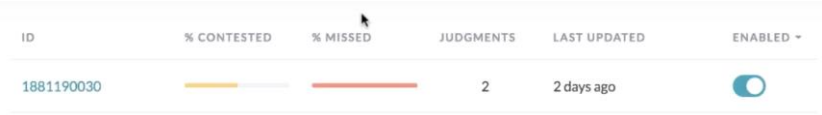
Project Proposal

Clifton Zhuo

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 problem is to detect the presence of pneumonia on X-ray scans. By using ML, healthcare workers can fasten their Pneumonia detection process, but also use it as an extra verification of the presence of Pneumonia.
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?	I went with the binary classification with additional options for uncertainty. There is the "yes" and "no" label, but I also included the uncertain option. Since this is a binary classification, even if annotators select randomly, they have a 50/50 chance to annotate correctly. To solve this uncertainty, I have put two additional labels. When the annotator chooses "yes" or "no", they must answer an additional question on how confident they are from 0-5. Optionally, there is also a blank open text box for annotators to provide their reasoning on their answer. The uncertain option also leaves room for unknown cases.

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?	8 with an even answer distribution. (Yes and No)
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?	 Augment the instructions, include more examples or redesign the job.
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.)	 All areas that are below 3.5 should be improved but starting with the lowest rated area. In this example, it would be ease of the job.

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?	I notice data contained 117 rows; this is an uneven number. This will ultimately lead to an unbalanced result. The data seems to specifically target Pneumonia. However, there may be similar unknown cases that are visually very similar to Pneumonia.
Designing for Longevity How might you improve your data labelling job, test questions, or product in the long term?	As data increase, more likelihood increases on other diseases. I may have to include other potential diseases and discard the binary classification model. As a result, I will also need to provide more test questions and train on this new data point.