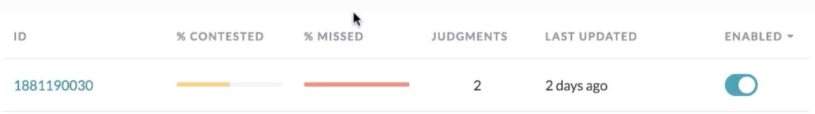
Project Proposal

Bakary Badjie

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	The aim of this project is to assist pulmonologist (i.e. Doctors who specialize in diagnosis and treatment of respiratory) in rapidly identifying pneumonia cases in infants. The work guide is designed to assist pulmonologist in rapidly identifying healthy patients and alerting them to the possibility of pneumonia. Machine Learning algorithms performs well at learning from huge amounts of unstructured data and image recognition, making them credible for the role of categorizing healthy vs unhealthy images.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	There several labels in the given dataset, however, we have added three extra labels on our data which are as follows: ➢ Healthy ➢ Unhealthy ➢ Unknown There are three possible results from our chest x-ray image classification: which are; 1. A patient has pneumonia which is labeled as "unhealthy", 2. A patient does not have pneumonia which is labeled as "healthy", and 3. We have no idea if a patient has pneumonia or not, which is labeled as a "unknown". However, the degree of our confidence that a patient has pneumonia or not is range from low to high, which can be translated as follows: low indicates that our confidence level is low in making our decision, while high indicates that our confidence level is very high on our decision.

Test Questions & Quality Assurance

Number of Test Questions Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?	We created 12 test questions which will essentially help us to prepare for the launching of our annotation job.
Improving a Test Question Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?	 The steps we would take: 1. We could first of all increase the size of our chest x-ray dataset to a huge amount of data. 2. To definitely highlight the regions of abnormality\ opacity in the chest x-ray image available to us, through enhancing the idea that will help to understand the explanation sentence; 3. We would also create and add as many similar test questions as possible, in order to equip an annotator example questions.
Contributor Satisfaction Say you've run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.)	 Firstly, we will go through the Test Questions because they have the minimum score. Inside the Test Questions, we will look over the accuracy of the given query as well as the explanation to the questions. In addition, we'll compare the directions and the examples to make sure they're right. As a result, we'll improve the Test examples by highlighting the many areas with issues.

Limitations & Improvements

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The sample size of 127 cases is sufficient for a preliminary review of pneumonia cases; however, this is our opinion. But there is a possibility of bias due to the absence of written examples of pneumonia on the stone, and the x-ray images can be misleading. Nonetheless, after the annotation is done, the final validation of the effectiveness of the provided method may be accomplished. As a result, it will be used to reflect the sum of bias found in the dataset.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	This Use Case necessitates the refinement of the data collection as well as the development of the Golden Corpus data. It could be achieved by comparing a pneumonia case to the rest of the results. Confirmed pneumonia cases which provide a clearer indication of which regions of the image annotator should be prioritized in the first instance.