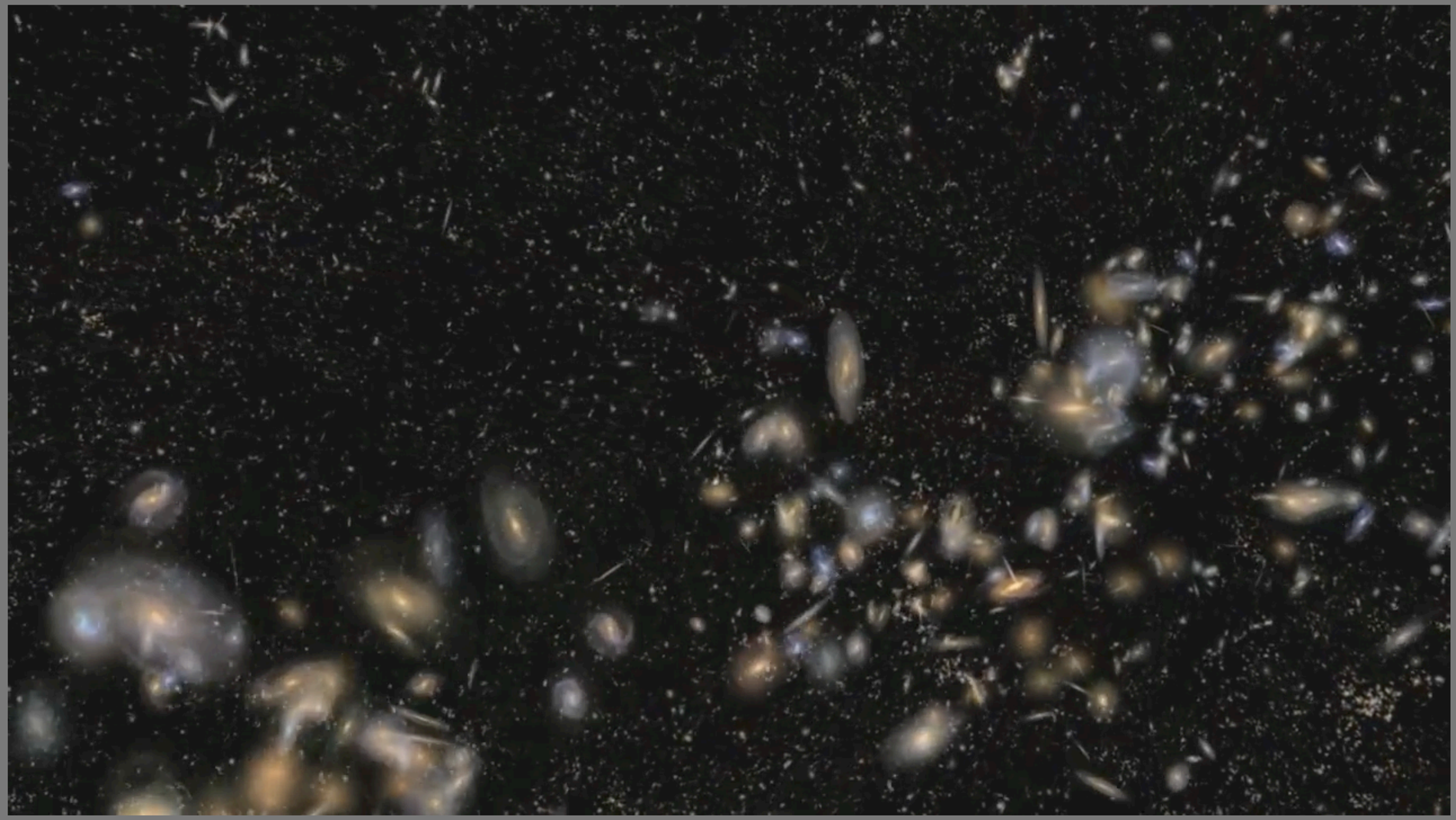
https://github.com/igc5972/comp_meth/tree/master/comp_final

Logistical Regression

For Color-Mass Diagrams

Isabella G. Cox
ASTP 720 Computational Methods
Fall 2020

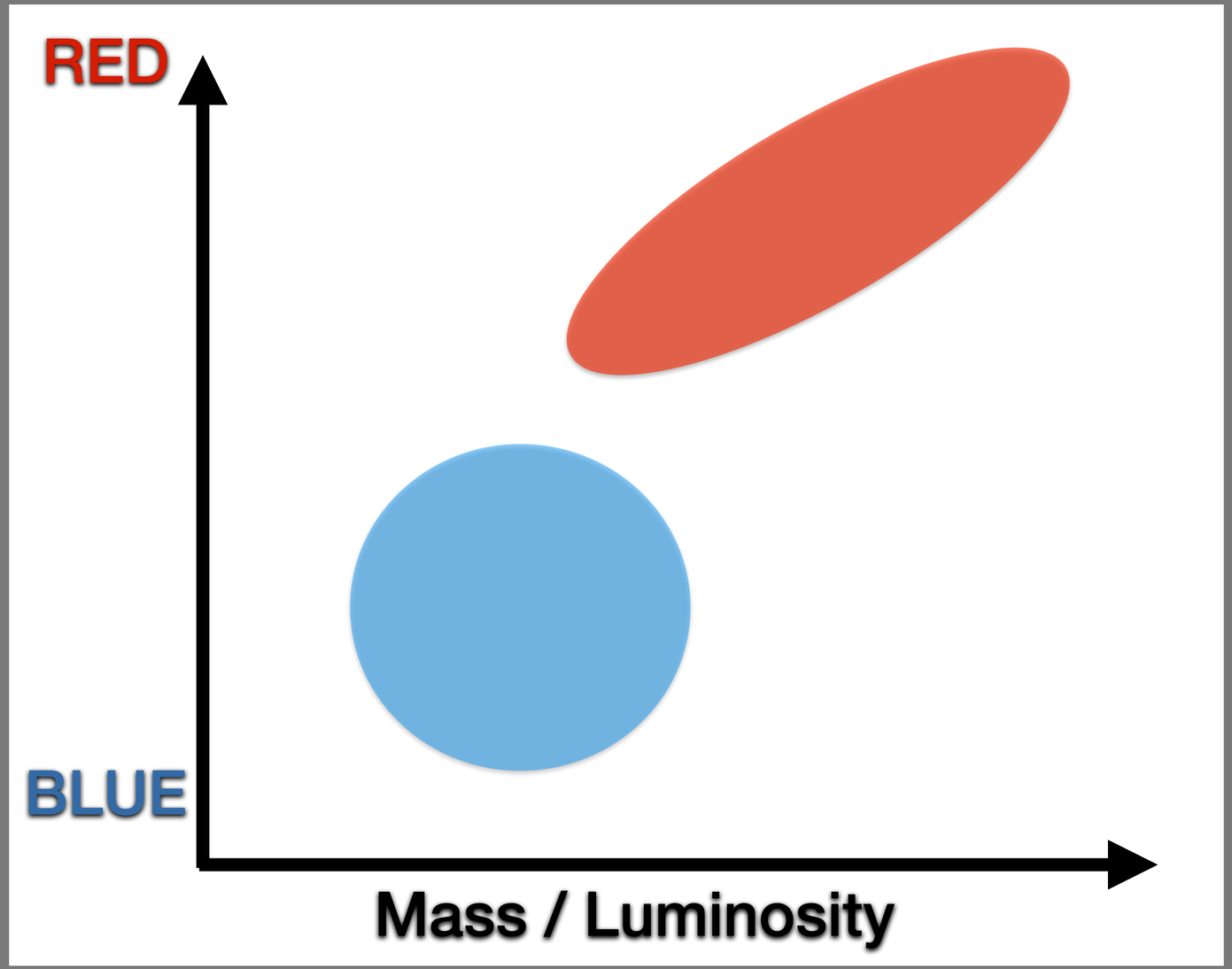
Sloan Digital Sky Survey



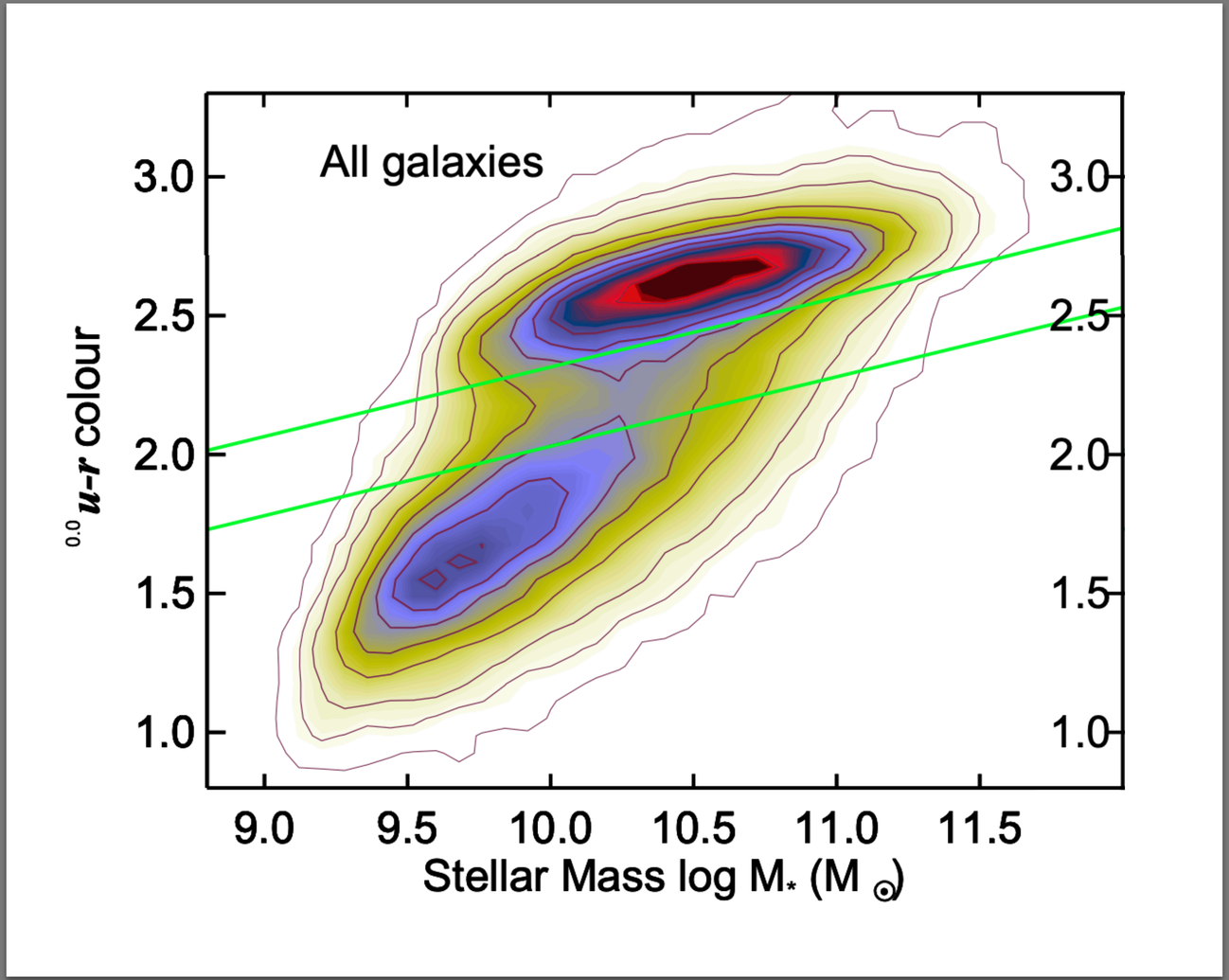
How do we sort through this vast amount of scientific data?

Look at *samples* of many galaxies, instead of individuals.

Color Bi-Modality of Galaxies

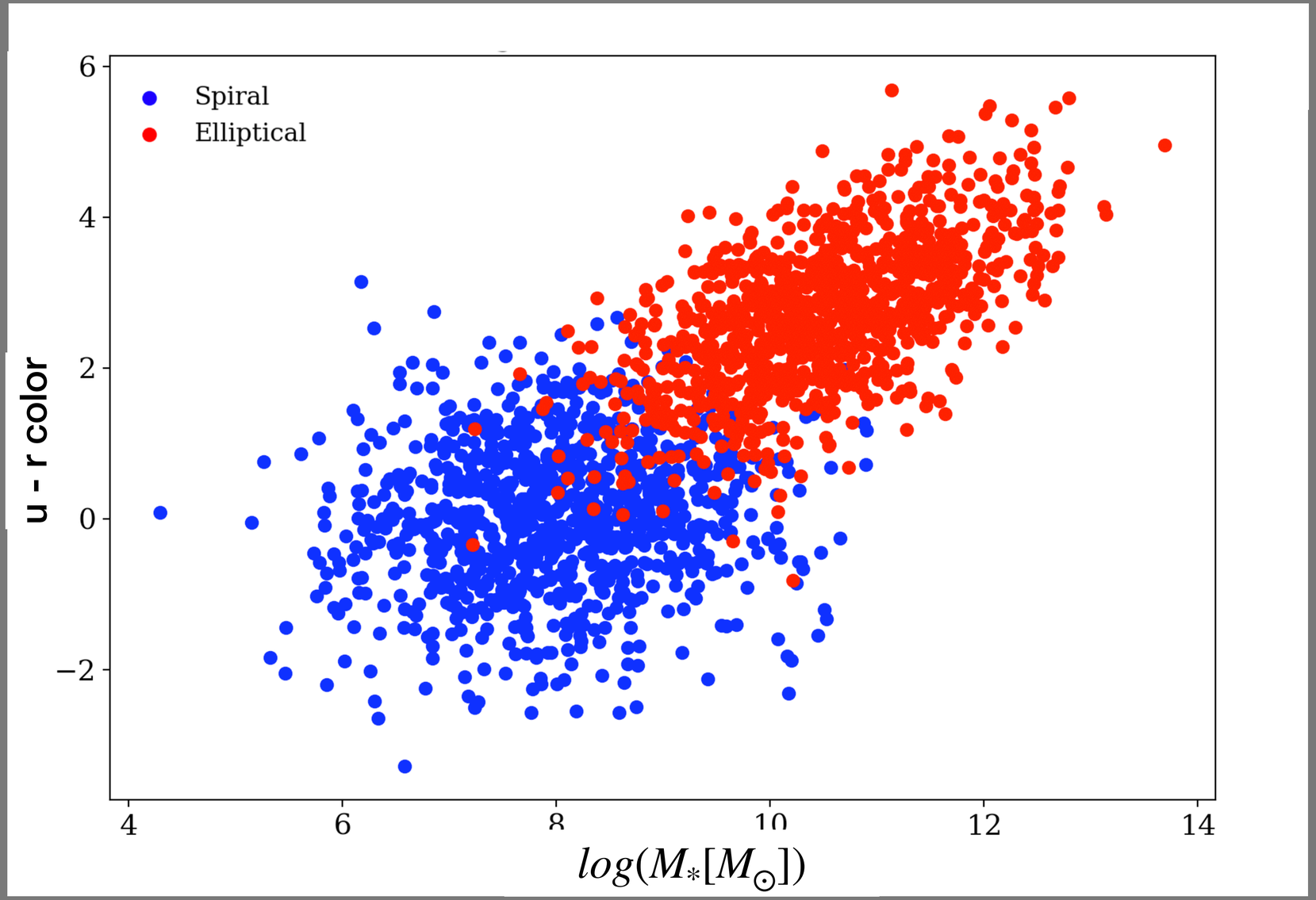


Color - Mass Diagram



Schawinski et. al, MNRAS 400.1, 2014

My “Simulated” SDSS Data



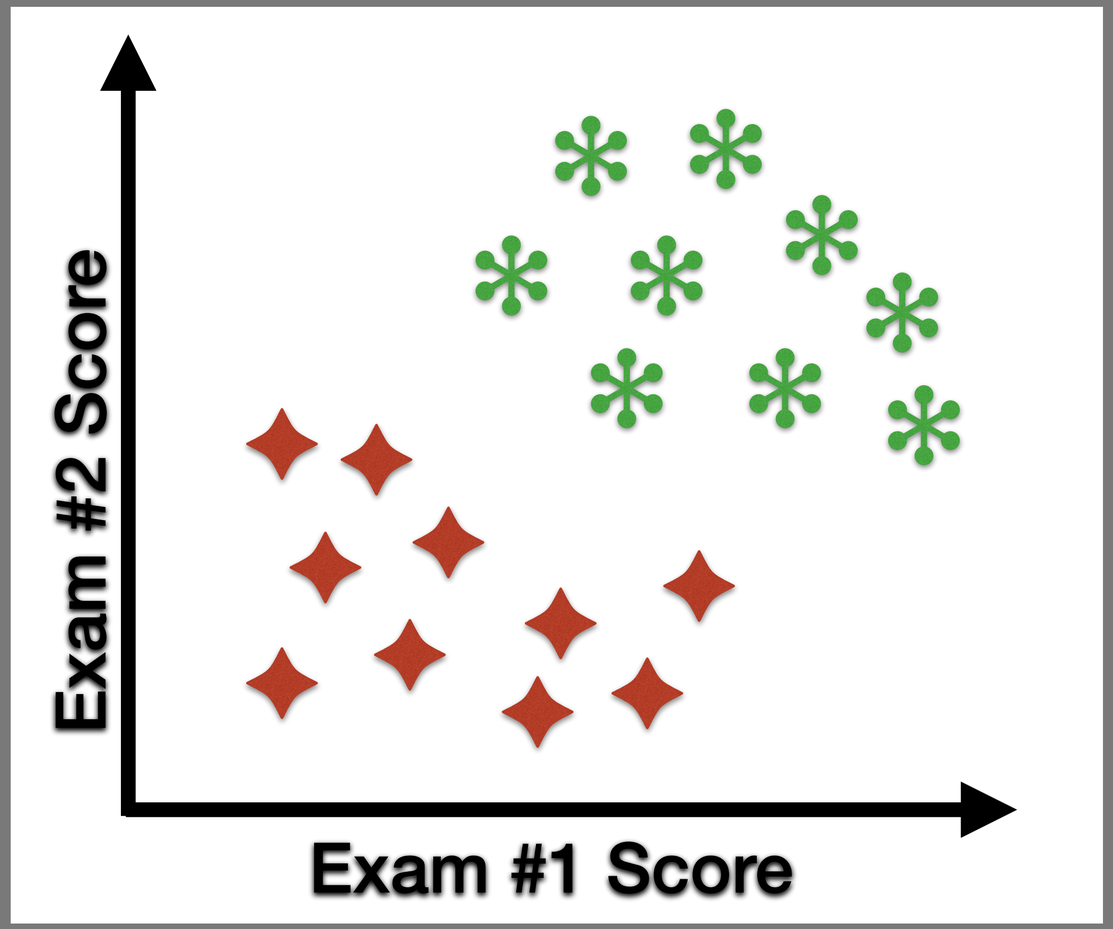
Science Question

If I assign labels to the simulated galaxies, can I correctly predict their label by drawing a line demarcating the data of a color-mass diagram into two populations?

Logistic Regression Example

Classification Algorithm. The dependent variable is predicted label for each datapoint, determined with respect to demarcation line separating populations

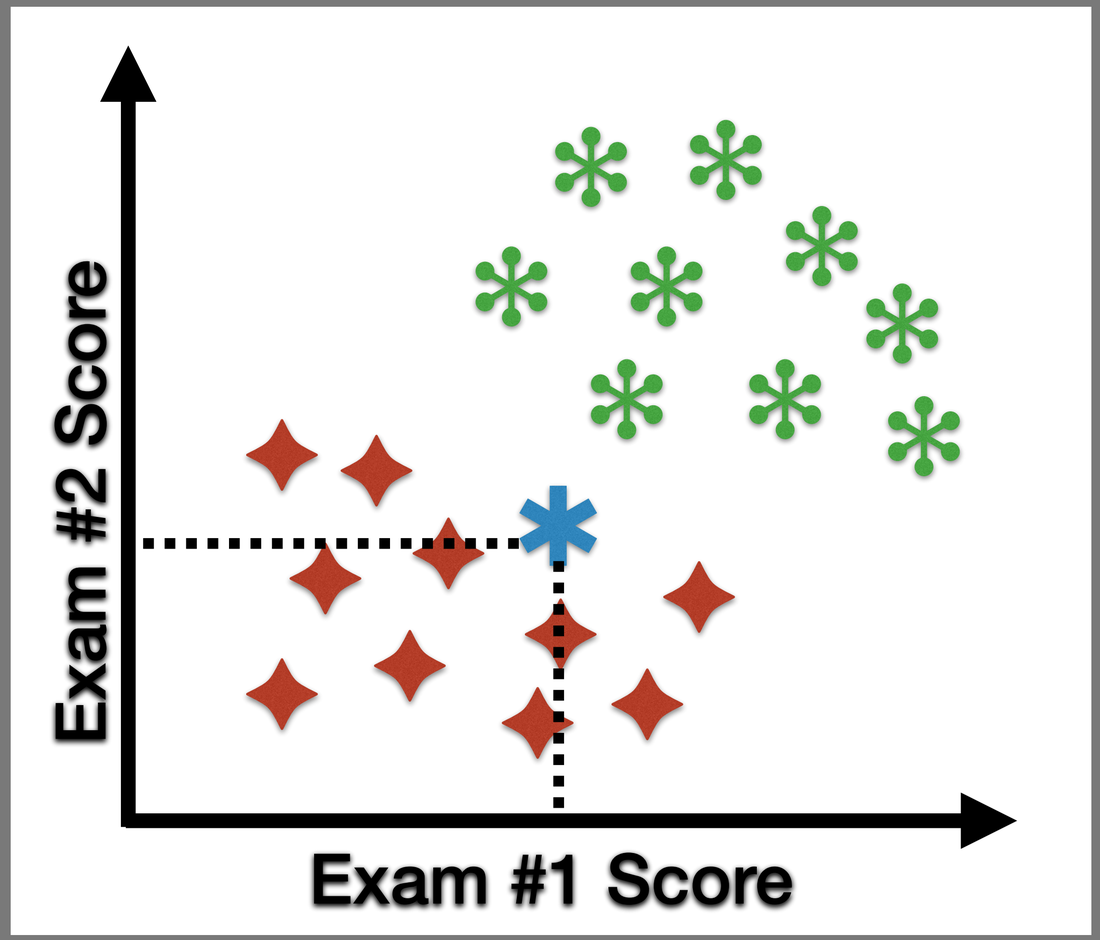
- Admitted Students
- Rejected Students



Logistic Regression Example

Classification Algorithm. The dependent variable is predicted label for each datapoint, determined with respect to demarcation line separating populations

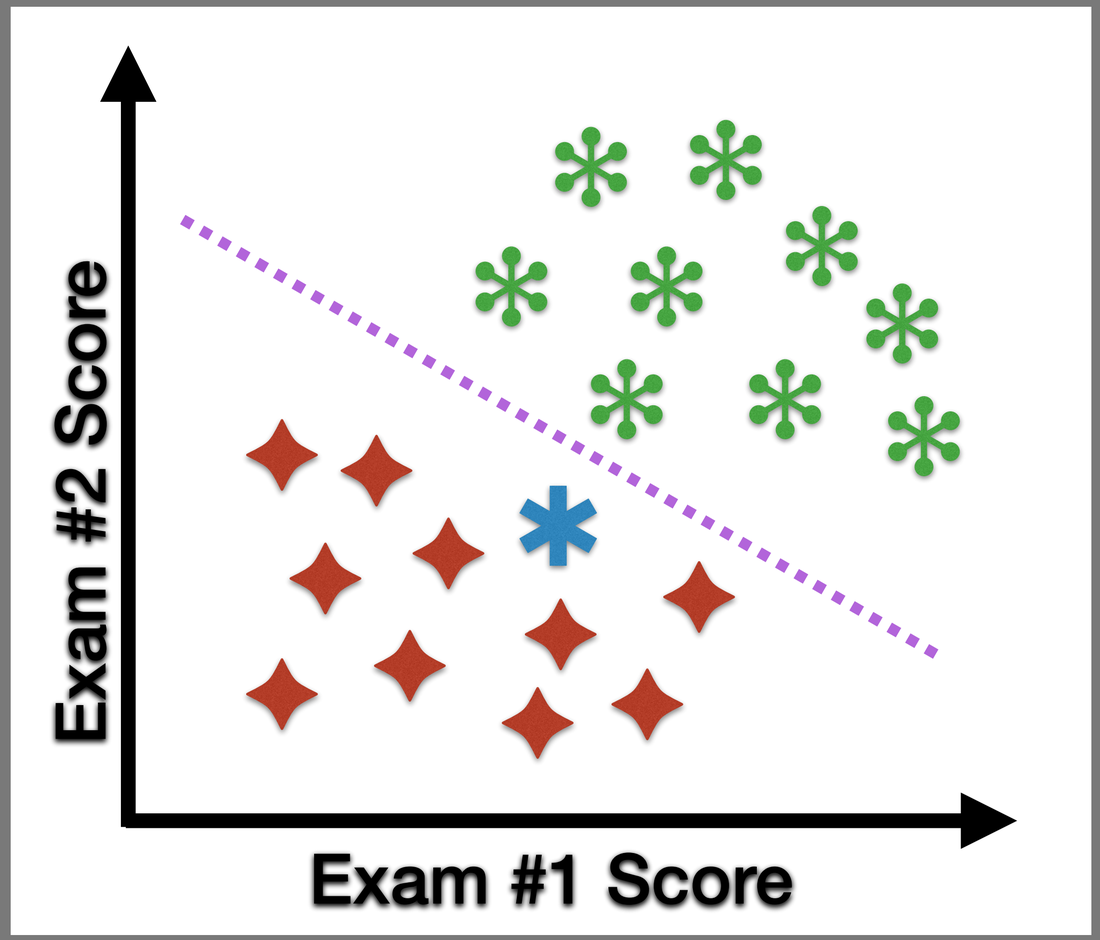
- Admitted Students
- Rejected Students
- New Student Data



Logistic Regression Example

Classification Algorithm. The dependent variable is predicted label for each datapoint, determined with respect to demarcation line separating populations

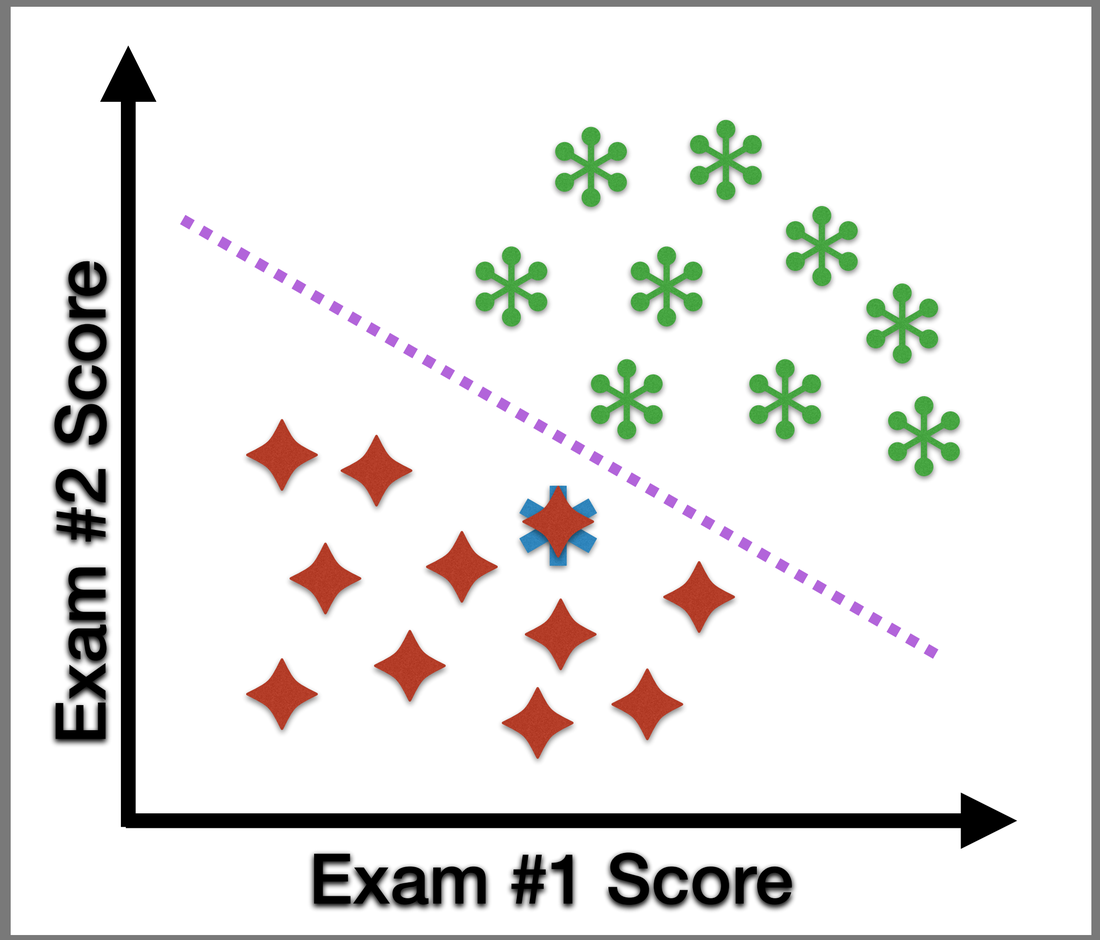
- Admitted Students
- Rejected Students
- New Student Data



Logistic Regression Example

Classification Algorithm. The dependent variable is predicted label for each datapoint, determined with respect to demarcation line separating populations

- Admitted Students
- Rejected Students
- New Student Data



Terminology

θ = parameters	y = labels	X = features
$\theta_1, \theta_2, \theta_3$ Will be used to construct eq. of line for decision boundary.	Elliptical [1]  Spiral [0] 	X_1: Color [float] X_2: Stellar Mass [float]

The Plan

Go through a large number of iterations, and for each step…..

The Plan

Go through a large number of iterations, and for each step…..

(1) Calculate the **sigmoid functional value** based on current θ values

The Plan

Go through a large number of iterations, and for each step…..

(1) Calculate the **sigmoid functional value** based on current θ values
(2) Calculate the corresponding **gradient rate** to minimize the **cost function**

The Plan

Go through a large number of iterations, and for each step…..

(1) Calculate the **sigmoid functional value** based on current θ values
(2) Calculate the corresponding **gradient rate** to minimize the **cost function**
(3) Decrease θ by the gradient rate times the learning rate

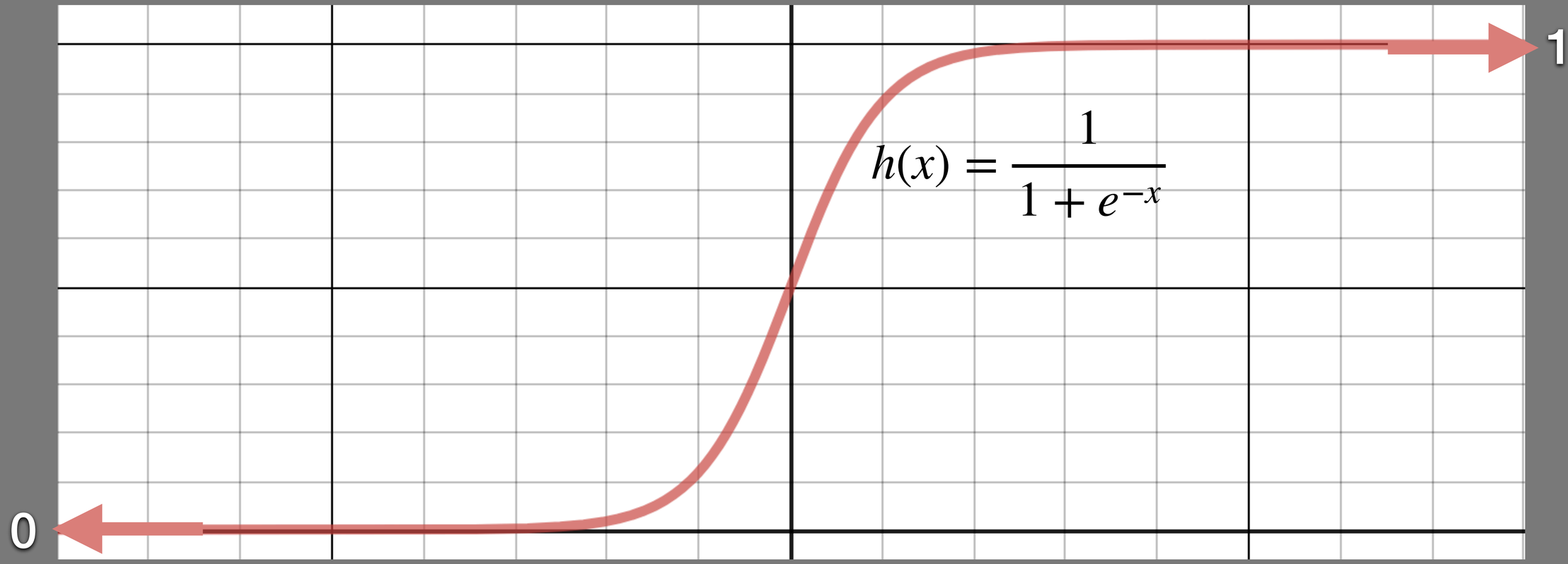
The Plan

Go through a large number of iterations, and for each step…..

(1) Calculate the **sigmoid functional value** based on current θ values
(2) Calculate the corresponding **gradient rate** to minimize the **cost function**
(3) Decrease θ by the gradient rate times the learning rate

Once we have final θ values, we can make predictions by calculating sigmoid functional value for these final θs and then use that to cast each entry to 0 or 1 (the binary label).

Function: Hypothesis Function

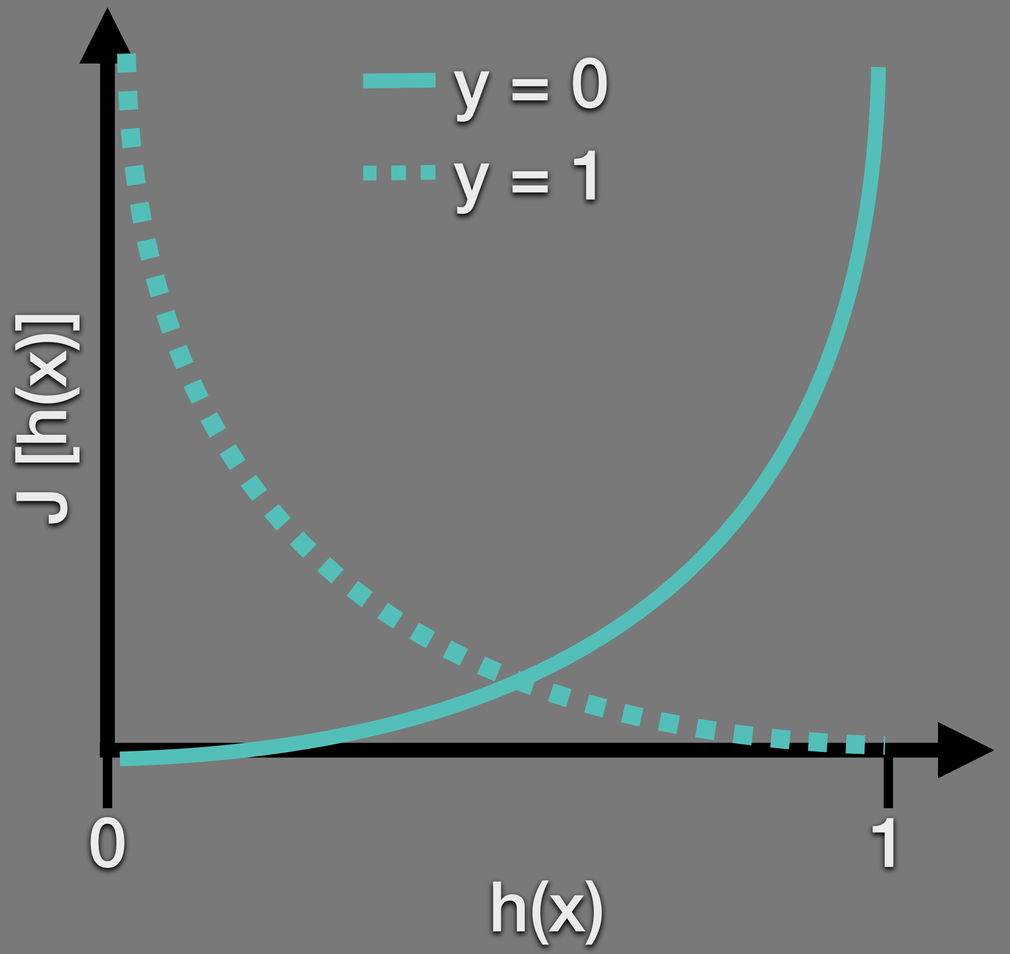


Function: Cost Function

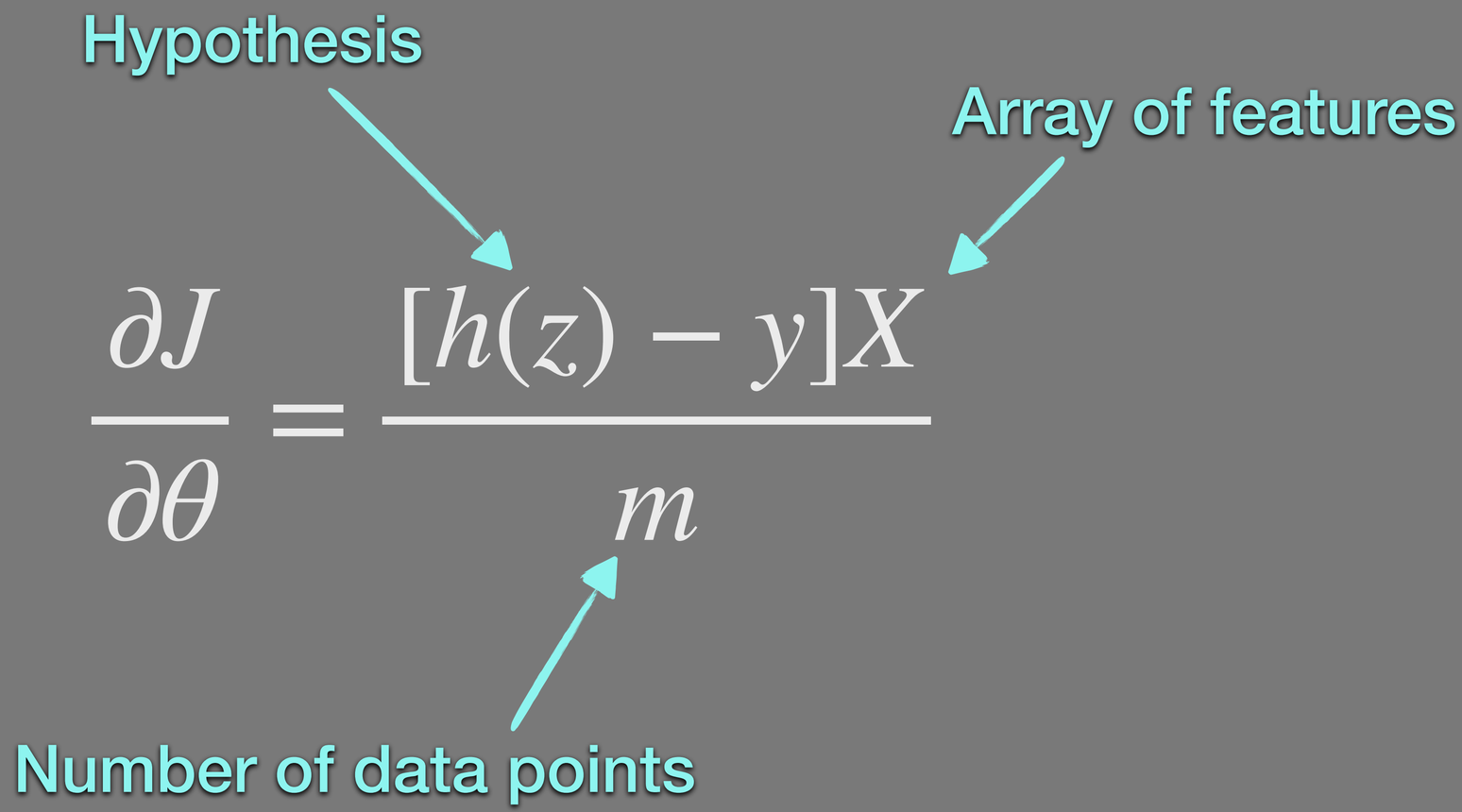
$$J(\theta) = \begin{cases} -log[h(x)], & if\ y = 1 \\ -log[1 - h(x)], & if\ y = 0 \end{cases}$$

Function: Cost Function (another view)

$$J(\theta) = \begin{cases} -log[h(x)], & if\ y = 1 \\ -log[1 - h(x)], & if\ y = 0 \end{cases}$$



Function: Gradient



Putting it all together

for i in range(no_iter):

    z = Xθ^T

    h = sigmoid(z)

    grad = gradient(X, h, y)

    θ -= rate * grad

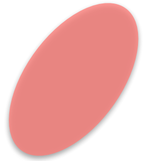
OUTPUT: θ has been populated

Making Predictions

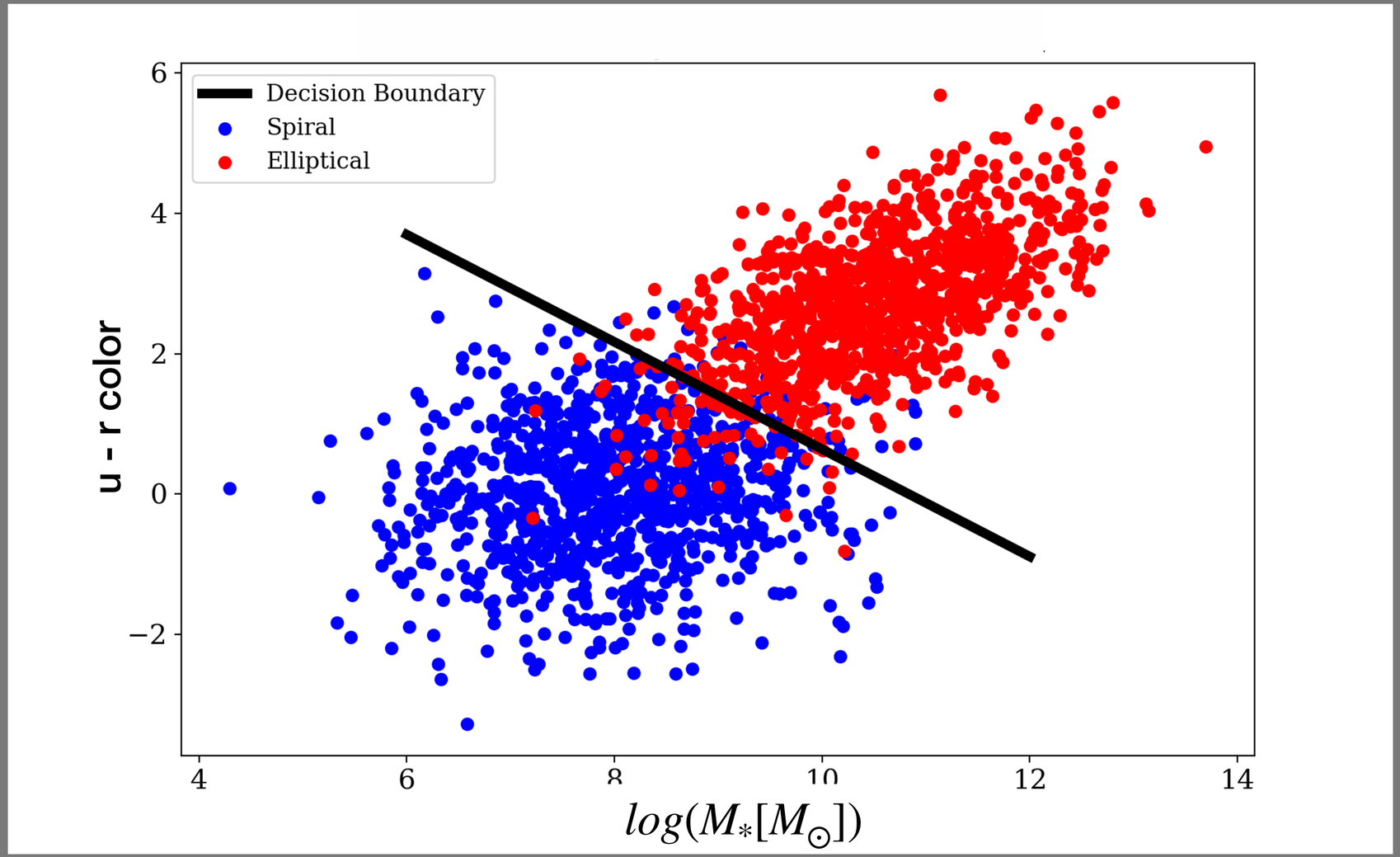
$z = X\theta^T$

s = sigmoid(z)

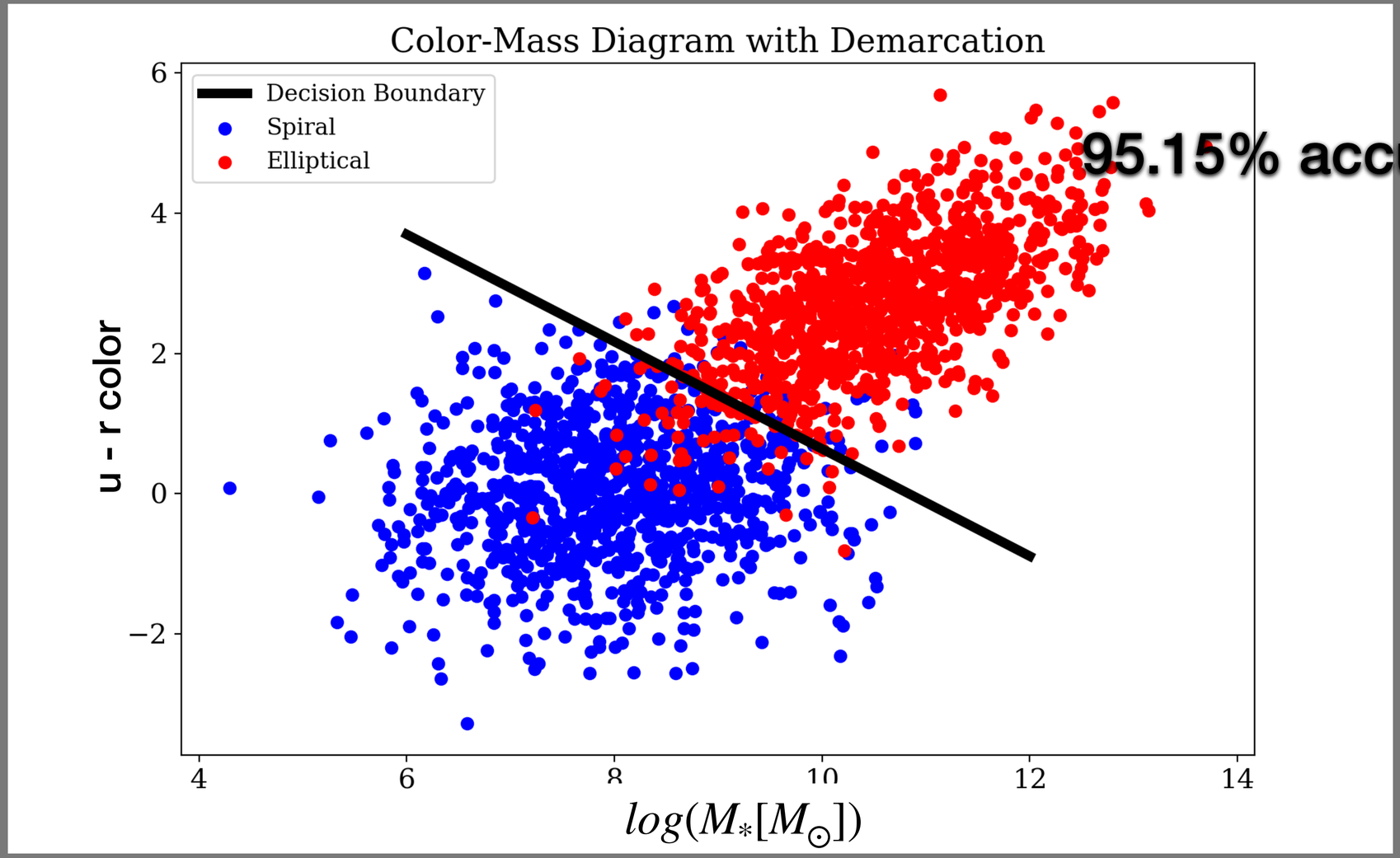
```
for i in range(len(X)):
if s[i] >= 0.5:
    prediction = 1

elif s[i] < 0:
    prediction = 0
```



Results: Plot

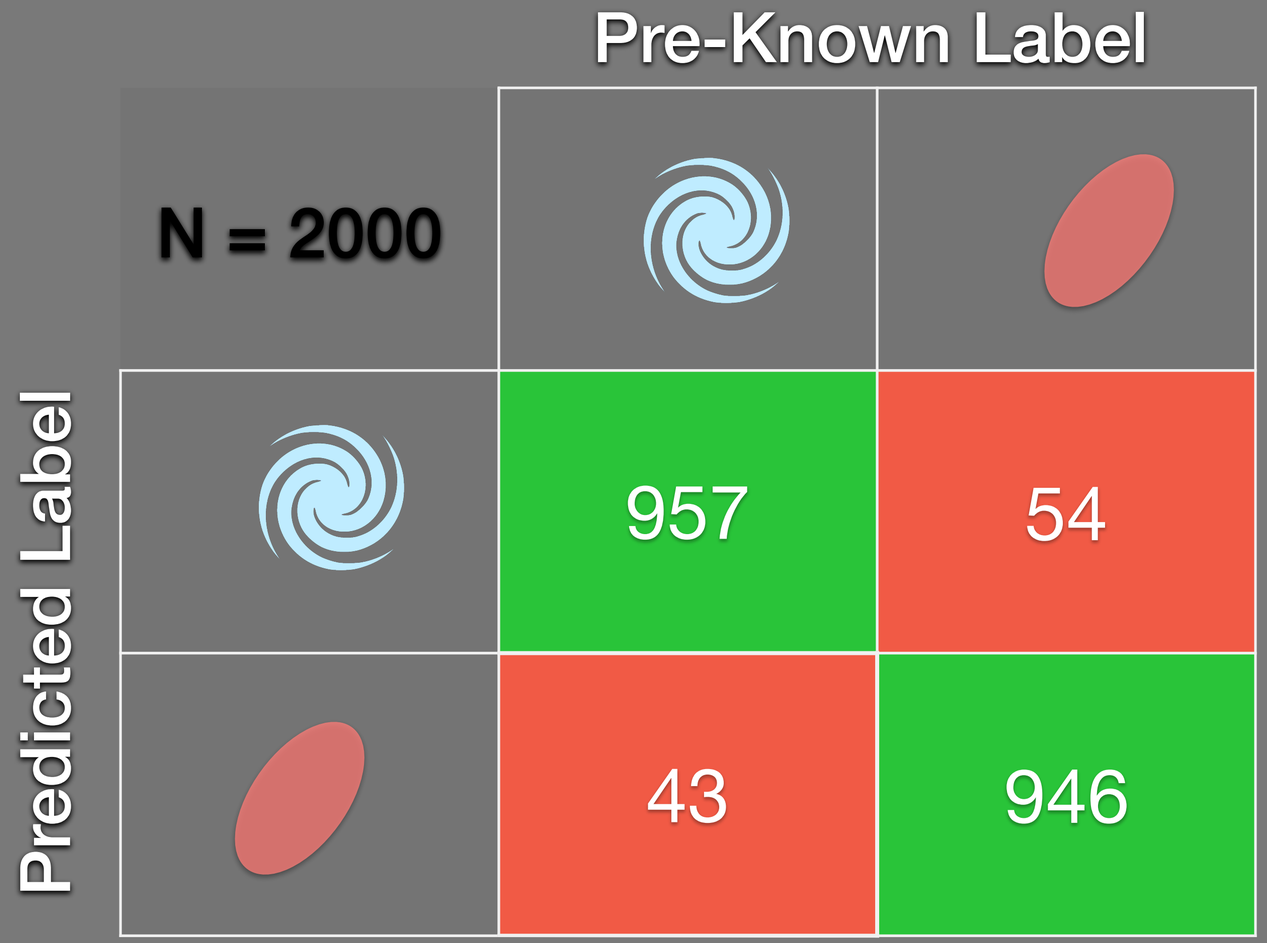


Results: Plot



95.15% accuracy

Results: Confusion Matrix



Future Direction

- Extend to multiple feature (from two features —> N features)
  - Find an equation that predicts label based on many values of many features

Future Direction

- Extend to multiple feature (from two features —> N features)
  - Find an equation that predicts label based on many values of many features

- Use actual observational data from SDSS and Galaxy Zoo

Questions?

References / Sources Consulted

Chandrasekaran, Dinesh. ”Logistic Regression from Scratch Using Python.” (2019).

Jurafsky, Daniel. ”Logistic Regression.” (2019).

Schawinski, Kevin et al. ”The green valley is a red herring: Galaxy Zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies”. Monthly Notices of the Royal Astronomical Society 440. 1(2014): 889–907.

Ungar, Lyle. ”Logistic Regression.” (2020).

Function: Cost Function