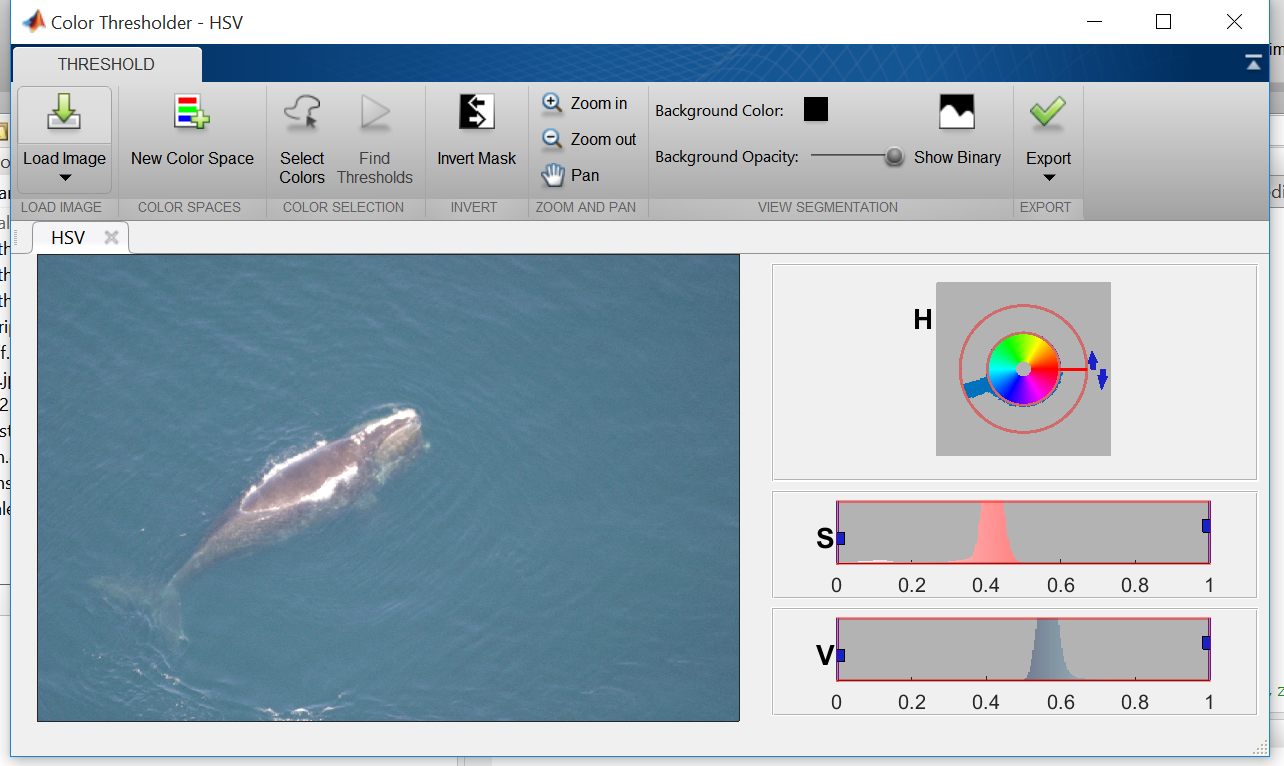
# Part 1: Define the objective and scope of the project. Gather and organize the data for the project.

## Conduct exploratory data analysis such as visualizing the data through graphs, tables, summary statistics, and other means to understand the data.

We have 11468 JPG images of right whales and a CSV file that associate 447 whales to 4544 images as a training set. All other tagging data has been stripped from the images. They are all aerial photos taken in all kinds of different lighting and ocean condition. For most of them they are shots of the whale’s backs with full body or only the head visible. The whales can also be submerged which distorts the image and spray from movement through the water and from the blowhole can obscure the image. Looking at the images in hsv we get the following typical color distribution using mathlab.

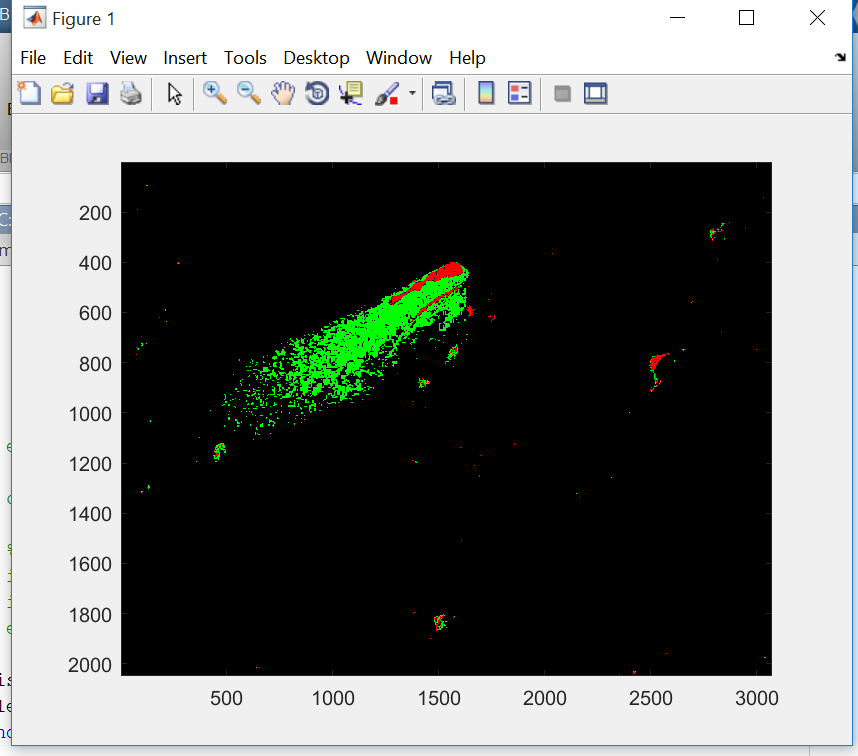
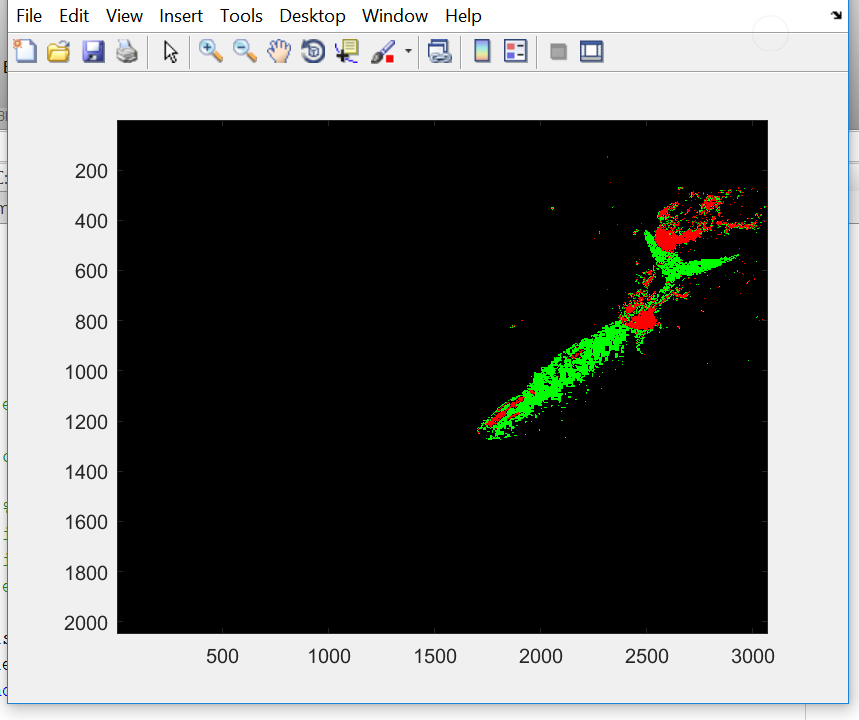


Most of the saturated color that you see in the h wheel is mostly in the blue green range and that is ocean background color and that constitutes most of the distribution that you see in the S scale between 0.3 to 0.5. The whale’s colors tend toward the unsaturated or gray scale so it is typically less than 0.3 on the s scale. Most of the color intensity v scale sits in a distribution range from 0.2 to 0.6 that is no more than 0.1 wide. The splash and the head orientation sit in a range above 0.8 on the v scale and below 0.2 on the s scale. The intensity on the head ornament, which is an object of interest can vary depending on light intensity and submersion of the head. Where the distributions are on the hsv scales vary but the distribution pattern remains consistent from photo to photo with few exception.

## Identify any issues associated with data gap, data size, data type, data manipulation, data storage and data retrieval for analysis. Structured or unstructured data?

Even though there is not any meta data besides the photos I consider this data to be very mess from the prospective of identification: All the 8Gb of photos very in size and resolution, there is a high amount of ocean to whale, spray and submersion obscure and distort the photos. Since the head ornament which will be very useful in the identification varies in intensity it is hard to separate it from the spray. Per-process will be an ongoing effort as we work out the features we want isolate for the identification. Data , reduction and normalization of the data is of high impotence as we determine the types of features we want to extract.

For example, from the color distributions I can get distributions in the photos as follows that can help us isolate the whale and the whale’s head. Where green shows the likelihood of whale and the red show the spray and head ornament.



## Describe the high level analytic problem needs to be resolved: supervised learning, unsupervised learning.

Ours is a supervised learning problem of whale recognition. We have 11468 JPG images with 4544 with target information on 447 whales. We are looking into object and facial recognition categorization models.