APA references  
Summarize your business problem/hypothesis  
Walk through your method  
5-8 Pages, 4000-8000 words  
Up to 3 Illustrations @ ¼ page each  
Up to 1 Illustration @ ½ page  
PDF  
1 Appendix  
Submit your 10 questions

# Problem

One of the best ways to fight Covid-19 is to simply wear a mask when out in public. This is true because it is a small step that everybody can do. In addition to masking up in the general public, (which ought to be happening anyways), there are locations where masks are required, and also must be work correctly. Common examples include hospitals, but could also include standard retail stores. As of right now, it is responsibility of employees to detect if a person is wearing a mask, and if they are wearing it correctly.

Luckily, a decent amount of the detection process can, in theory, be handed off to a well-trained machine learning model. Many businesses already have cameras monitoring the interior of said business, so it would be useful if they could be used to help boost public safety, and monitor mask wearing.

# Hypothesis

I hypothesize that an image classification model can be trained to detect not only the difference between a masked individual and an unmasked individual, but it can also predict if a person is wearing a mask incorrectly or not.

# Dataset

The dataset that will be used for training and testing a model can be found at the following URL:

https://www.kaggle.com/andrewmvd/face-mask-detection

This dataset contains 853 images, each of which contains images of at least one person’s face. These faces fall into one of three categories:

* mask\_weared\_incorrect
* with\_mask
* without\_mask

Each image has an associated annotation file, which is in .xml format. An example is shown here:

There is a bit of extra information in this dataset that is of little use to us, such as the “segmented” field, or the “pose”, “truncated”, “occluded”, or “difficult” fields. The data that will be useful is:

* filename: which image does this annotation correspond to
* size: dimensions of the source image
* name: classification of the person in the image
* bandbox: tells us what portion of the image contains the face in question.

Given this information, and the images for analysis, the next stage in the data pipeline is to prepare the data for analysis.

# Data Preparation

In order to load the annotation data, it was decided that a convenient way to store the data would be a simple .csv file. There are plenty of ways that one may accomplish this task, but for this project, a Powershell script was used. It loops through each annotation file, and within each file, it then loops through each