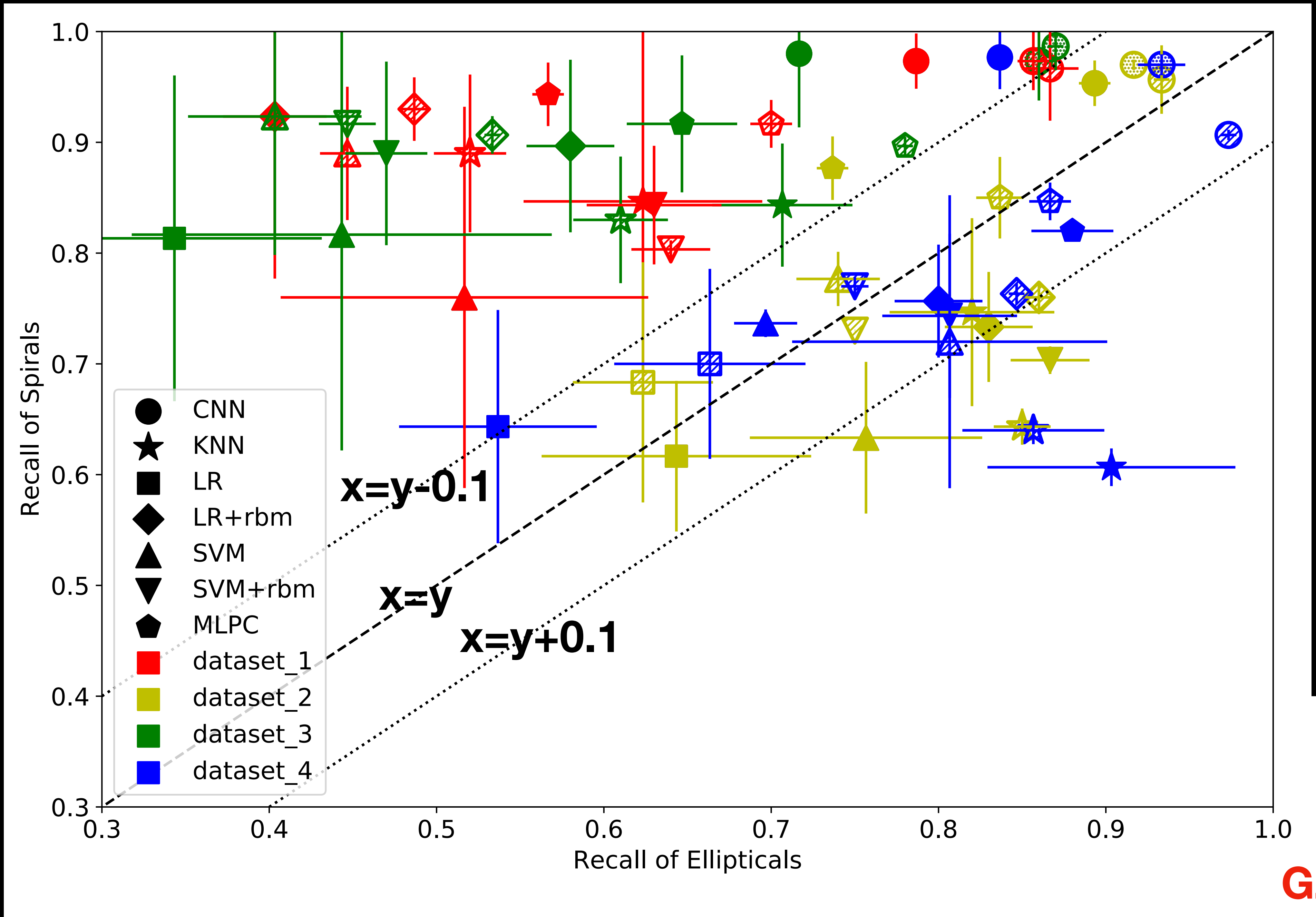
Connections & Classification

★ Dark Energy Survey (DES) Y1 GOLD data

★ Visual classification is from Galaxy Zoo1 project
(Classification with **agreement > 80%** for Ellipticals and Spirals)
(Lintott 2008, 2011)

★ Total number of matching sample between them is **~2800 (Number ratio E:S~1:3)**.





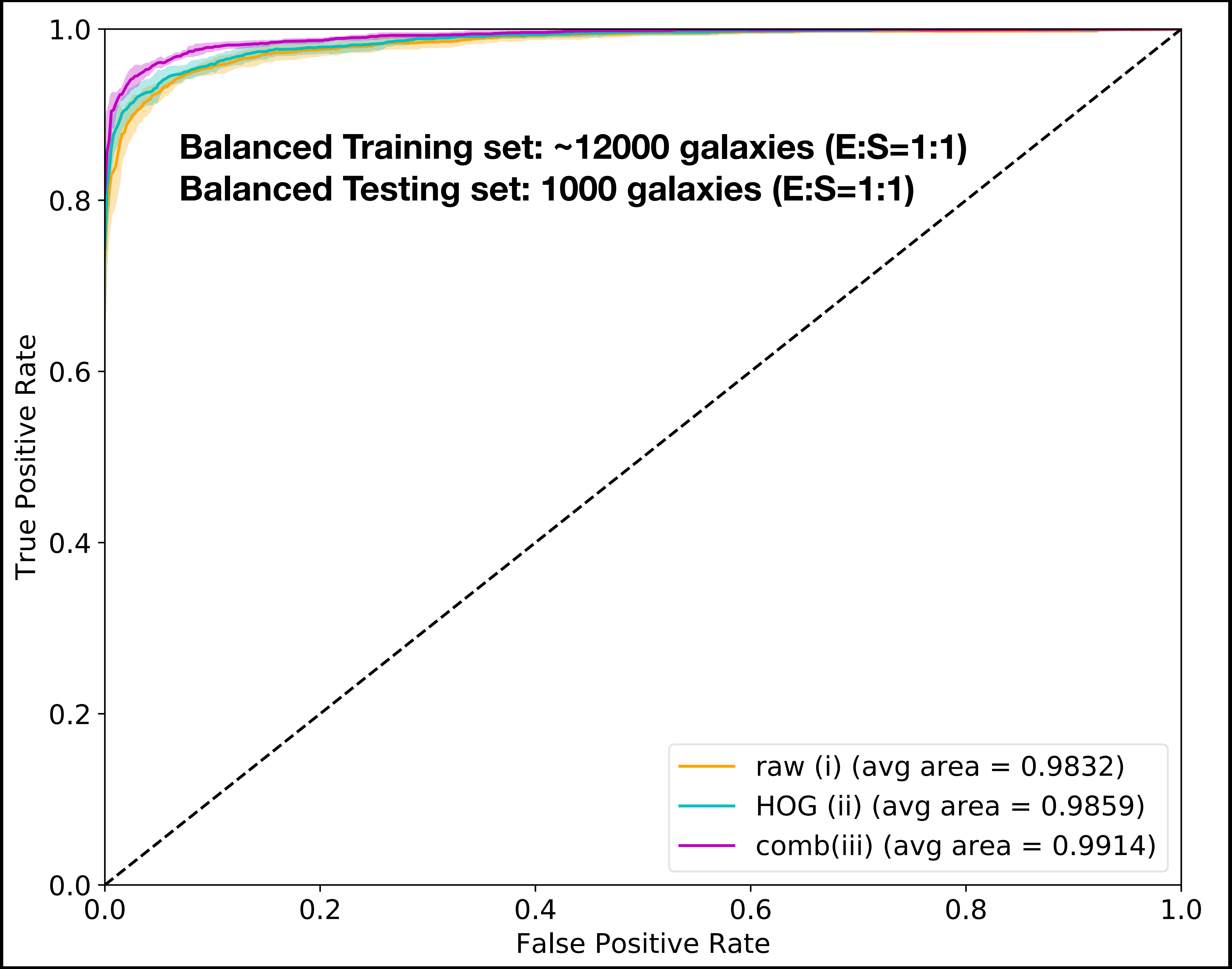
$$Recall(E) = \frac{TN}{TN+FP}$$

$$Recall(S) = \frac{TP}{TP+FN}$$

CNN **Predicted label**

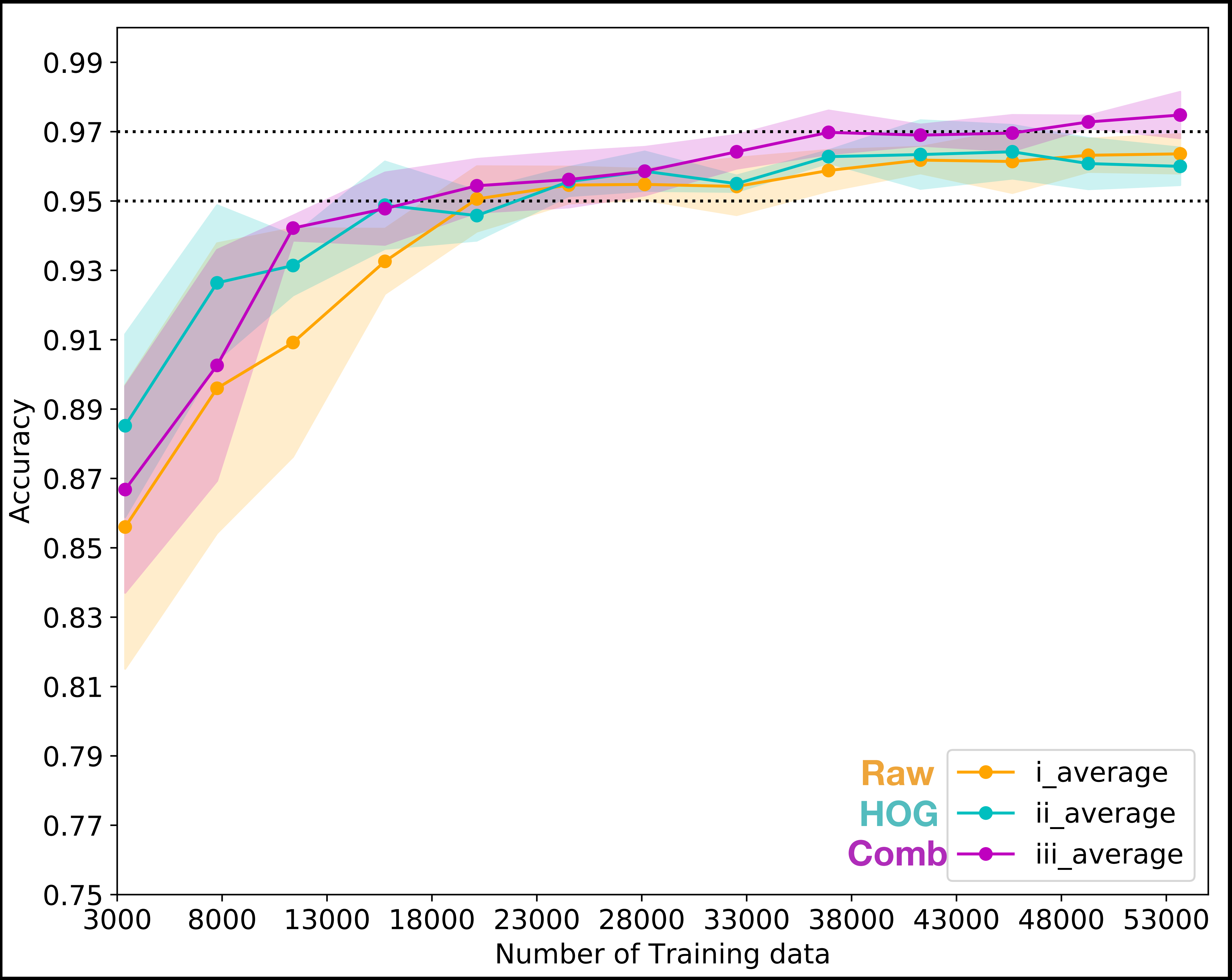
GZ

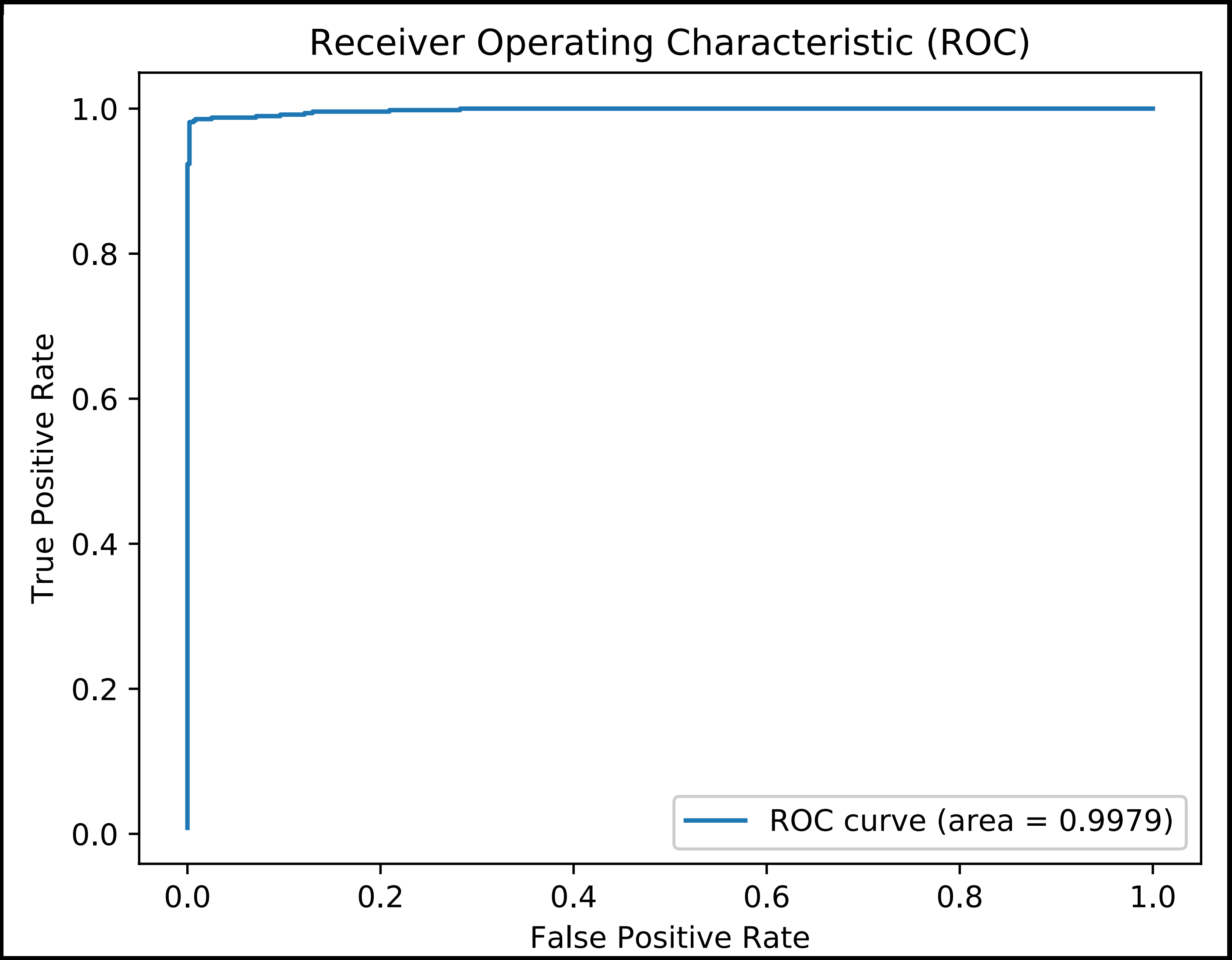
True label		0	1
	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True positive (TP)



Balanced Training set: ~12000 galaxies (E:S=1:1)

Balanced Testing set: 1000 galaxies (E:S=1:1)

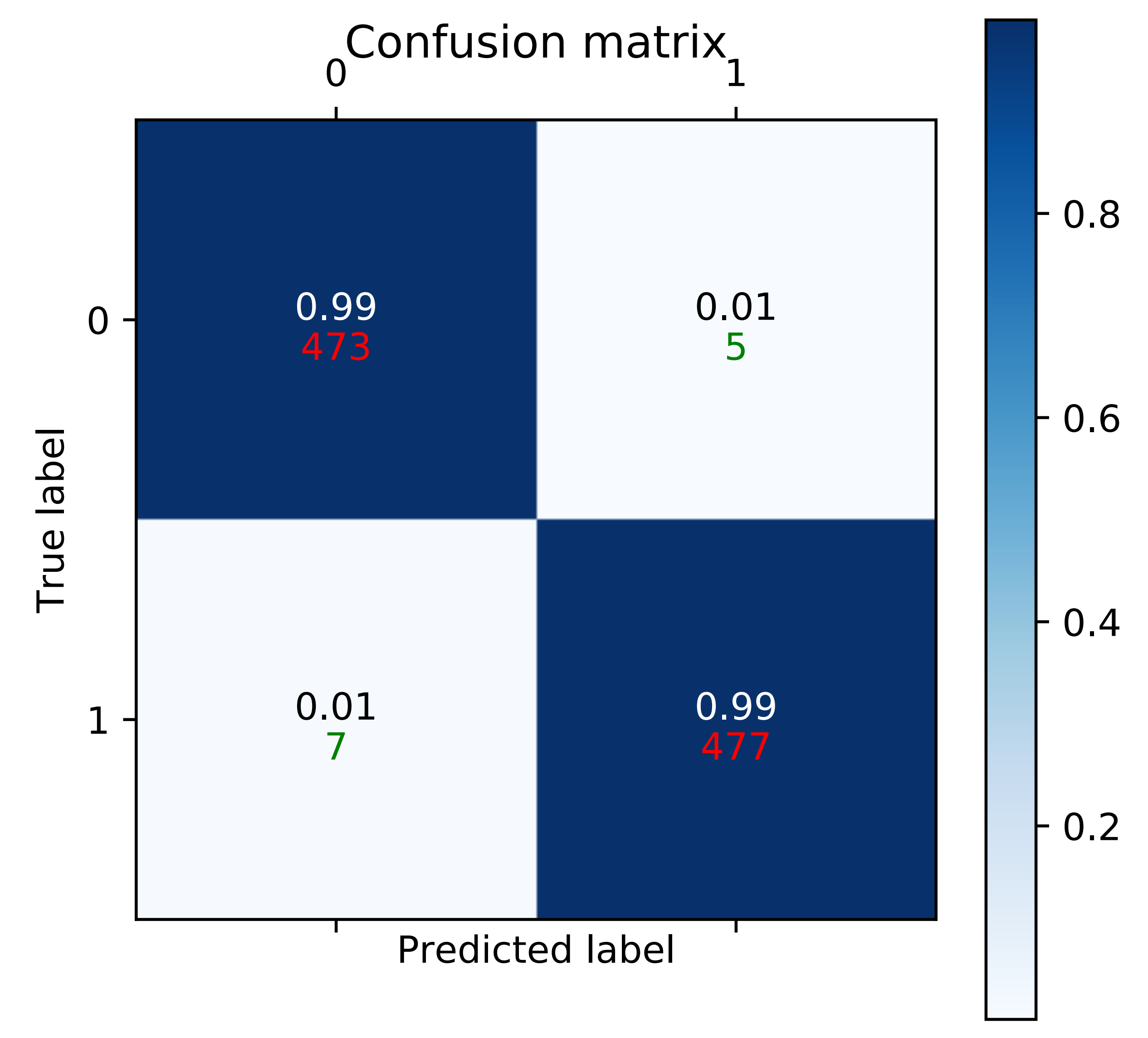

Accuracy
0.99
0.97
0.95
0.93
0.91
0.89
0.87
0.85
0.83
0.81
0.79
0.77
0.75
3000
8000
13000
18000
23000
28000
33000
38000
43000
48000
53000
Number of Training data
Raw
HOG
Comb
i_average
ii_average
iii_average



Balanced Training set = 53663
Balanced Testing set = 1000

Probability criterion: p=0.8

Accuracy = 0.987
Classifiable galaxies = 96%
non-classifiable galaxies = 4%
(Uncertain type)



Why did I fail??

? 1. What are failures?

? 2. What cause failures?

? 3. What can failures tell us?

What are failures?

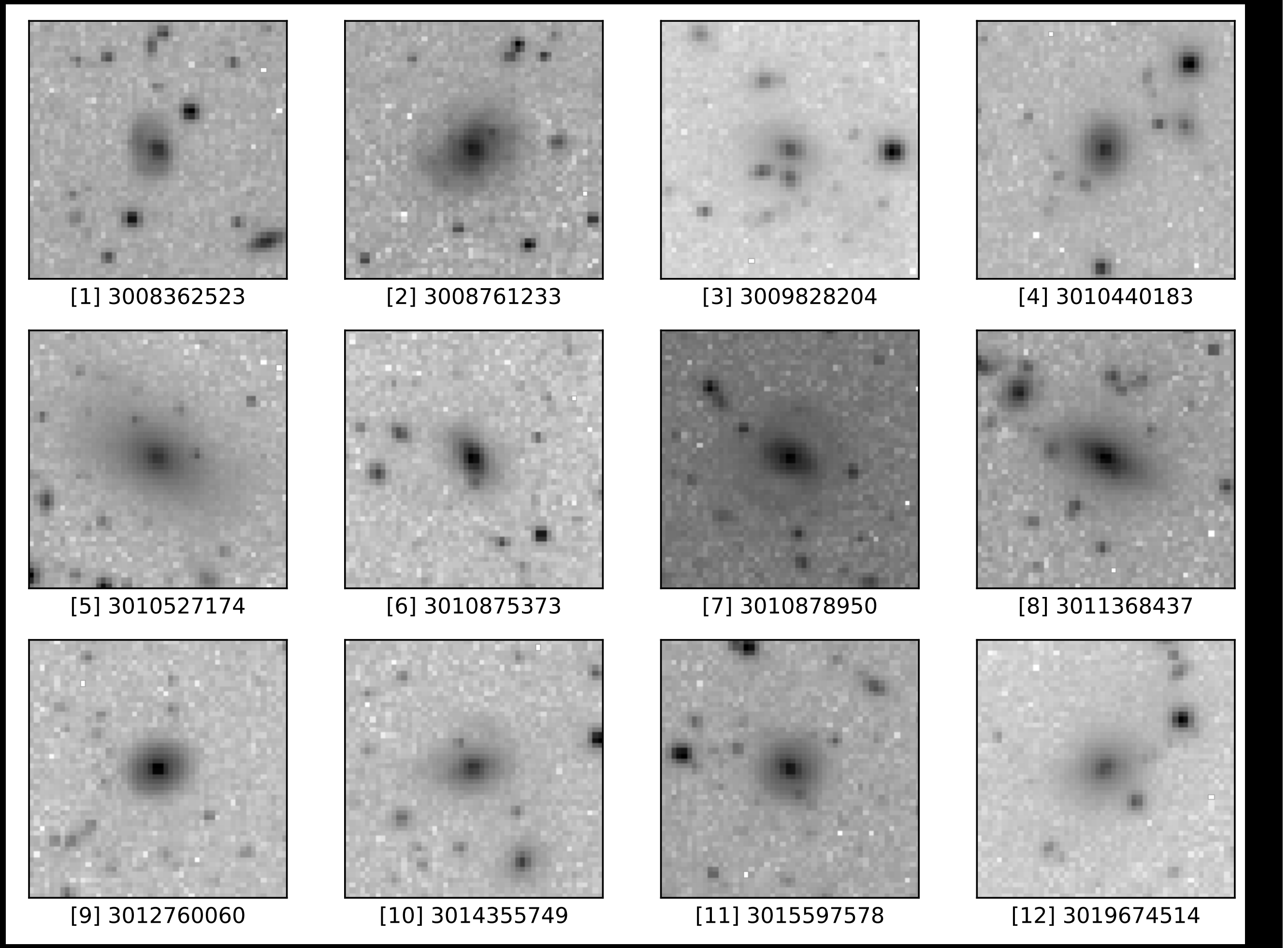
★ Low predicted probability ($p<0.8$) (Uncertain Type)

★ High predicted probability ($p\geq0.8$) but misclassified by our CNN

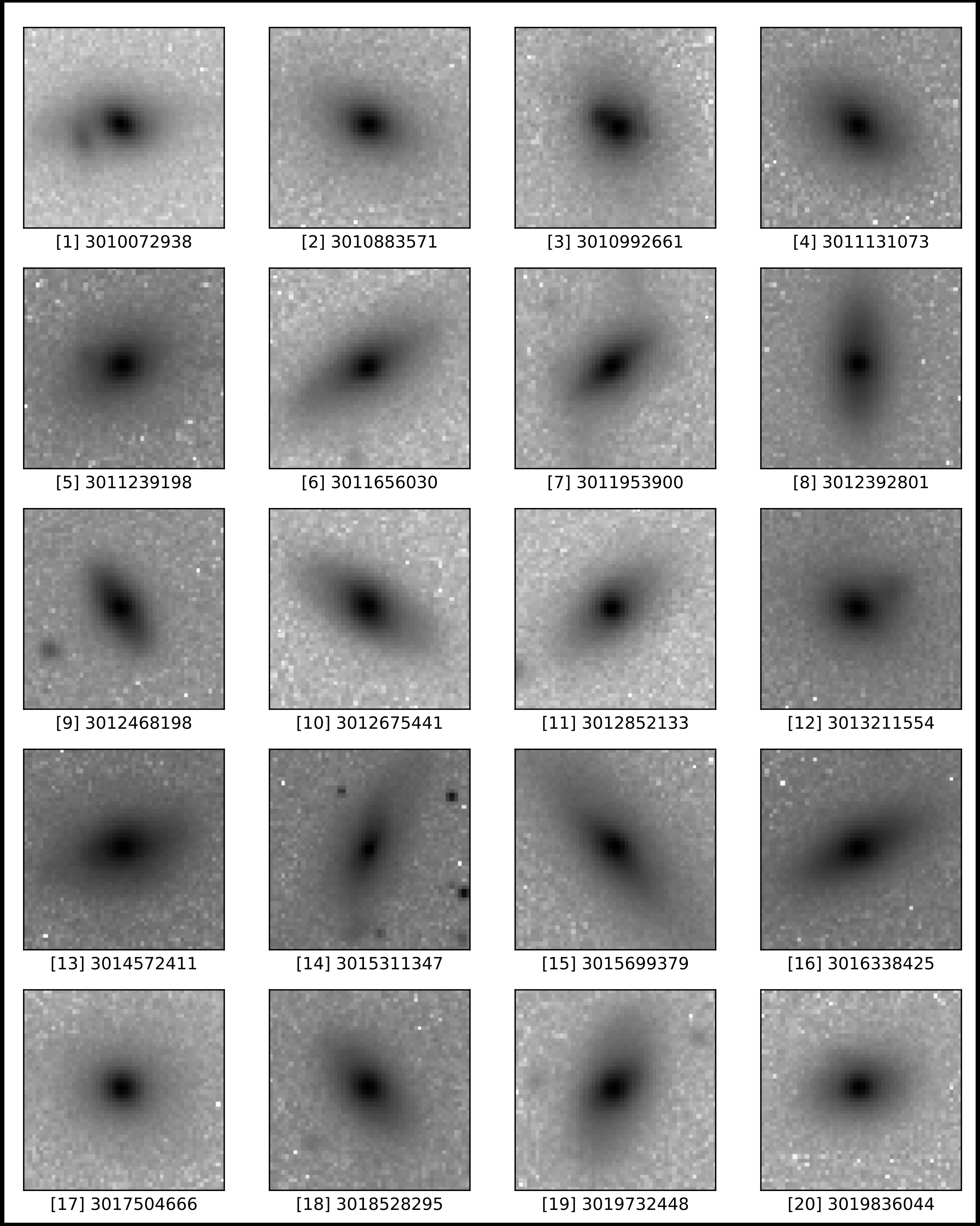
What are failures?

★ Low predicted probability ($p<0.8$) (Uncertain Type)

★ High predicted probability ($p≥0.8$) but misclassified by our CNN



[1] 3008362523
[2] 3008761233
[3] 3009828204
[4] 3010440183
[5] 3010527174
[6] 3010875373
[7] 3010878950
[8] 3011368437
[9] 3012760060
[10] 3014355749
[11] 3015597578
[12] 3019674514



[1] 3010072938
[2] 3010883571
[3] 3010992661
[4] 3011131073
[5] 3011239198
[6] 3011656030
[7] 3011953900
[8] 3012392801
[9] 3012468198
[10] 3012675441
[11] 3012852133
[12] 3013211554
[13] 3014572411
[14] 3015311347
[15] 3015699379
[16] 3016338425
[17] 3017504666
[18] 3018528295
[19] 3019732448
[20] 3019836044

What cause these failures?

There are three sources of the failures:

★ Difficult images

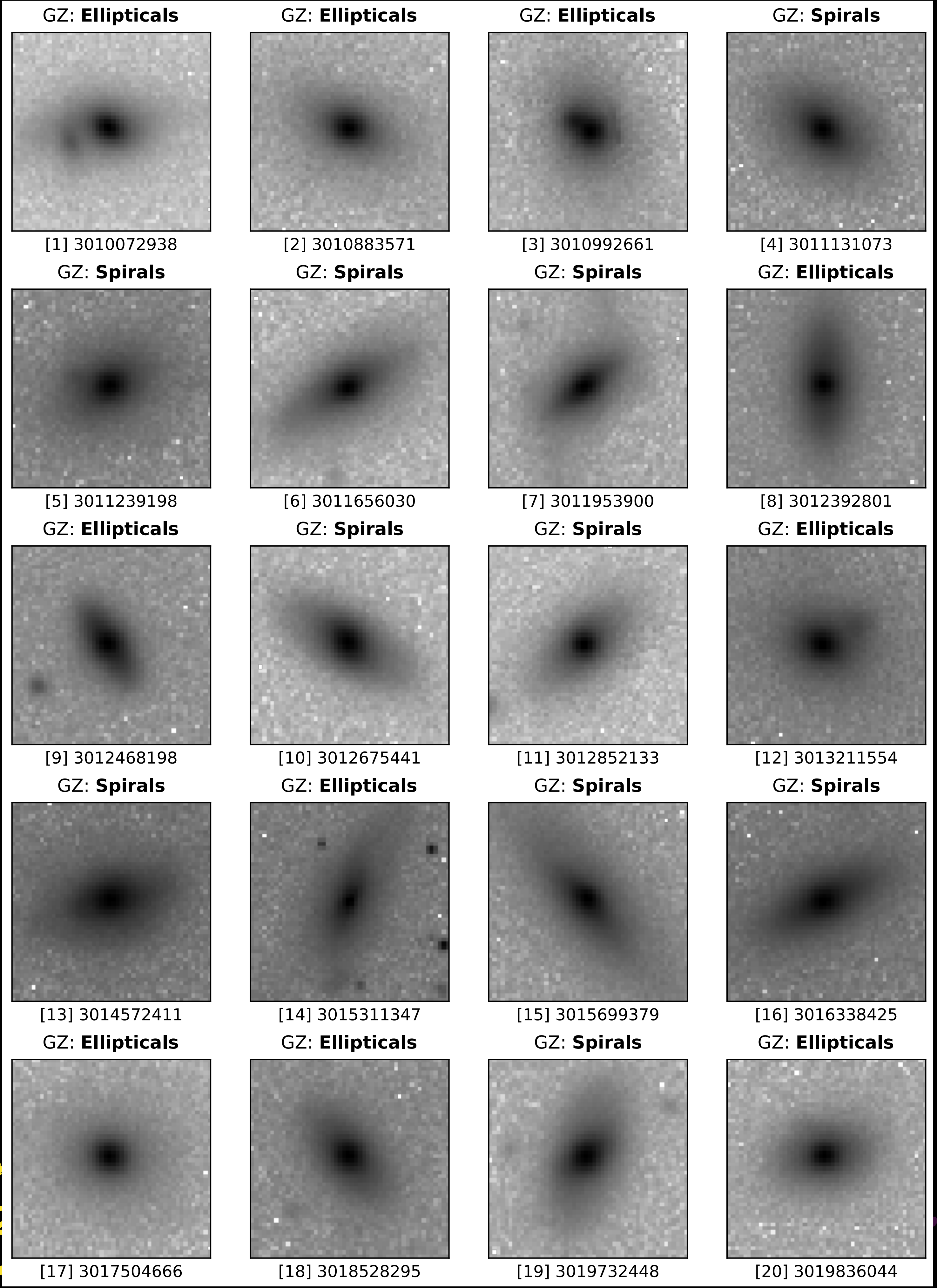
★ The problems from the initial labels

★ The problems from our CNN

★ Difficult images

Galaxy Zoo projects DO NOT have class for lenticular galaxies (S0).

Half of them are classified as Ellipticals, and half of them are Spirals.
(Examples: →)



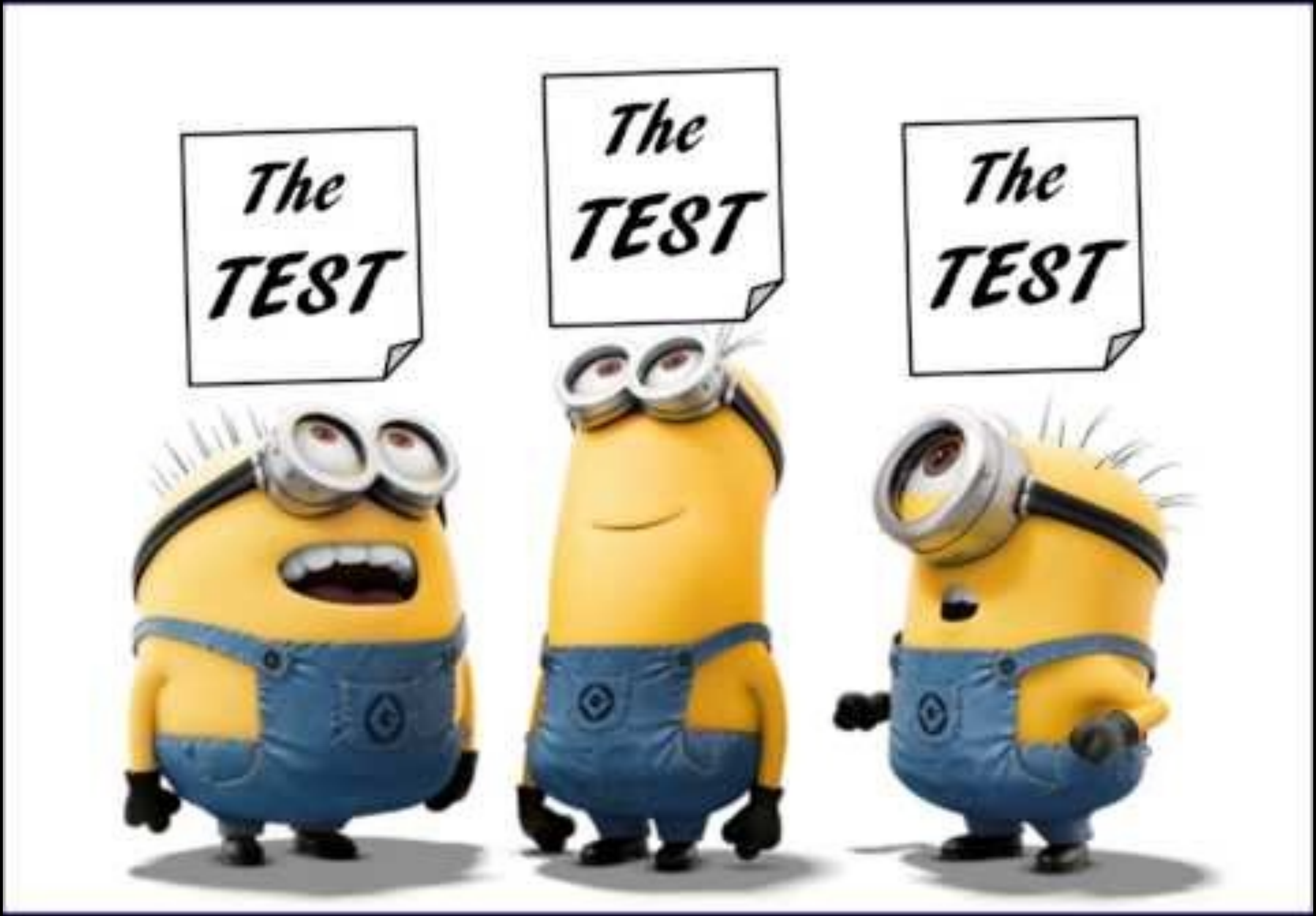
There are three sources of the failures:

★ Difficult images

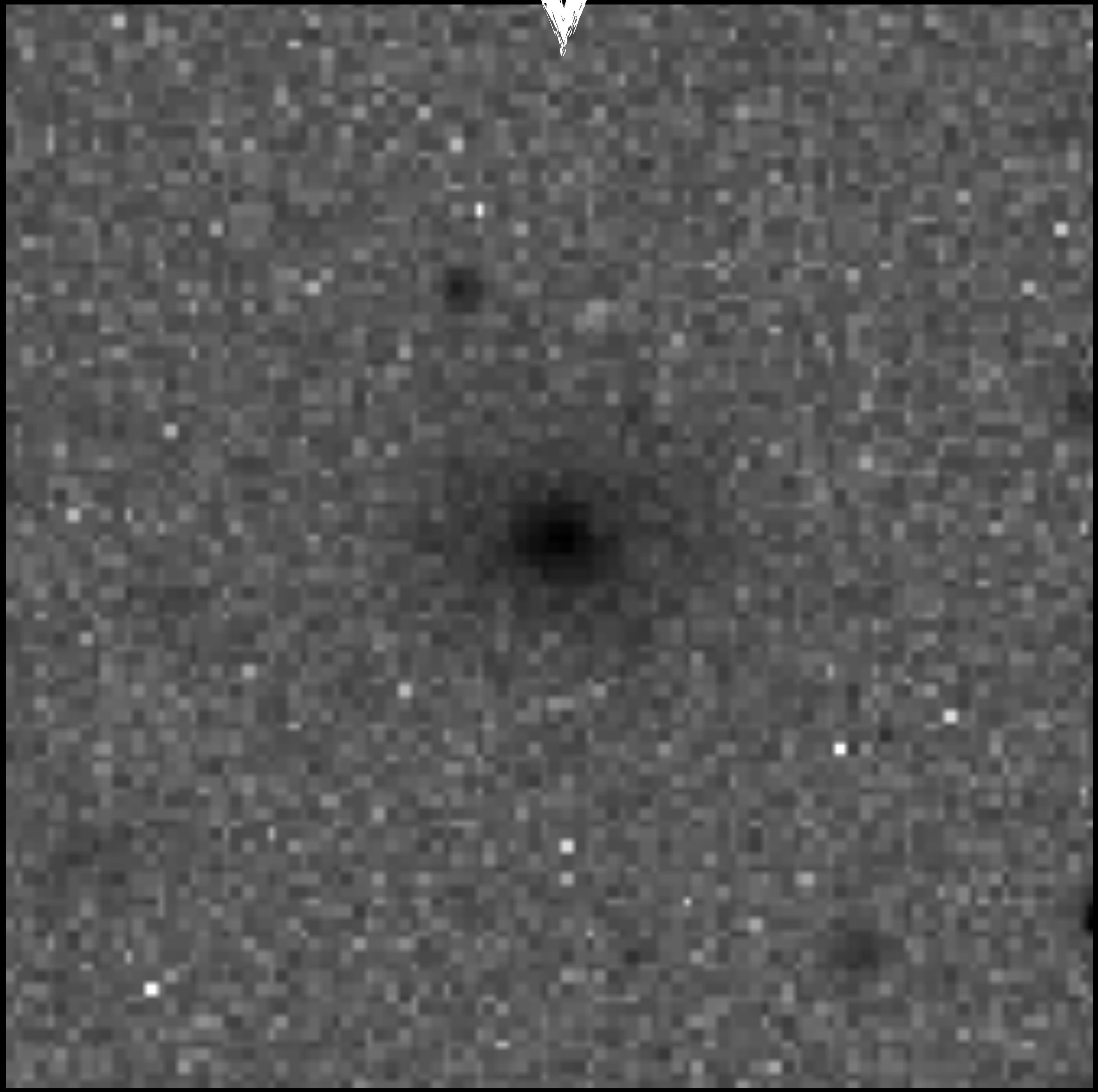
★ The problems from the initial labels

— The lack of the class of lenticular galaxy (S0)

★ The problems from our CNN

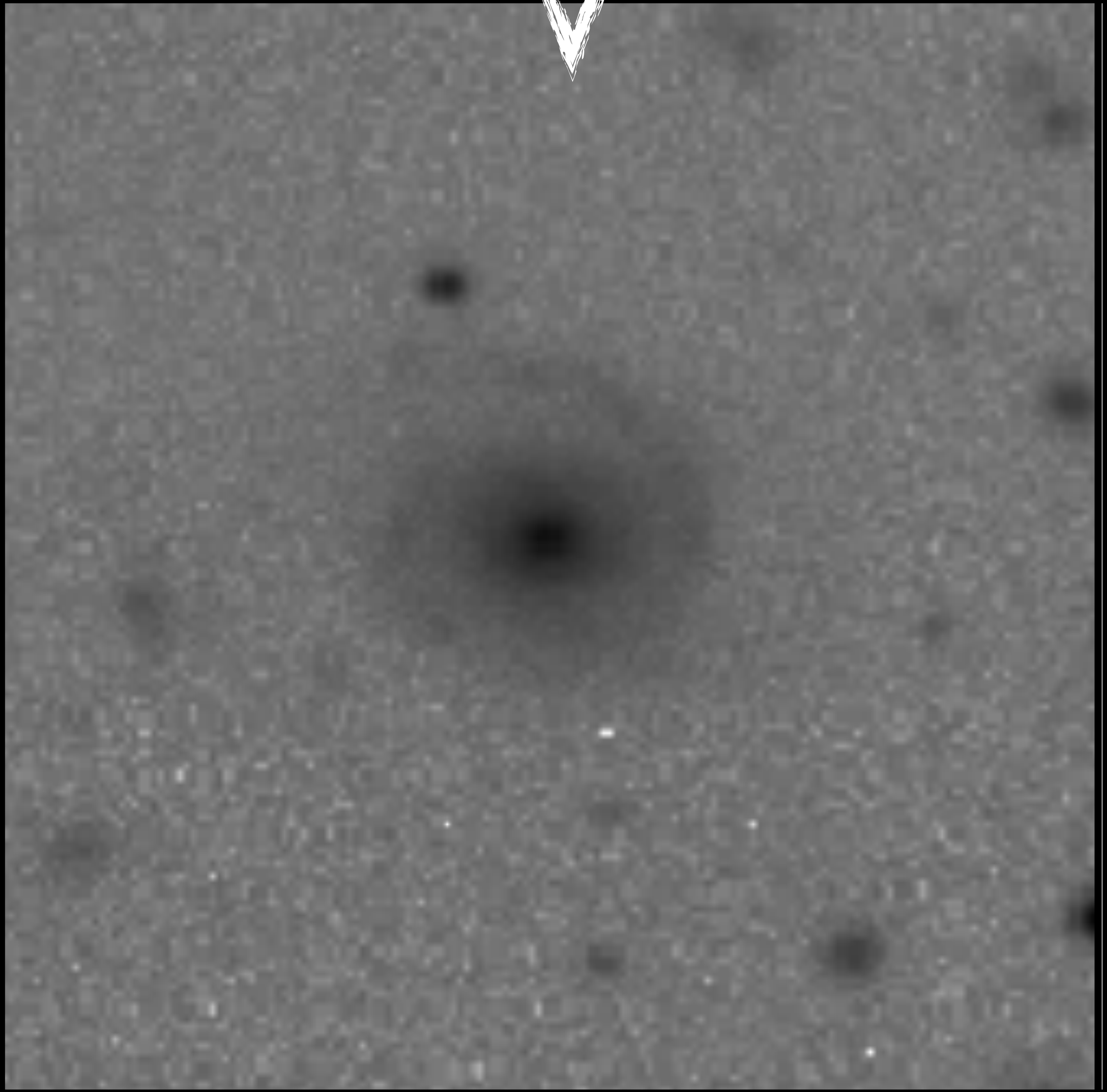

The TEST
The TEST
The TEST

Who am I?

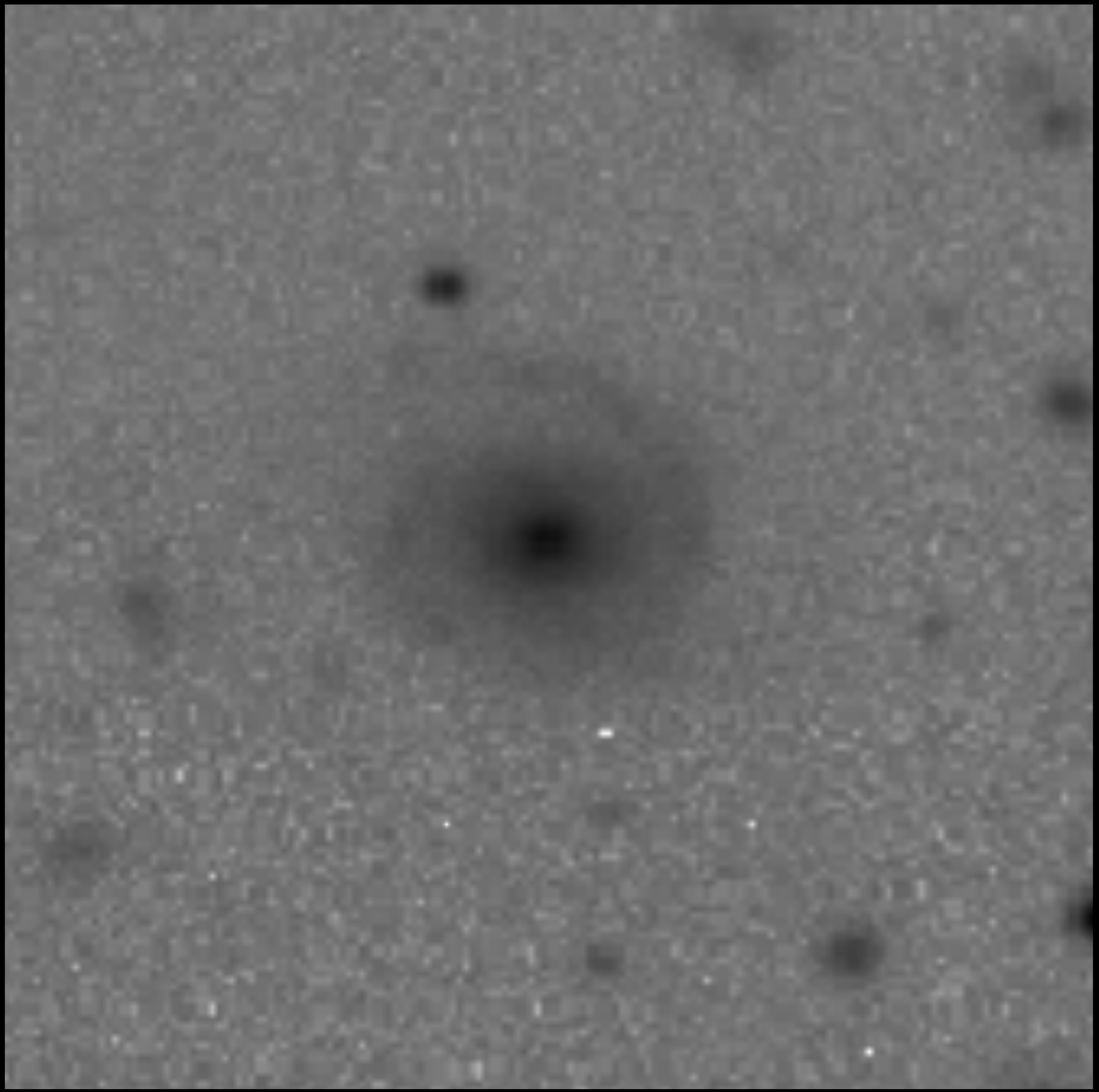


?

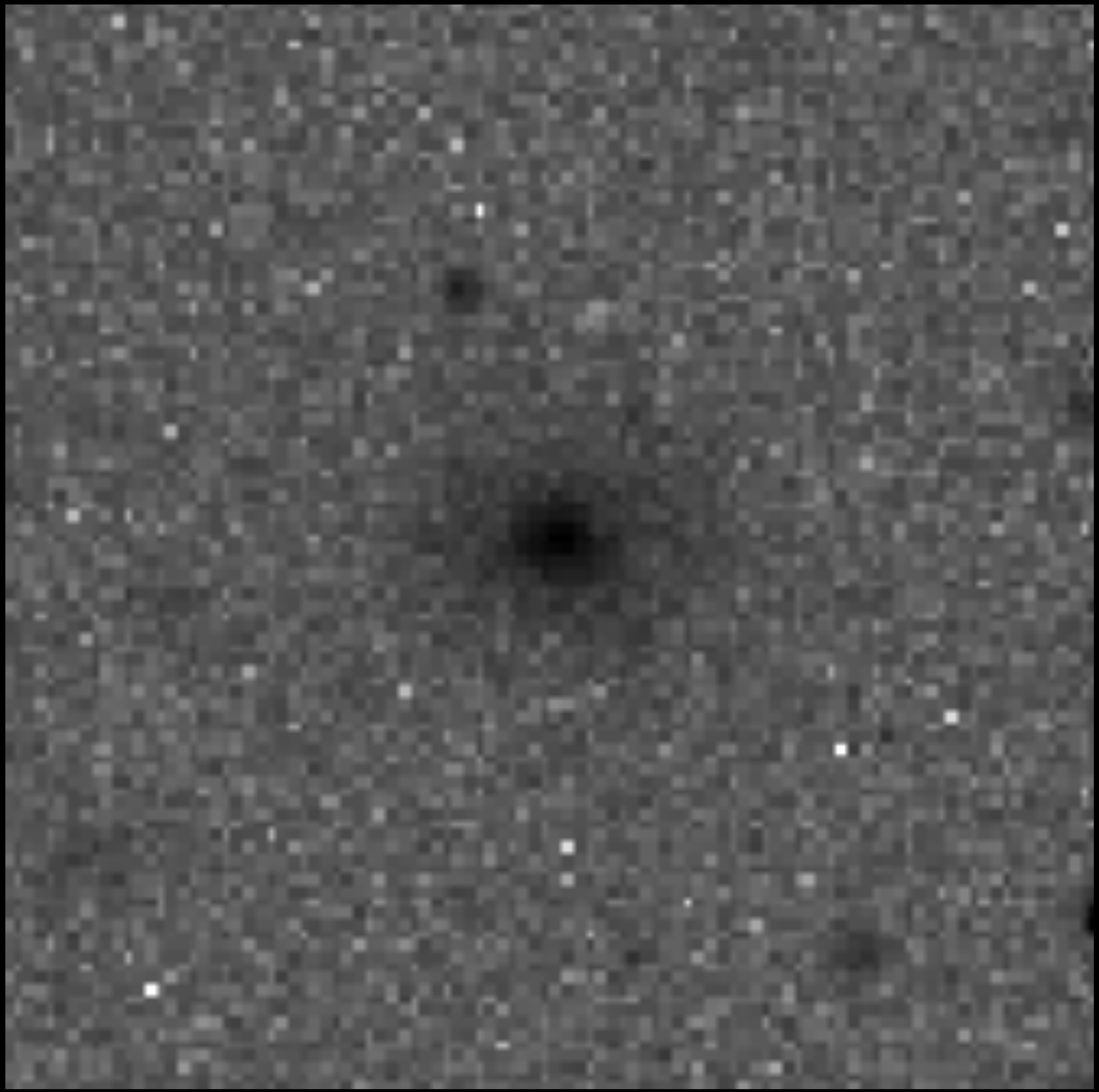
How about me?



?



Dark Energy Survey (DES)
Classification: Spirals
(By our CNN)



Sloan Digital Sky Survey (SDSS)
Classification: Ellipticals
(By Galaxy Zoo)

There are three sources of the failures:

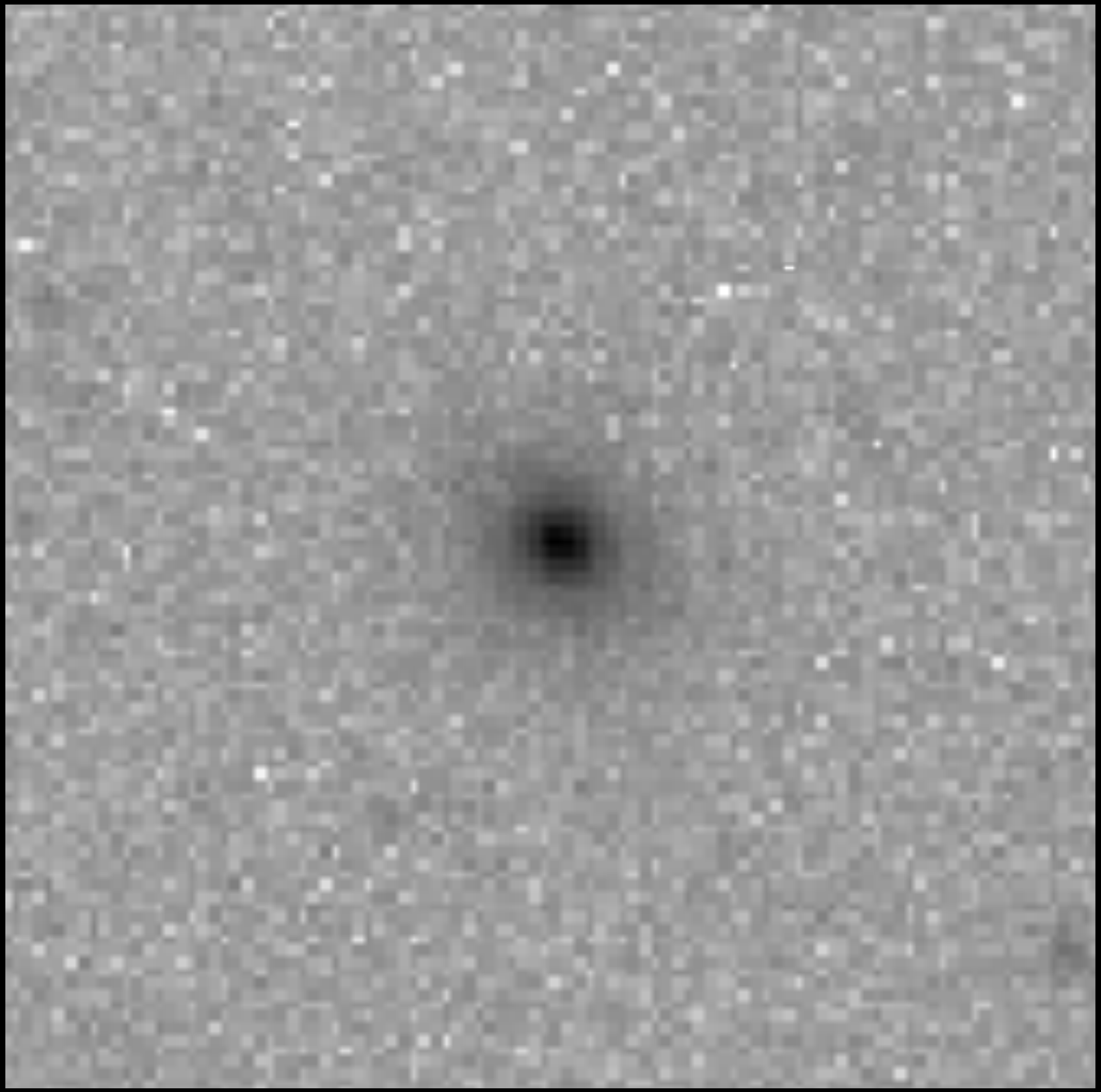
★ Difficult images

★ The problems from the initial labels

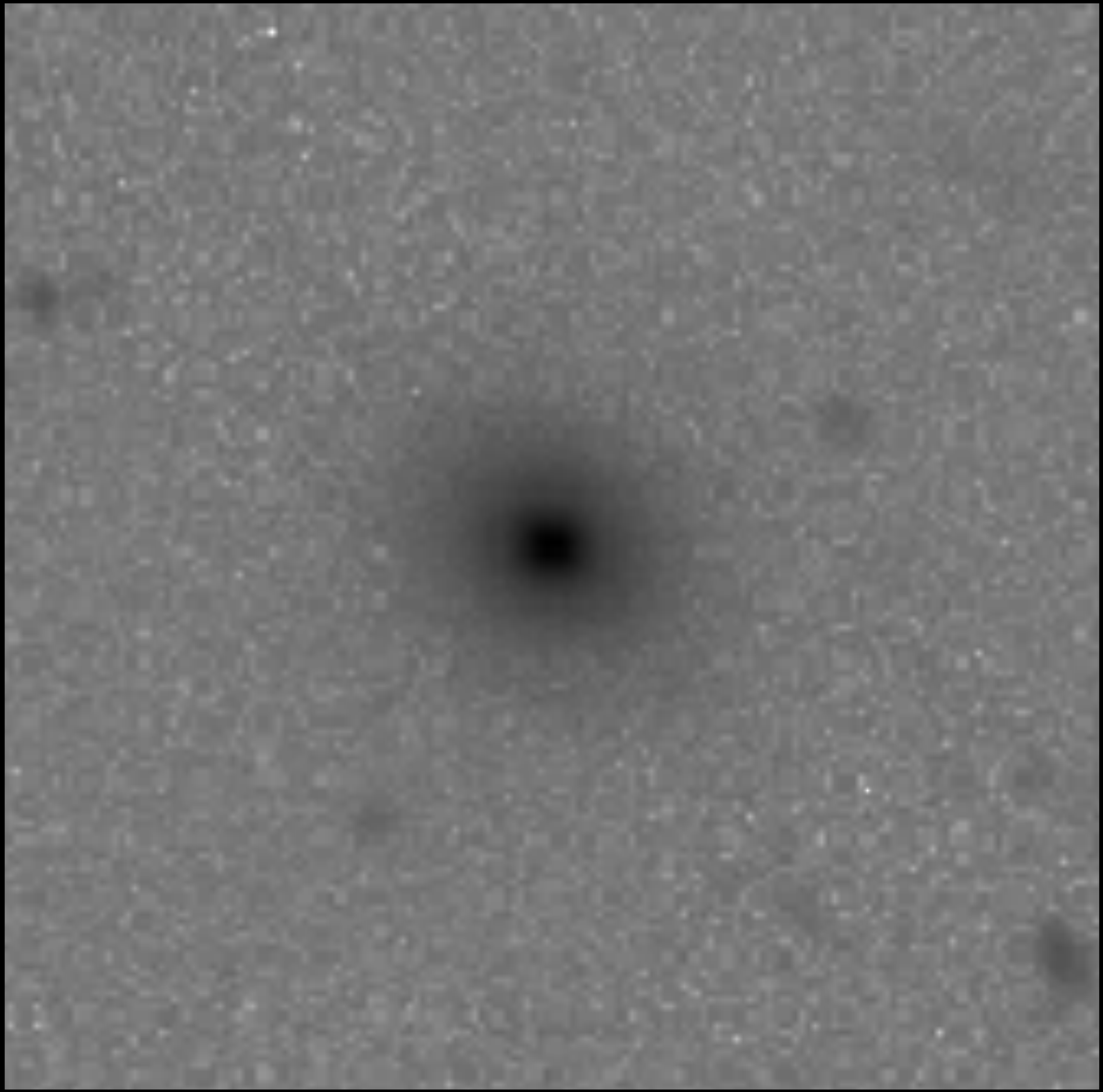
— The lack of the class of lenticular galaxy (S0)

— Better resolution of DES data reveals new features

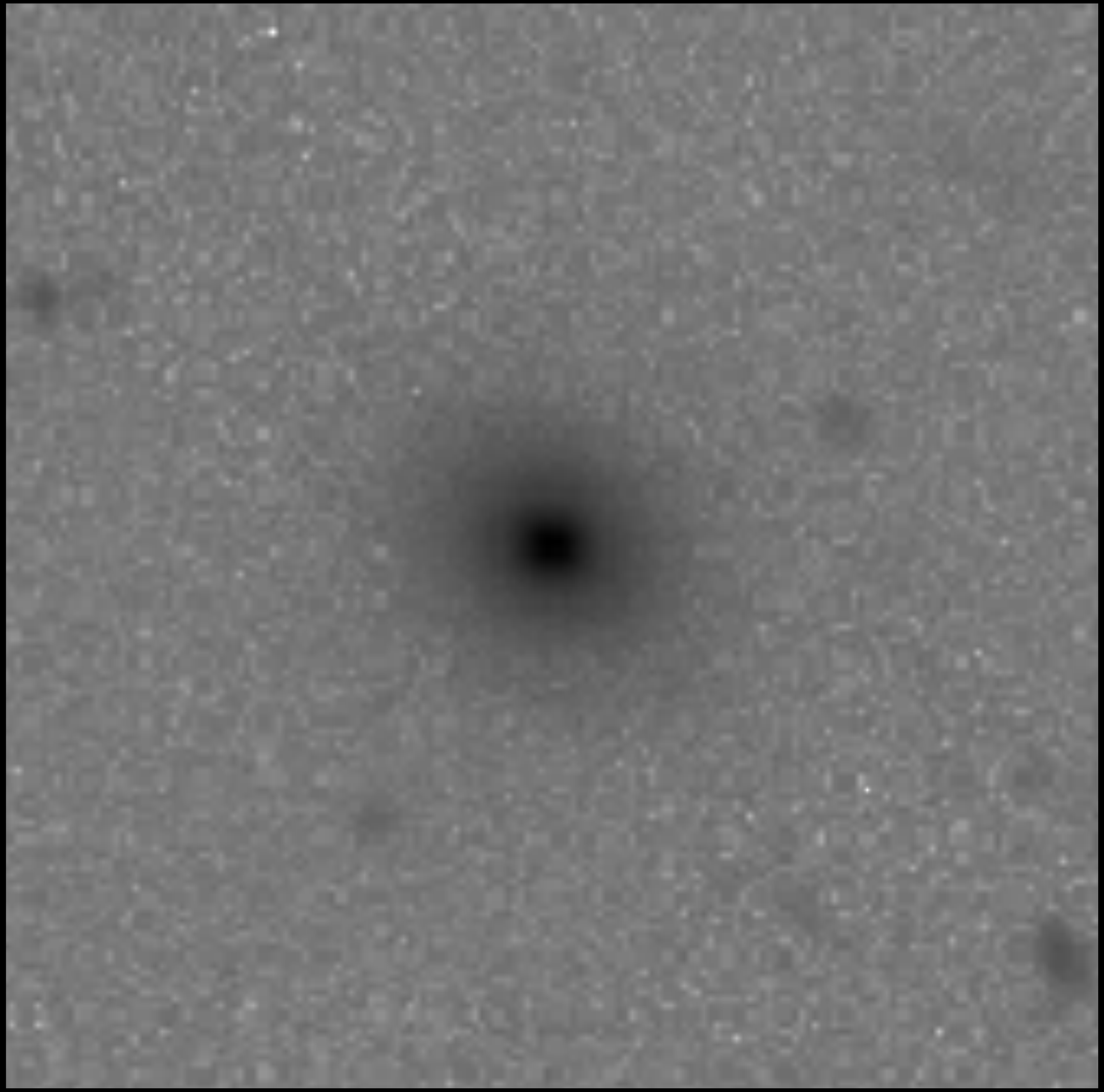
★ The problems from our CNN



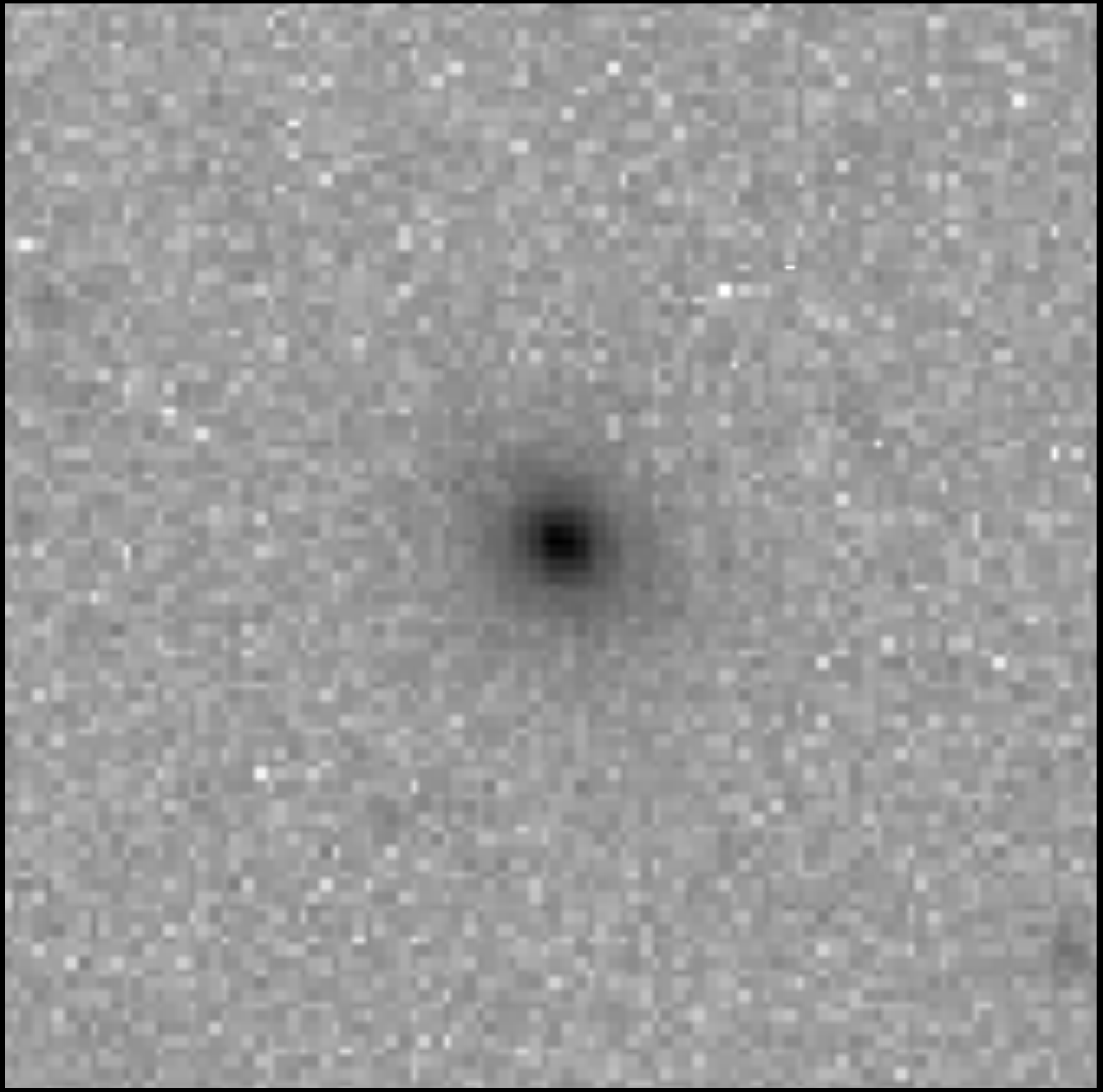
?



?

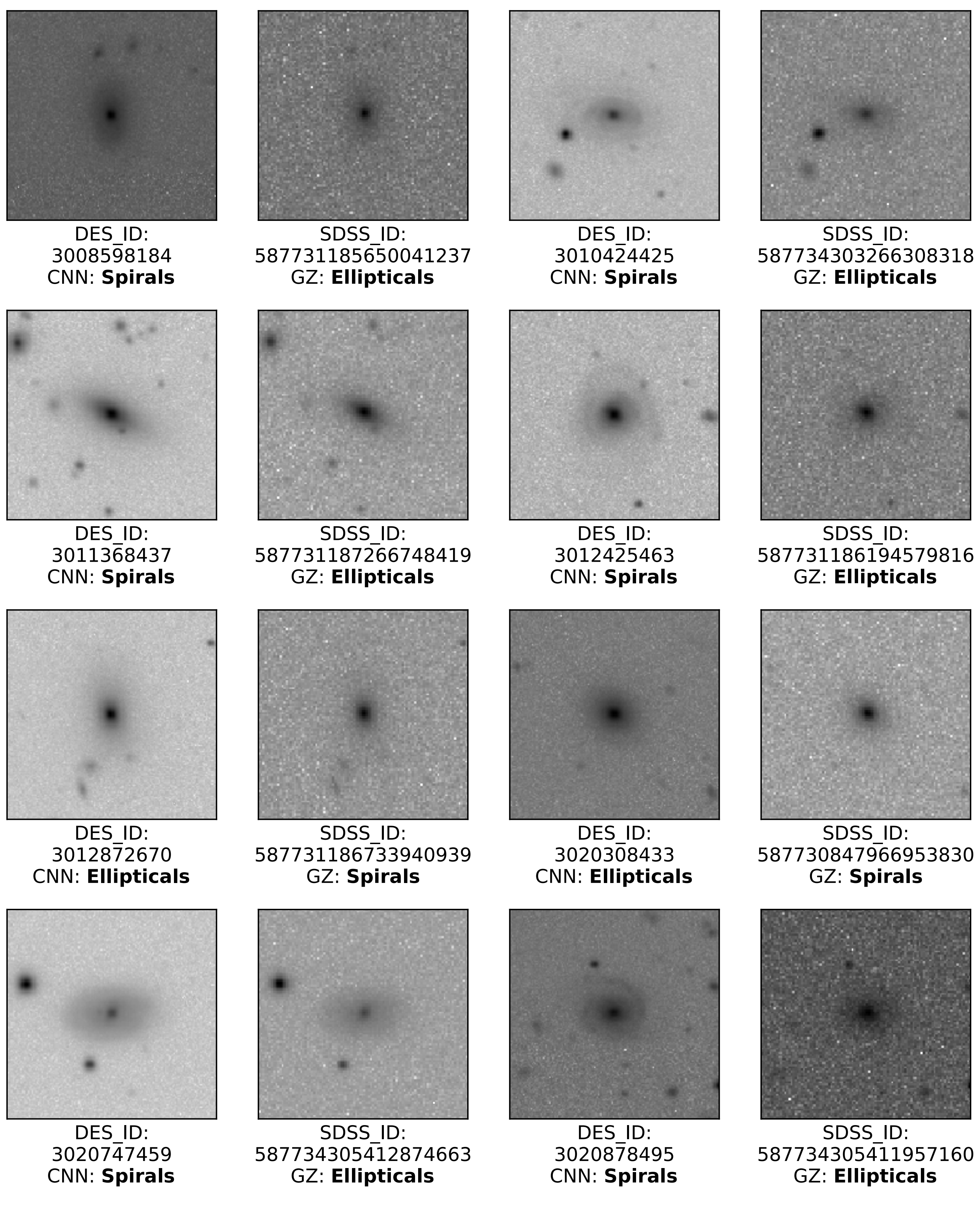


Dark Energy Survey (DES)
Classification: Ellipticals
(By our CNN)



Sloan Digital Sky Survey (SDSS)
Classification: Spirals
(By Galaxy Zoo)

Examples of the misclassification by Galaxy Zoo project :



There are three sources of the failures:

★ Difficult images

★ The problems from the initial labels

— The lack of the class of lenticular galaxy (S0)

— Better resolution of DES data reveals new features

— The misclassification by the Galaxy Zoo project

★ The problems from our CNN

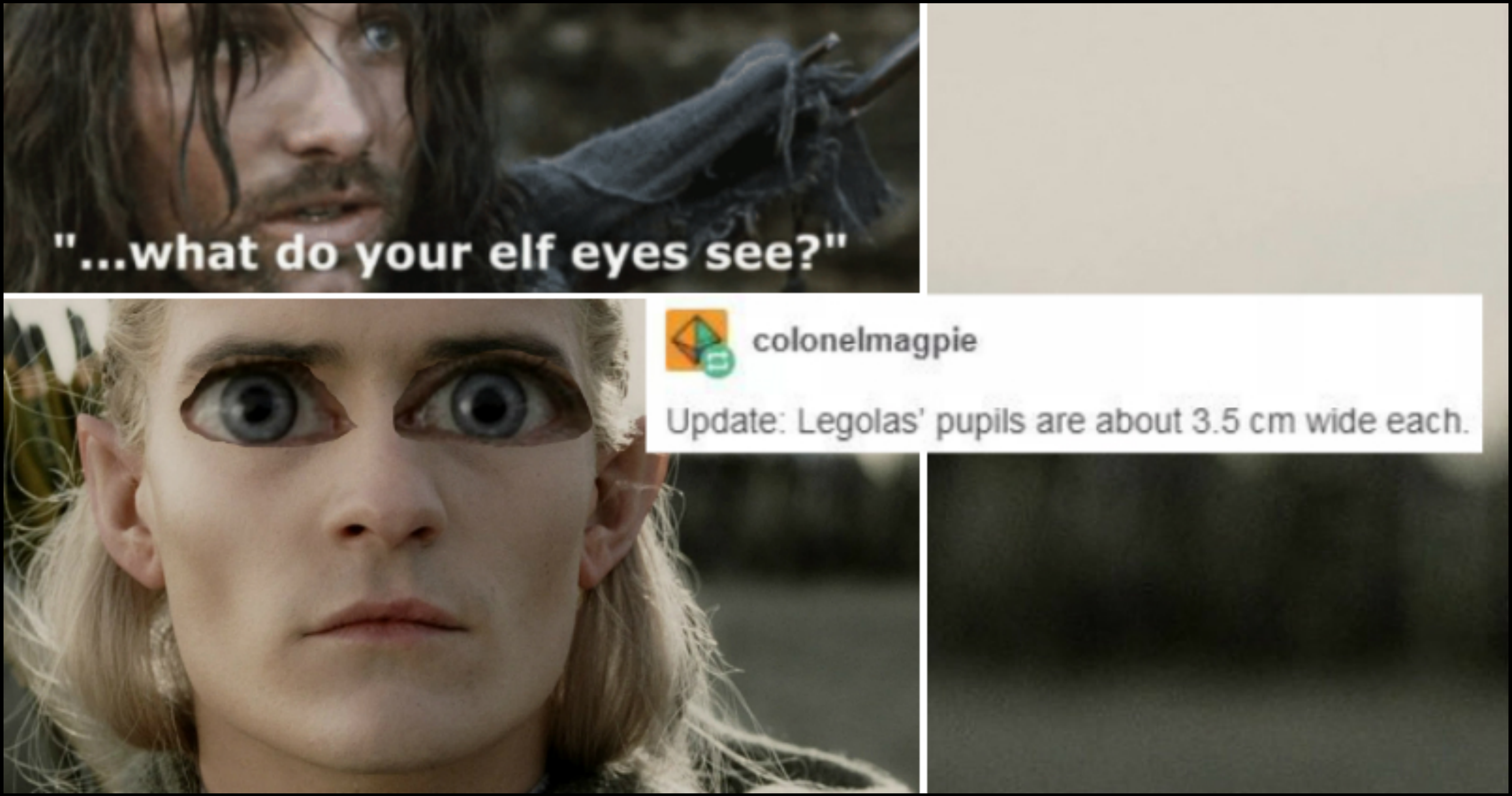
— The contamination in training set

— There is an uncertainty in CNN

What can we learn from these failures?

★ The limits of human visual classification.

– Tiny detail detection to the appearance of galaxy


"...what do your elf eyes see?"
colonelmagpie
Update: Legolas' pupils are about 3.5 cm wide each.

★ The limits of human visual classification.

— Tiny detail detection to the appearance of galaxy

➤ What is the difference between human mistakes and machine mistakes?

— Lenticular galaxy

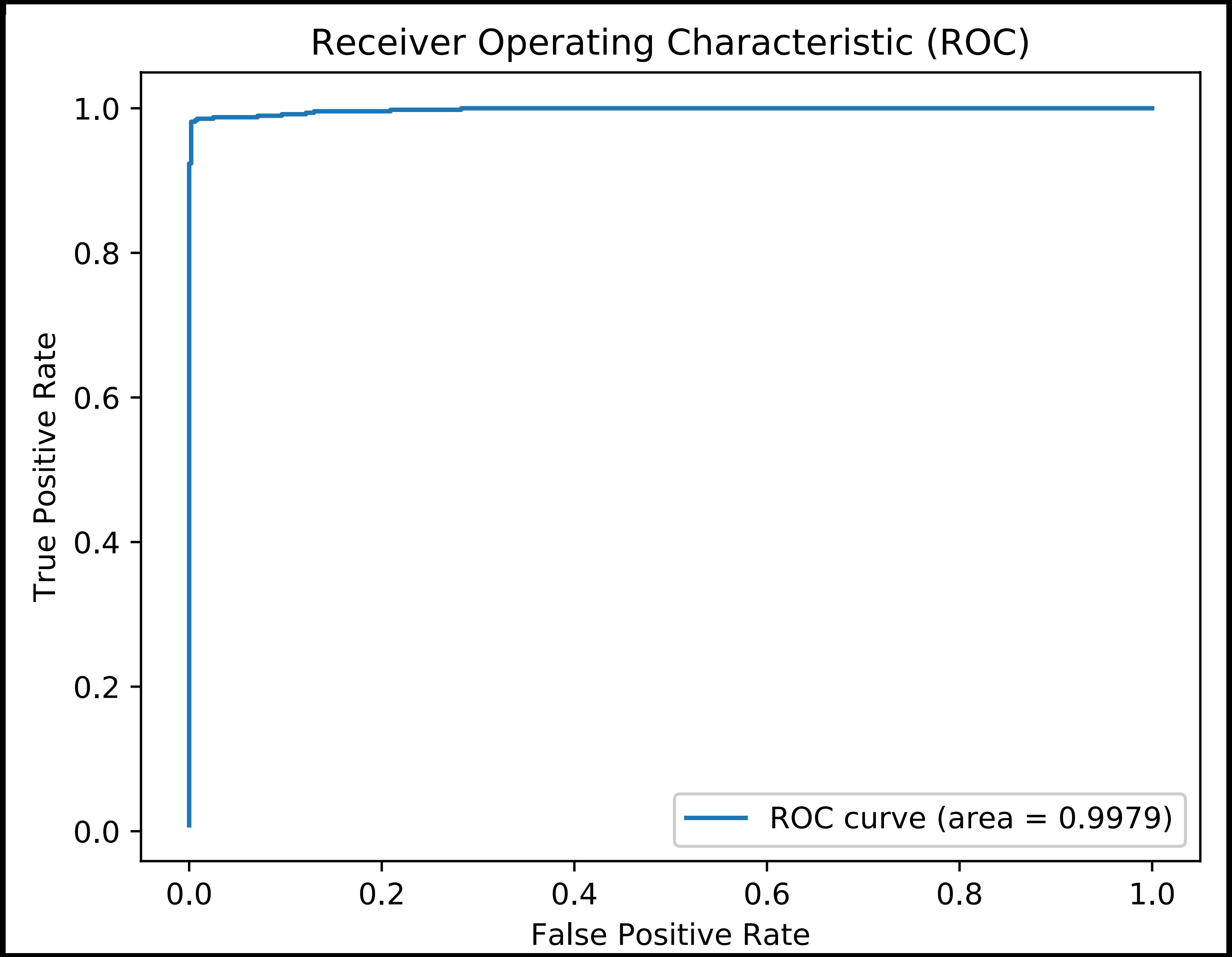
➤ Can we use the failures to create the class for lenticular galaxy?

(Make machine learn from the mistakes?)

★ To purify our training set

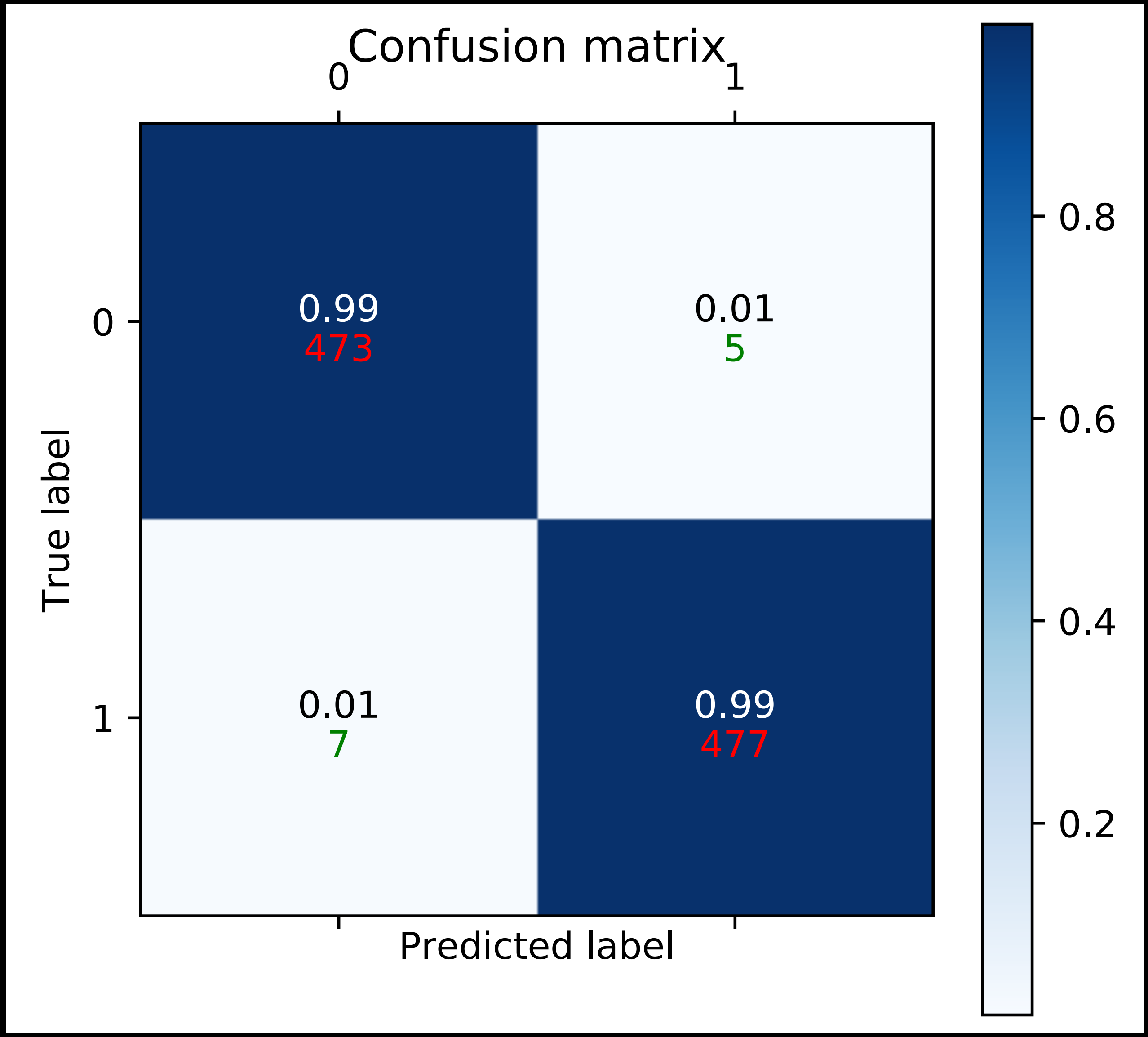
— Excluding the suspected misclassified galaxies by Galaxy Zoo project (both “resolution problem” and “error”), but keep potential lenticular galaxy.

— Retraining + retesting



Balanced Training set = 53663
Balanced Testing set = 1000

Probability criterion: p=0.8

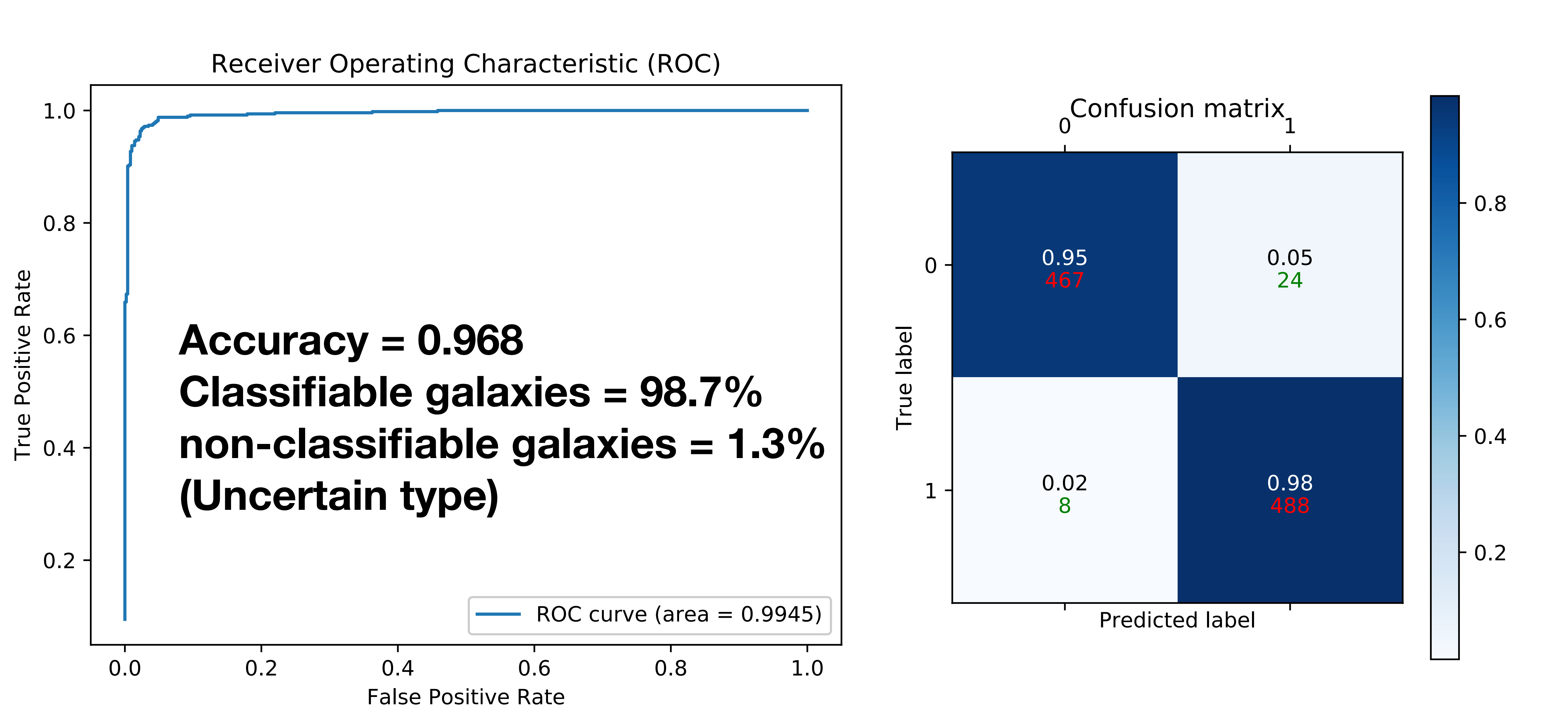


Accuracy = 0.987
Classifiable galaxies = 96%
non-classifiable galaxies = 4%
(Uncertain type)

★ To purify our training set

— Excluding the suspected misclassified galaxies by Galaxy Zoo project.

— Retraining + retesting

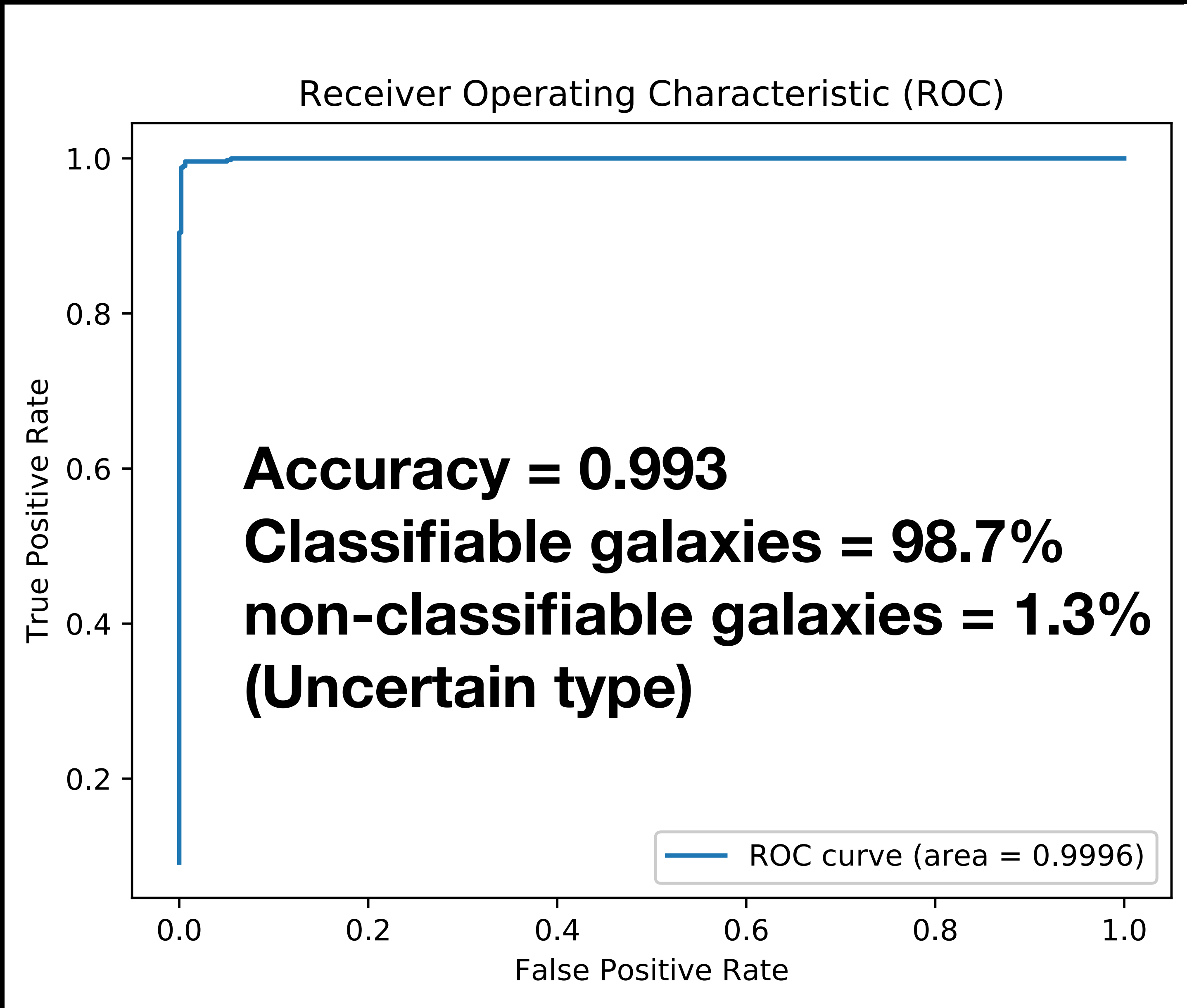


Accuracy = 0.968
Classifiable galaxies = 98.7%
non-classifiable galaxies = 1.3%
(Uncertain type)

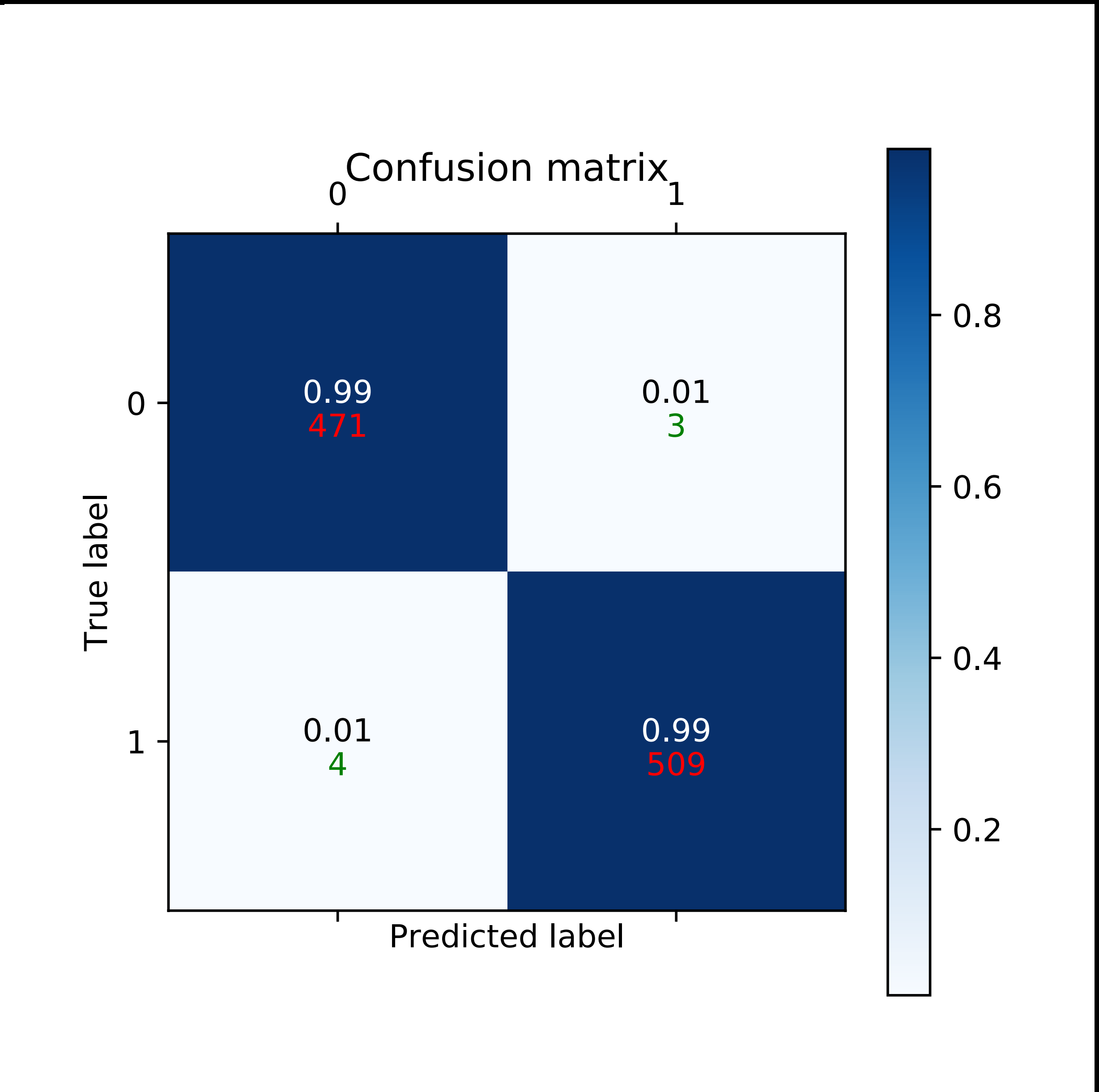
★ To purify our training set

— Excluding the suspected misclassified galaxies by Galaxy Zoo project.

— Retraining + retesting



Accuracy = 0.993
Classifiable galaxies = 98.7%
non-classifiable galaxies = 1.3%
(Uncertain type)



★ To purify our training set

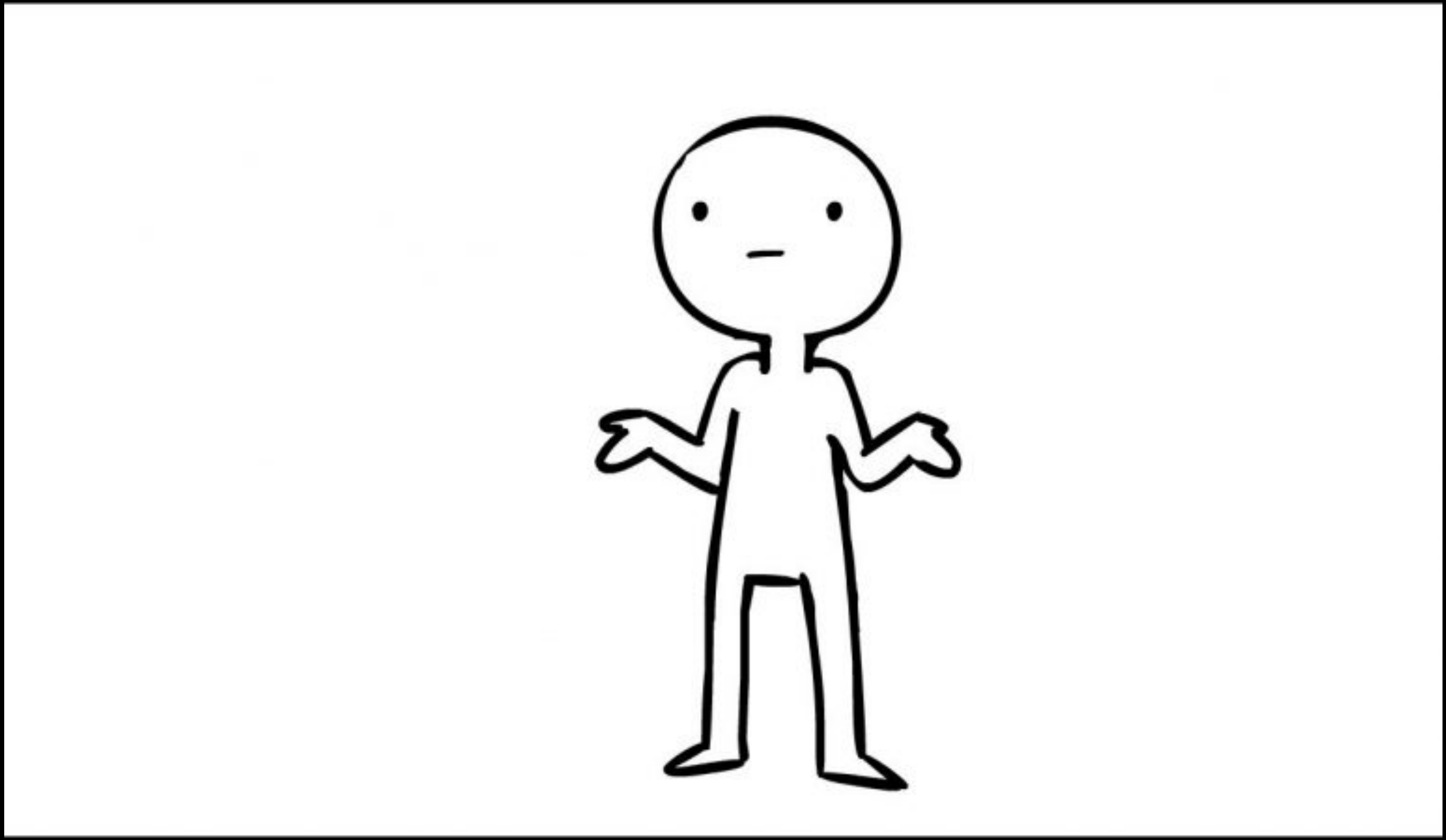
— Excluding the suspected misclassified galaxies by Galaxy Zoo project (both “resolution problem” and “error”), but keep potential lenticular galaxy.

— Retraining + retesting

— Showing up ≥ 3 times in failures with high probabilities within 5 reruns

→ **Misclassification (by GZ):**
Testing set: ~42 (~4.2%)
Training set: ~54 (~2.9%)
Total: ~3.35%
(1.36% Spirals by GZ/ 1.99% Ellipticals by GZ)

lenticular galaxy?



★ After purifying our training set, it can improve our CNN results from accuracy ~0.987 to ~0.993. The number of uncertain type decrease from 4% to 1.3%

★ We found some classifications from Galaxy Zoo project need to be updated.

— Can we use these failure investigation to modify them?

★ A class for lenticular galaxy

— The setting of classification system is of great importance.

— What do you want your machine learning to do?

— What does your machine learning actually do?

☛ All the results and discussion will be published in my first paper! (**Cheng et al. in progress**)
(btw, my real name is Ting-Yun Cheng.)

☛ We are building on a catalogue of galaxy morphology for Dark Energy Survey images data by our CNN.
(I am still trying to find a way to separate a group of S0.)

☛ We are working on the Unsupervised Machine Learning, e.g. Fuzzy K-mean, Self-Organised Map, etc.

Enjoy your trip in Machine Learning!
Thank you for the listening.