Data organization:

In order to try and limit the number of for loops, I propose we pre-process the images in the following way:

1. Place them all in the same folder when we are ready to start using the files.
2. Naming them “covid_#” or “nonCovid_#”

Find Ideal Parameters:

1. Covid to Non-Covid pictures ratio:

According to various websites and articles, having a much higher number of non-covid pictures like we do could bias the model towards giving non-covid answers. To remedy this, we must decide on a ratio we will apply to all test sets (i.e. ½ pictures will be covid pics, or 3/10 will be covid pics etc...) To do this we need to create a model, and see how its cross-entropy changes as we change the ratio.

2. Number of pixels:

Having a very high number of features (70K+ columns) versus instances (rows in the hundreds) is yet another source of error for our model. Thus we need to further resize the pictures to have the smallest number of pixels possible while retaining the highest variance possible. To do this, we need to create a graph which shows the variance of the dataset in function of the number of pixels, and select a number à la Elbow Curve. Once the model is created, we will create a second graph which shows the cross-entropy in function of the number of pixels.

Train Model

Once we start building the model, I think we should pick our covid and non-covid images at random.

Validate Model

Test Model

In the testing phase, I think the program should write a file which contains track of this iterations parameters(ratio, pixel#) and the metrics (variance, cross-entropy). This way we can keep track of which set of parameters is the best.

Code Optimization:

Naturally our code is going to get messy, despite our best efforts. This is why we will need to dedicate some time to commenting, documenting, optimization and vectorization, to ensure that our code is as simple and clean as possible.