

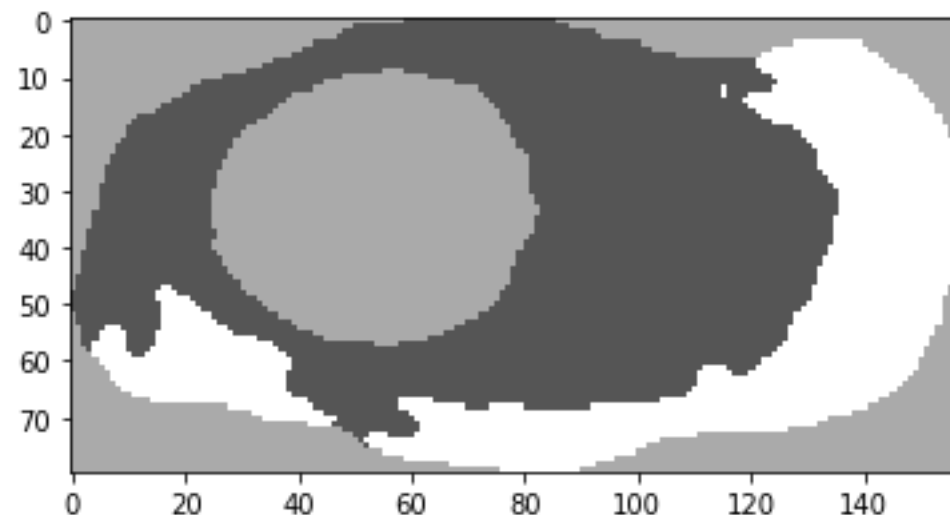
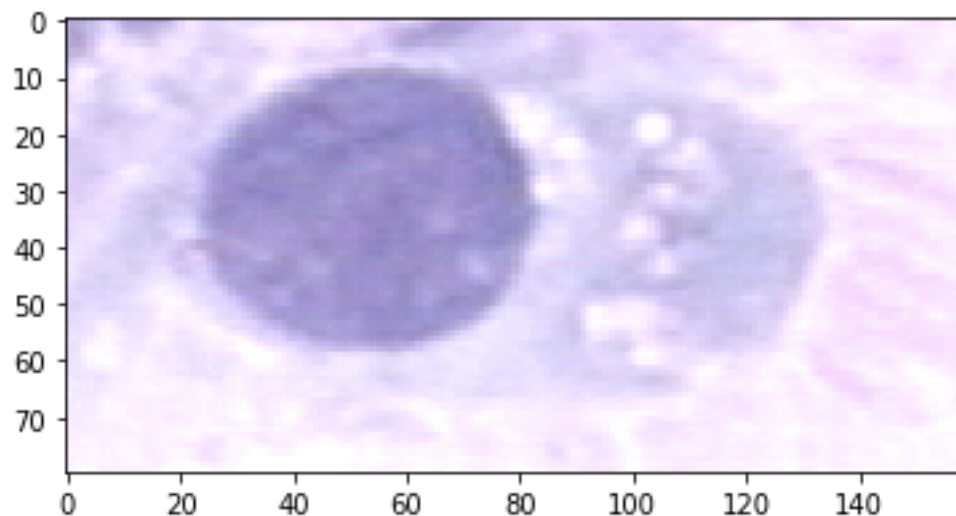
Image Analysis of Pap-Smear dataset.

George Batchkala

EDA

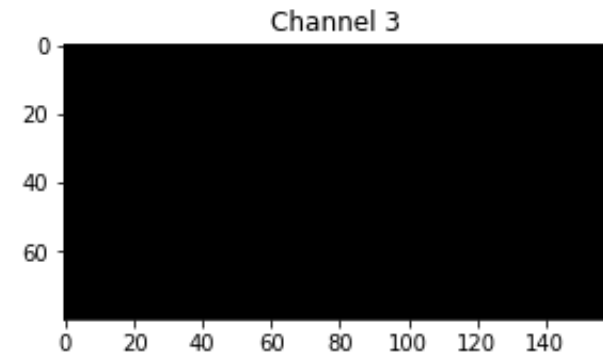
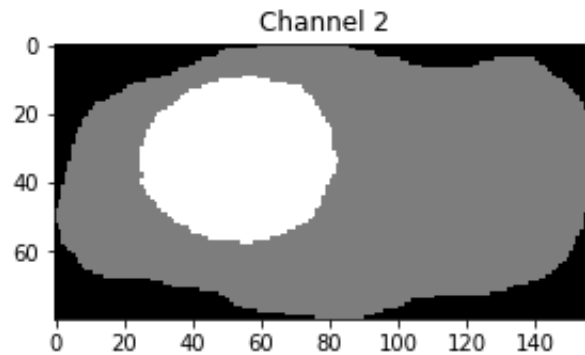
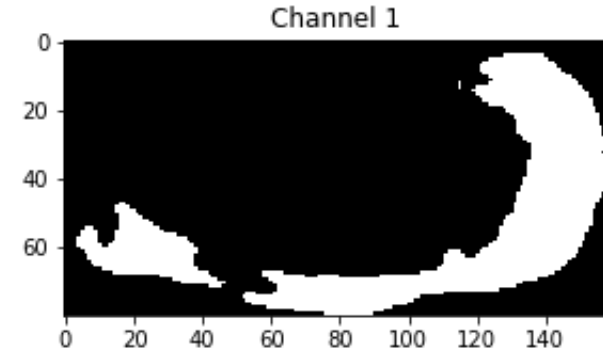
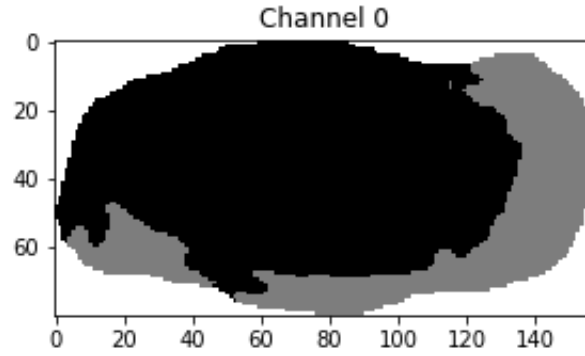
- The data is split into 7 classes
- Images are given in 3 channels
- Masks are given in 4 channels

Example of an image from `normal_columnar` set (left) with the corresponding mask (right).



Choosing How to Work with Data: 1

- Masks contain 4 channels and segment many images in 4 regions while only 3 regions representing the nucleus, cytoplasm, and the background are desired.

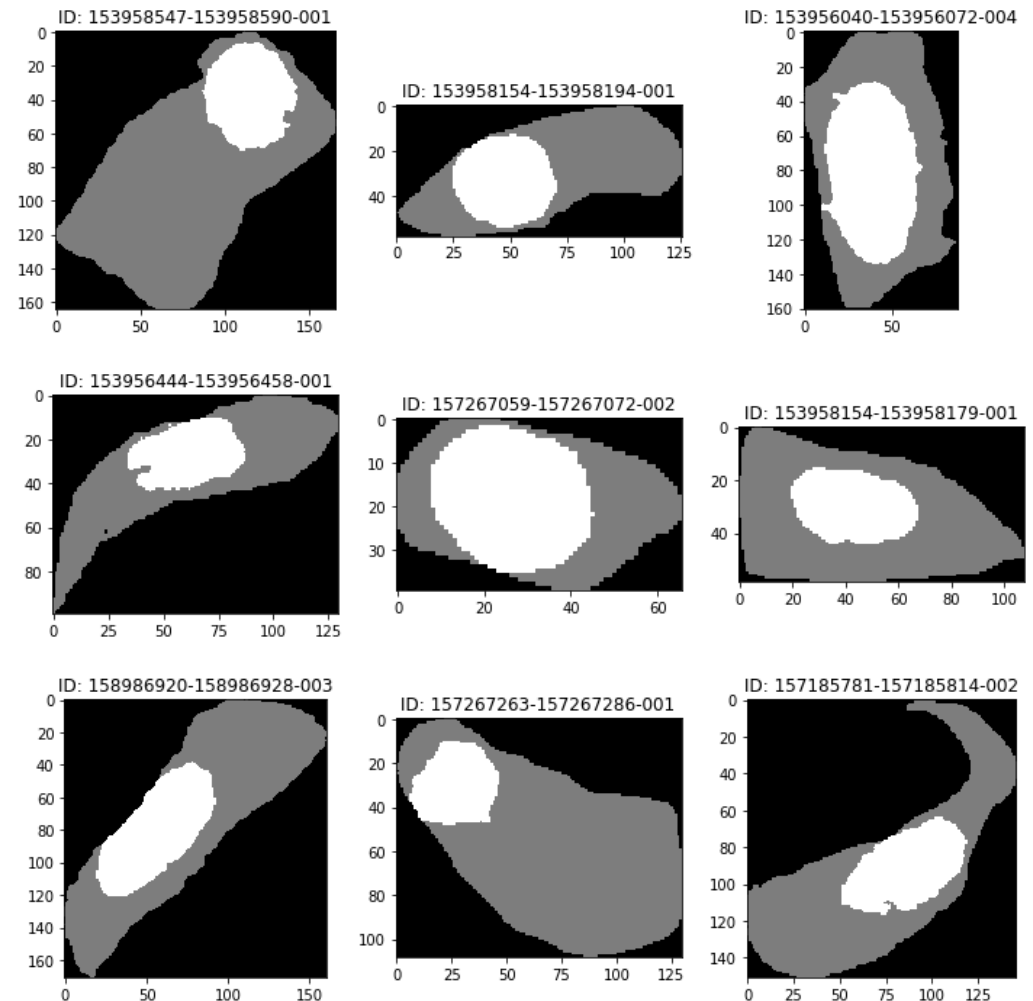


Choosing How to Work with Data: 2

Inspection of all masks in
`normal_columnar` set
shows that channel-2 has
exactly 3 distinct values of
intensity:

- Background: 0
- Cytoplasm: 128
- Nucleus: 255

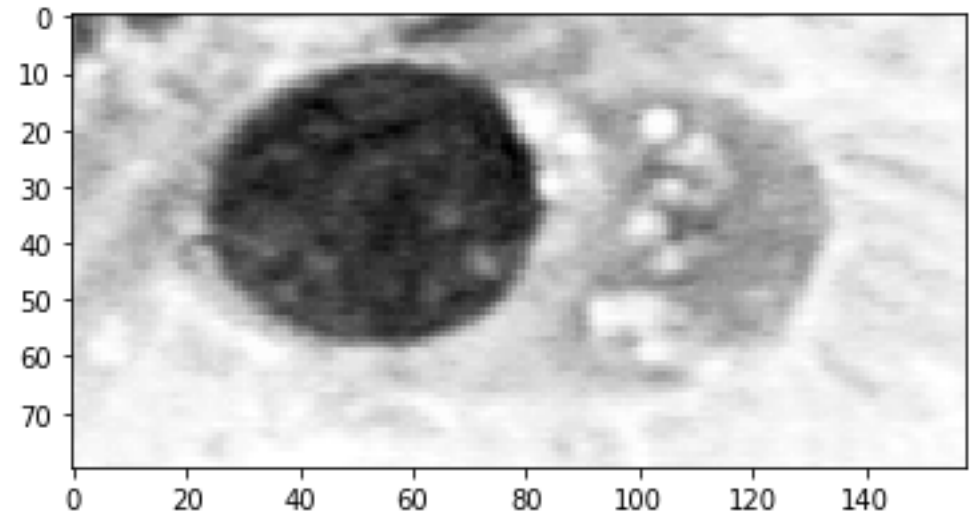
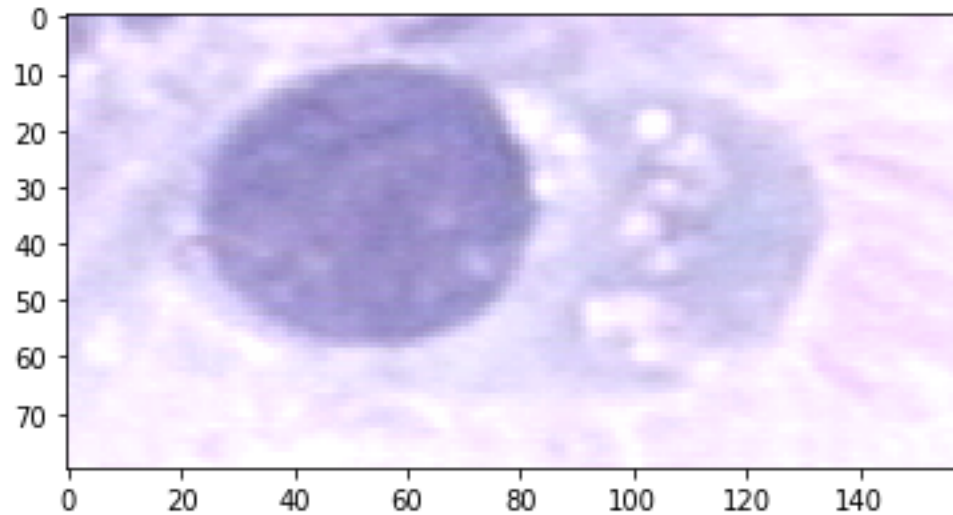
9 randomly selected channel-2 masks across all sets.



Thresholding

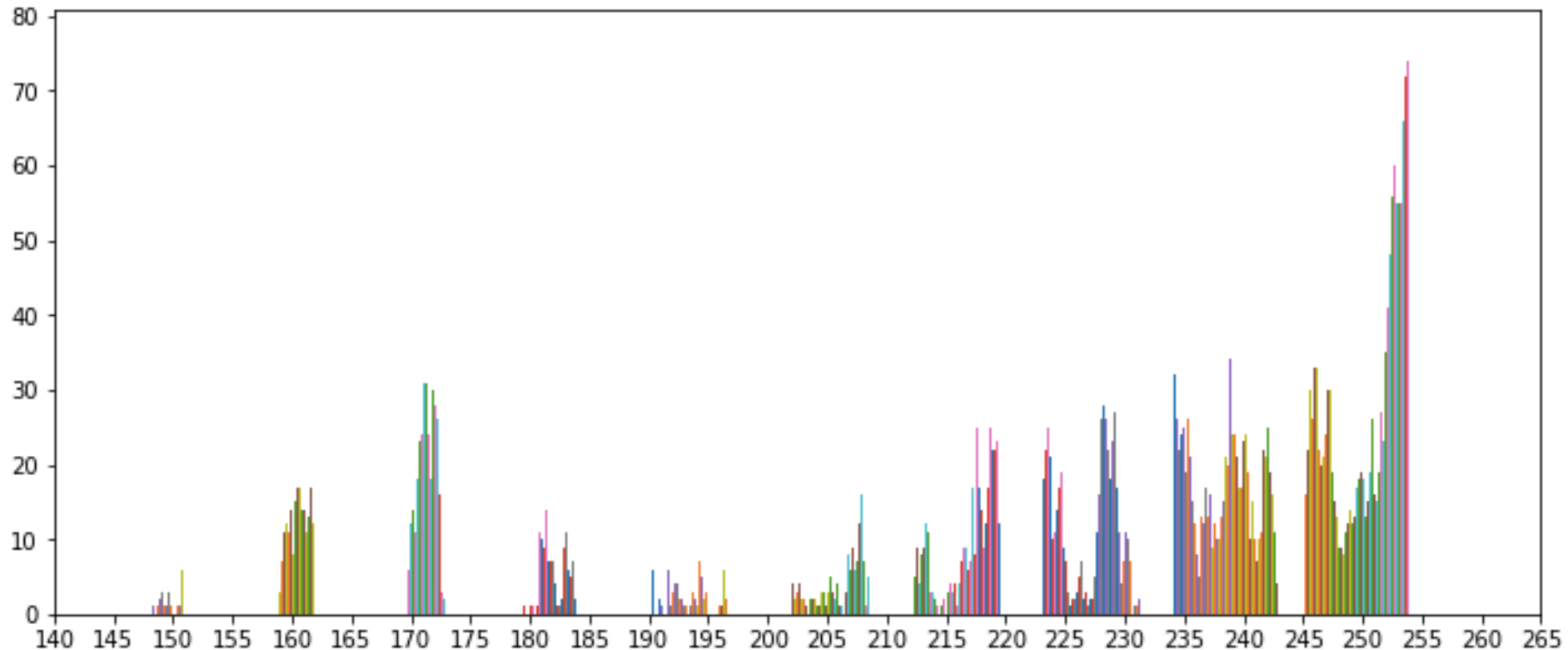
- To perform the segmentation using thresholding original images were first converted into one channel by averaging the intensities.

Example of an image from `normal_columnar` set (left) and the corresponding image averaged across channels (right).



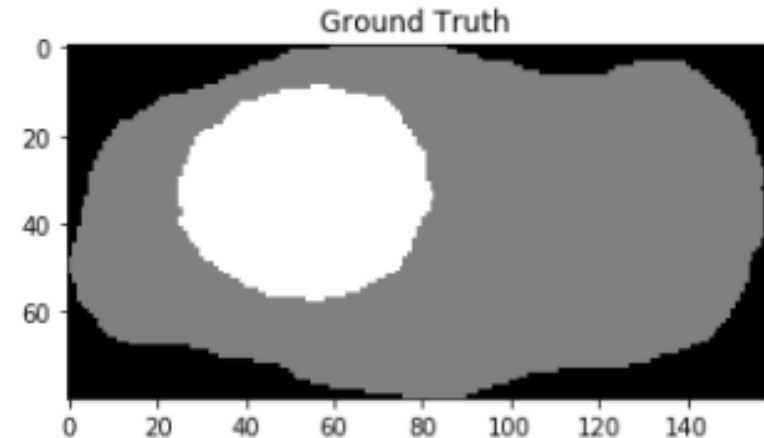
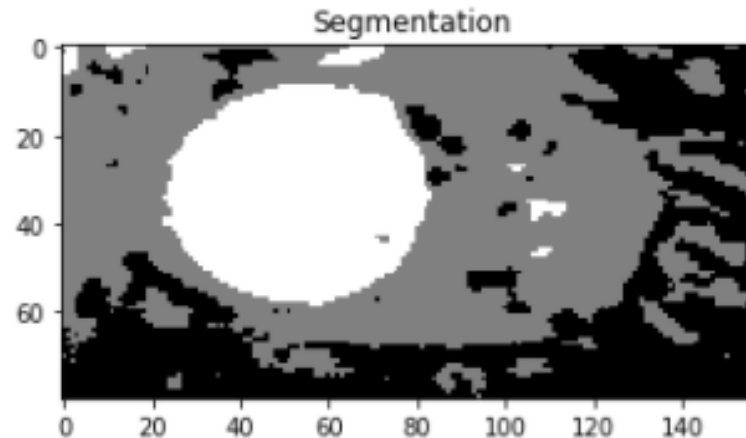
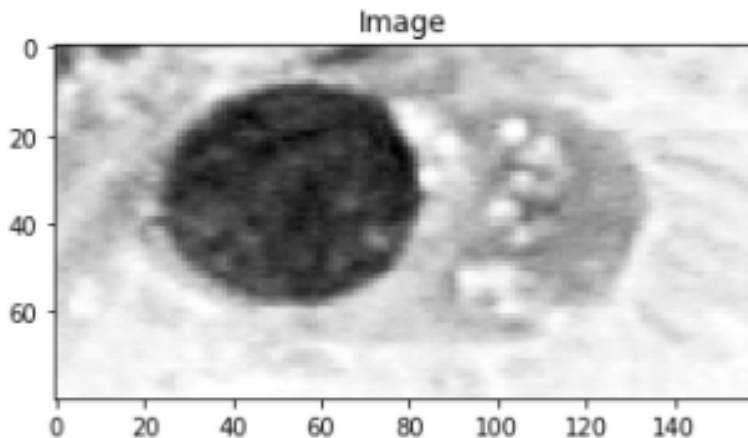
Manual Thresholding: 1

- First an attempt of thresholding the image manually using the histogram of intensities was performed.



Manual Thresholding: 2

- I decided to cut off at intensities of 200 and 245. Result below:



- Dice Scores for Nucleus, Cytoplasm, and Background respectively :
 - This Particular Image: 0.94, 0.74, 0.49
 - normal_columnar all: 0.40, 0.17, 0.09

As expected, I overfitted when choosing thresholds by looking at the histogram

Otsu Thresholding into Multiple Classes

- Next, I decided to do exactly the same, but now use an automatically selected thresholds (otsu) for each image. Note, that 2-threshold method was used: https://scikit-image.org/docs/dev/auto_examples/segmentation/plot_multiotsu.html
- Dice Scores for Nucleus, Cytoplasm, and Background respectively :
 - First image: 0.95, 0.51, 0.45
 - All images: 0.72, 0.58, 0.47

This resulted in an improvement of the mean dice-score across all images compared to using manually selected thresholds at a computational cost of computing thresholds for all images separately.

Otsu Thresholding, Median Filter

- Now I decided to use the median filter for denoising original images before applying the otsu-thresholding method. However, the size of the filter had to be chosen – it was overfitted to the dataset.
- The filter radius was run from 0 (no change) to 15 included.
- Both for the 1-image (with disk of radius 9) and all-images cases (with disc of radius 4) there was an improvement in the dice scores. See Table of the results for exact numbers.

Otsu Thresholding, Other Filters

- Finally, I decided to explore the 4 filters shown in the lectures for vessel detection:
 - Sato;
 - Meijering;
 - Frangi; and
 - Hessian

None of them improved upon the results of otsu-thresholding.

Results: Binary Dice Coefficient

One image from `normal_columnar` set

All images from `normal_columnar` set

Method	Nucleus: 1	Cytoplasm: 1	Background: 1	Nucleus: all	Cytoplasm: all	Background: all
Manual	0.94	0.74	0.49	0.40	0.17	0.09
Otsu	0.95	0.51	0.45	0.72	0.58	0.47
Otsu + median	0.97	0.62	0.45	0.73	0.59	0.48
Otsu + sato	0.00	0.46	0.06	0.05	0.37	0.09
Otsu + meijering	0.00	0.27	0.00	0.08	0.27	0.04
Otsu + frangi	0.06	0.04	0.00	0.13	0.15	0.02
Otsu + hessian	0.41	0.28	0.4	0.37	0.28	0.42

Median filtering with disk radius 4 followed by **otsu-thresholding** with 2 thresholds produced best results for the normal_columnar subset of the pap-smear dataset.

Possible Improvements

- K-Means will be very interesting to explore, however the automatic procedure to map clusters to the corresponding labels (nucleus, cytoplasm, background) must be established before trying to assess the performance
- Using SLIC Superpixels is another pre-processing method to consider
- Segmentation via Deep-Learning U-Net-type models will be interesting to perform