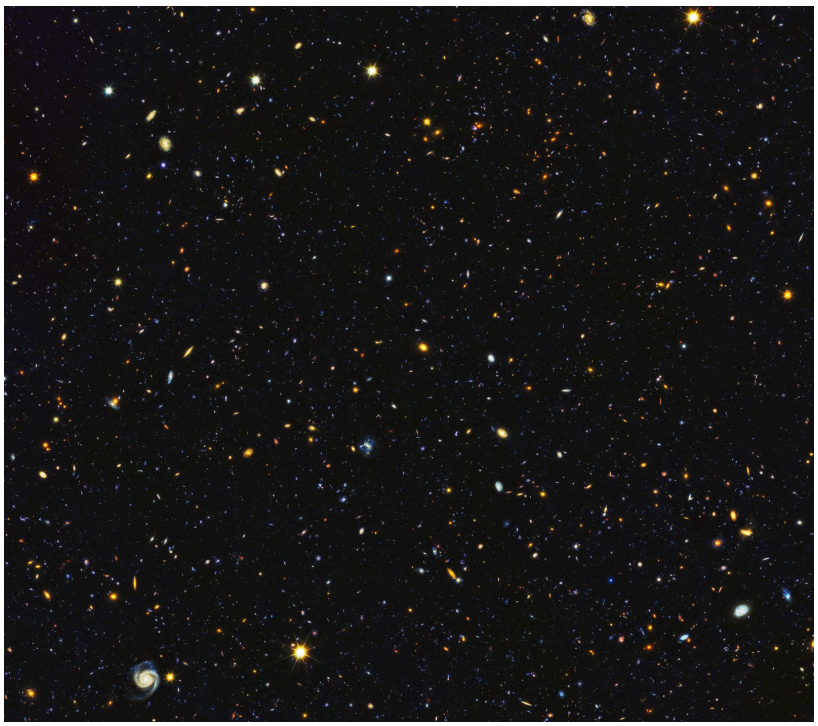
Naming the Stars

Using Convolutional Neural Networks to Classify Galaxies



The number of galaxies in the observable universe is currently estimated to be more than 2 trillion 1. The number of galaxies in the image above alone is estimated to be over 15,000.

Evaluating and classifying the staggering array of galaxies continuously being discovered by telescopes is a daunting task which, if performed by human observation and labelling, could never be completed.

In 2007, a data set was published by the Sloan Digital Sky Survey consisting of over a million images of galaxies 2. These were eventually added to other data sets of galaxy images by a team of scientists known as Galaxy Zoo, which facilitated the classification of these galaxies. Much of these data was eventually released as a part of a Kaggle competition in 2013 3, which provided galaxy images along with the probability of classifications applying to each galaxy.

Since 2013, options for computation classification of galaxy images has greatly expanded, with significant developments in the area of image classification using

convolutional neural networks. It is my aim in this project to use the crowd-sourced data obtained by Galaxy Zoo to perform the following:

1) Distill the crowd-sourced data into a 4-class version of Hubble's galaxy classification system (elliptical, spiral, irregular, and star/artifact).
2) Using images labelled according to these classifications, use Keras to train a convolutional neural network to classify galaxies accordingly.
3) Evaluate the accuracy of the neural network using a validation set of galaxies.

This can then be refined and expanded on to create a more thorough and accurate classification system for galaxies - one that can feasibly keep up with, and perhaps exceed, the rate at which new galaxies are discovered.

I intended to deliver the code which accomplished the steps outlined above, as well as a blog post explaining the process.

Fortunately, the data itself is uniform, has no missing values, and requires no wrangling apart from image processing, which will be incorporated into the modeling pipeline. Because the labelling of the images was crowdsourced data, however, it is worth keeping in mind that the data associated with galaxy class assignments for each image indicates what percentage of data labellers assigned that class to the image. There is a possibility that some images are difficult to classify and have a low degree of agreement between labellers.

Those labelling the data were given the tasks laid out in the image below, with each task potentially leading to other tasks, as indicated in the "Next" column. Each class above corresponds to task and an answer below. In other words an answer of "No" on task 3 would be assigned Class 3.2. The results in a percentage associated with each class for each galaxy image.

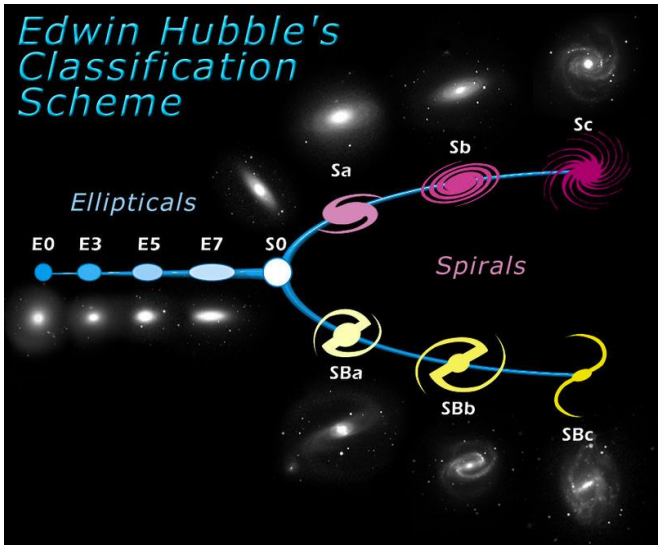
Task	Question	Responses	Next
01	*Is the galaxy simply smooth and rounded, with no sign of a disk?*	smooth	07
		features or disk	02
		star or artifact	**end**
02	*Could this be a disk viewed edge-on?*	yes	09
		no	03
03	*Is there a sign of a bar feature through the centre of the galaxy?*	yes	04
		no	04
04	*Is there any sign of a spiral arm pattern?*	yes	10
		no	05
05	*How prominent is the central bulge, compared with the rest of the galaxy?*	no bulge	06
		just noticeable	06
		obvious	06
		dominant	06
06	*Is there anything odd?*	yes	08
		no	**end**
07	*How rounded is it?*	completely round	06
		in between	06
		cigar-shaped	06
08	*Is the odd feature a ring, or is the galaxy disturbed or irregular?*	ring	**end**
		lens or arc	**end**
		disturbed	**end**
		irregular	**end**
		other	**end**
		merger	**end**
		dust lane	**end**
09	*Does the galaxy have a bulge at its centre? If so, what shape?*	rounded	06
		boxy	06
		no bulge	06
10	*How tightly wound do the spiral arms appear?*	tight	11
		medium	11
		loose	11
11	*How many spiral arms are there?*	1	05
		2	05
		3	05
		4	05
		more than four	05
		can't tell	05

Table 2. The GZ2 decision tree, comprising 11 tasks and 37 responses. The 'Task' number is an abbreviation only and does *not* necessarily represent the order of the task within the decision tree. The text in 'Question' and 'Responses' are displayed to volunteers during classification, along with the icons in Figure 1. 'Next' gives the subsequent task for the chosen response.

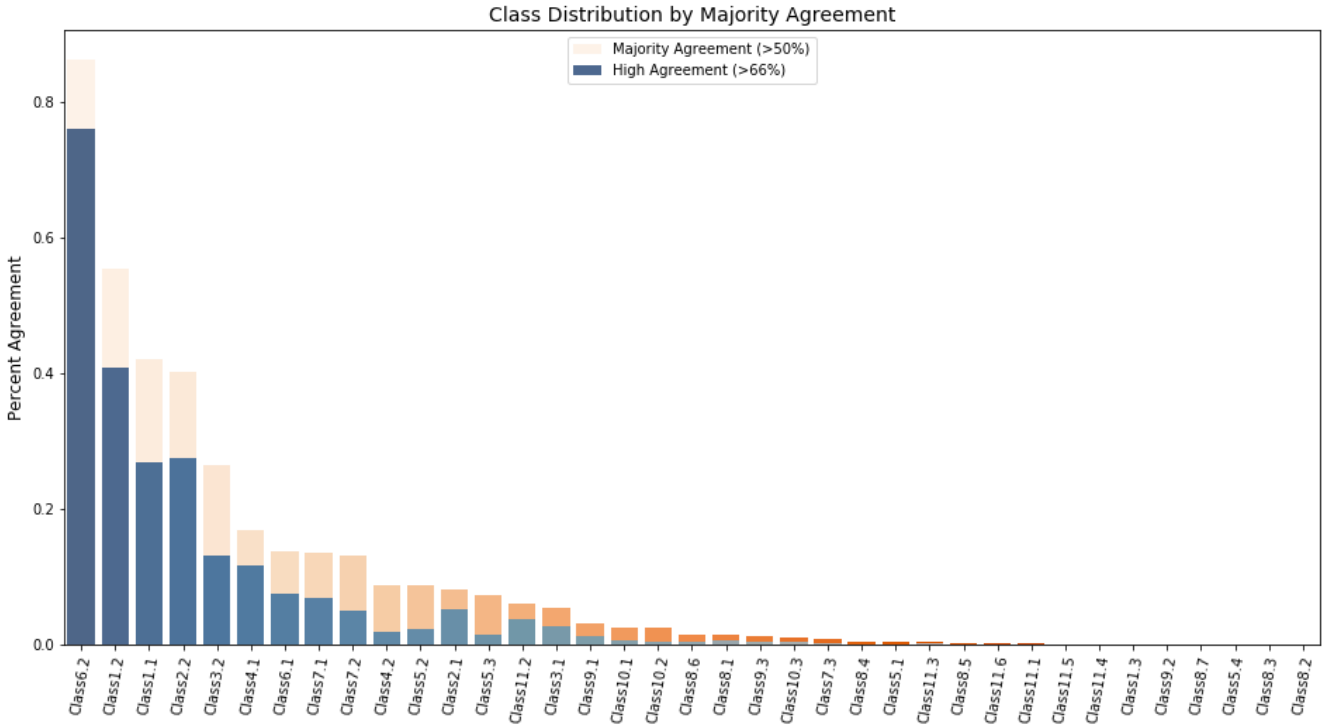
The above chart can, to at least some degree, be distilled into Hubble's galaxy classification system. As can be seen below, this divides galaxies into ellipticals, spirals, and spiral bar classes.

- Ellipticals would be indicated by class 1.1 in the above, though this may not be true if it is accompanied by a high classification percentage on class 6.2.
- Spirals would be indicated by classes 4.1 and 3.2 together (again, with 6.2 possibly affecting it).
- Spiral bar galaxies would be indicated by classes 4.1 and 3.1 together (again, depending on 6.2)

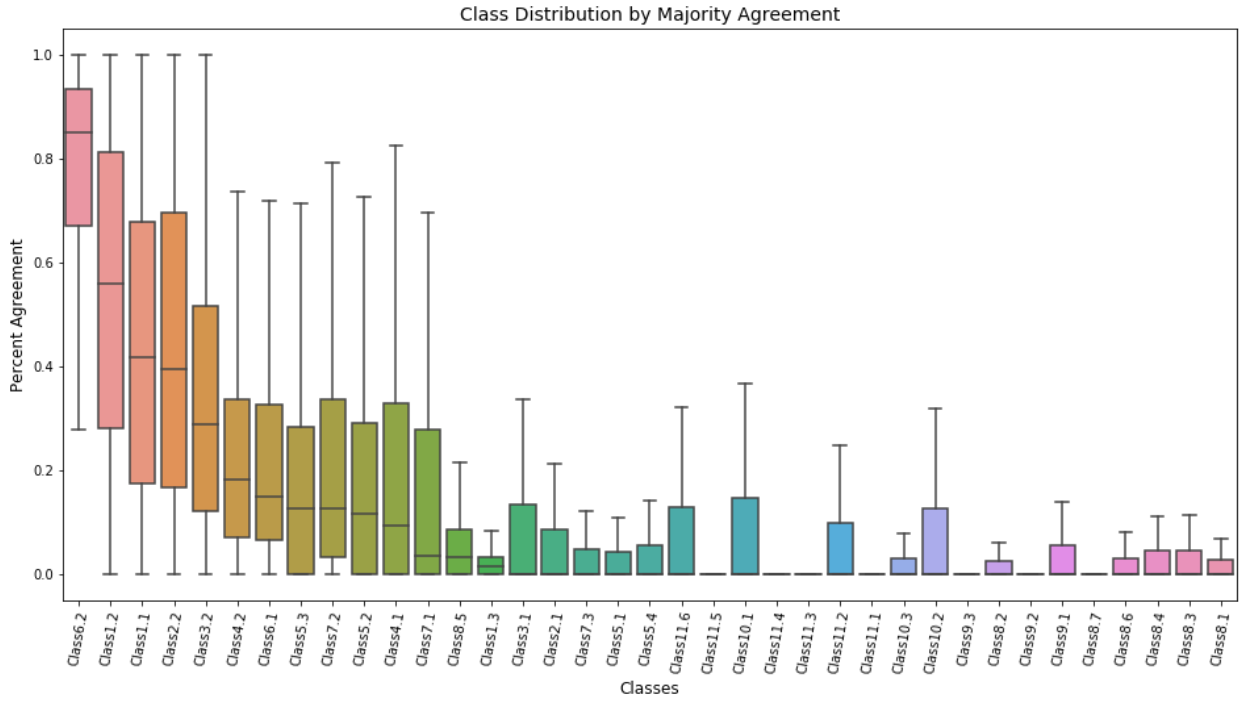
Subclasses of each of these would be indicated in classes 7 (degrees of elliptical), or 10 and 11 (types of spiral and spiral bar galaxies).



Keeping the above in mind, let's take a look at which categories tend to be assigned more frequently than others with a high degree of agreement among data labellers.



We can get additional information about the above by looking at a box plot to see the median values and quartile ranges of agreement for each class distribution (note that flyers were disabled from this class due to a particularly high number of them).

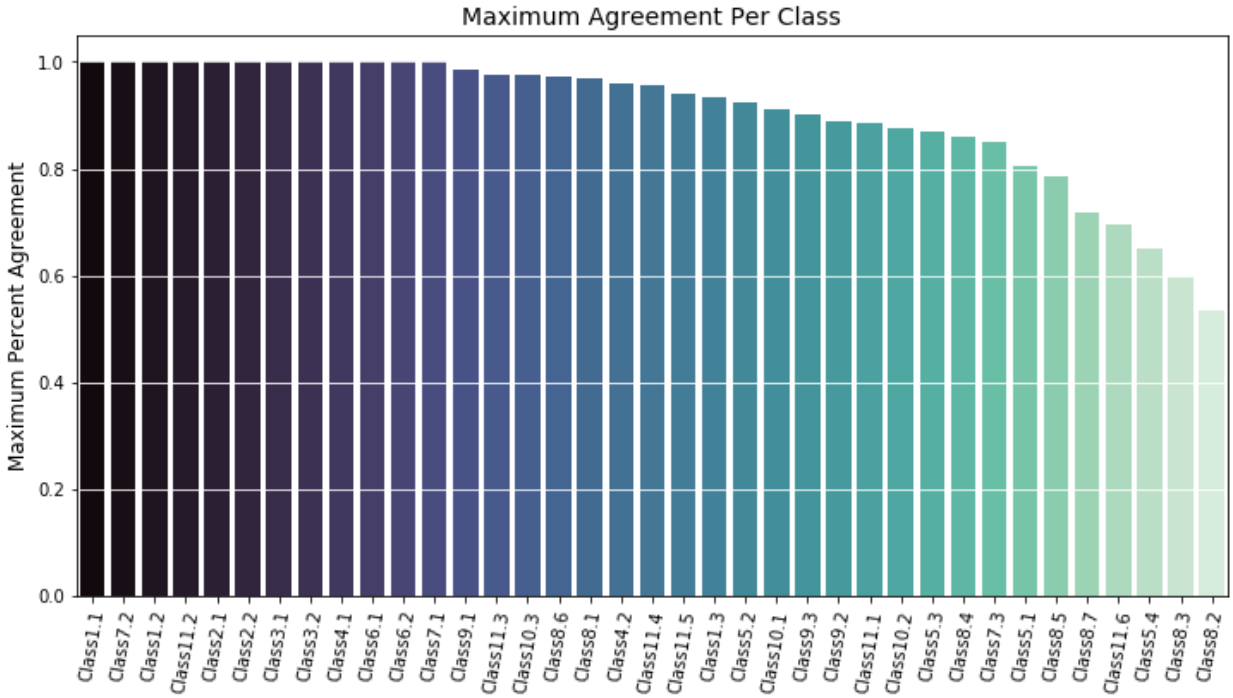


Important information that can be gleaned from these graphs, from left to right:

1. There is nothing "odd" about the majority of labelled galaxy images. (Class 6.2 compared to 6.1)
2. More than half of galaxy images include some kind of feature or disk, but a significant range is observed in every quartile (Class 1.1)
3. Around 40% of galaxies could *not* be disks viewed edge-on, though significant range is observed in every quartile. (Class 2.2 compared to 2.1)
4. Where galaxies are not disks viewed edge-on, most tend not have have a bar (Class 3.2 compared to 3.1)

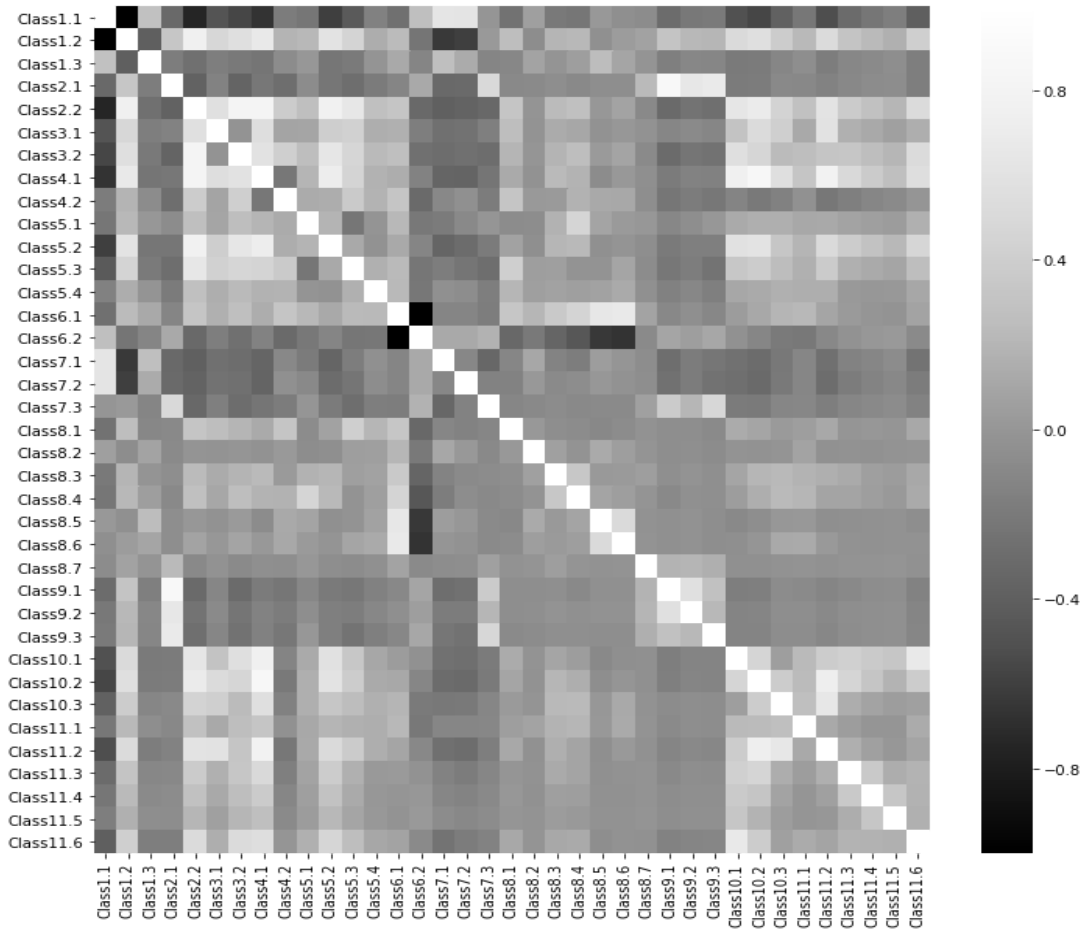
Overall, the median agreement for each class tends to be rather small, but it appears there are galaxies in most classes for which a significant number of labellers were in agreement.

Since the median agreement for many classes is relatively small, it may be helpful to evaluate the maximum percentage agreement for each class to see how many classes may have a low maximum. A large number of classes with a low maximum may indicate that agreement among class labellers was so low as to prohibit successful training of model.



Fortunately, it appears that most classes have at least some galaxies for which agreement was over 80%. Only 6 classes have a maximum agreement percentage lower than this, and 4 of these relate to the classification of “odd features”, which are not very common. The other 2 classes related to the presence of a “dominant bulge” (class 5.4) and an indeterminate number of spiral arms (class 11.6).

Because some tasks in the data labelling process lead to others, we would expect to see a strong measure of correlation between specific classes. This should be observable in a correlation matrix.



Looking closely at the above, we find about what we would expect. The starting and ending tasks (classes 1 and 6) with a “yes” or “no” distinction are clearly negatively correlated. The other “yes” and “no” classes vary since they are dependent on other tasks. Other classes can be evaluated one-by-one to see that they make since. For instance, since the first answer on task one would send a crowd-sourced data labeller on to task 7, Class 1.1 correlates highly with classes 7.1 and 7.2. As could be seen in the description of the dataframe, class 7.3 was assigned rather uncommonly (with a mean value of 0.05), which explains the lesser degree of correlation there. Stepping through each class similarly, we can observe the patterns defined by the data-labelling task list.

To finish up the exploring this data, it would be interesting to see if galaxies that are classified with a high degree of agreement visually correspond with their Hubble classification. Looking for elliptical, spiral, and spiral-bar galaxies given the criteria outlined above with over a 90% agreement yielded 1,673 elliptical galaxies, 234 spiral galaxies, and 77 spiral-bar galaxies. Choosing the first result for each of these yields the following images, in the order they were just listed: