**CSCI 597J**
**Project Pre-Proposal**
Carter, Drazic, Chuprinov, Gavin, Hansen

1. What problem are you trying to solve? Why should someone care if you solve it?

   Given an x-ray of someone's lungs to determine whether or not that person has pneumonia either viral or bacterial, or not at all. Detecting what kind of pneumonia a patient has would be immensely helpful in determining what course of action to take for the nurse or doctor.

2. (a) What data could you use to work on this problem?
   A dataset containing 5,800+ X-ray chest images.

   (b) Where will you get it from?
   This data set will be supplied by Kaggle.
   https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

   (c) Are there restrictions or other obstacles to obtaining or using it?
   None exist. It is available on Kaggle with a CC BY 4.0 license. No personal information is included that would violate a patient's privacy.

   (d) How big is it (in as specific of a sense as possible)?
   5800+ X-Ray images (JPEG) and 2 categories (Pneumonia/Normal)

   (e) What kind of labels, if any, are available. Give an honest assessment of any known strengths or flaws in the data set(s) you identify.
   Labels available: Viral Pneumonia, Bacterial Pneumonia, and No Pneumonia
   Strengths:
   The images were screened for quality control and evaluated by experts
   The images are relatively high resolution/zoom
   The data is anonymized
   Weaknesses:
   Only concerned with pneumonia as opposed to all lung conditions
   A relatively small dataset
   The images have different resolutions
   Descriptions of files are not always included
   Origin of image and update history of the dataset are also not provided

3. What other variants of the problem could you tackle? How do these variants compare to your main problem in terms of effort, impact, etc.?

   Additional variants or extensions of the problem include ranking the severity of a patient's case of pneumonia, determining when the pneumonia first appeared based on the properties of the case, or even identifying the regions within the lungs/xray of where the pneumonia is contained.

   These additional variants would be significantly more difficult as we would have to learn more about the different types and properties of pneumonias in order to better approach the task. Our current problem is much simpler thanks to the data being clearly labeled for non-domain experts like us. However, variants like the pneumonia region identification would require new, manually labeled, images with regions outlined by experts where the pneumonia is contained. This is a much larger task to tackle as opposed to classification.

4. Who is on your project team so far, would you like more members and would you like me to help find those new members?

   Chris Drazic, Ivan Chuprinov, Raleigh Hansen, Alex Gavin, Matthew Carter