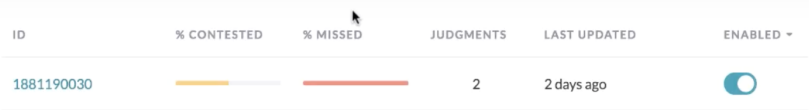
Project Proposal

<Ying Shan>

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 industry problem is that doctors need help on identifying pneumonia in children, so that a doctor can quickly make a diagnosis to help more children, meanwhile, decrease the human error rate. The reason of using ML in resolve the task is because there are some specific features/symptoms (on chest X-ray image) for pneumonia diagnosis, which is a typical image recognition task for CNN.
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 decide to choose the method of having annotators to label an x-ray image as whether it is pneumonia or not, or uncertain. Because these labels are more close to the real world task: identifying whether the child has pneumonia or not. Also, true or false question is relatively objective, which maximumly decrease the subjective judgement of annotators.

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?	I developed 11 test questions, which is about 10% of total dataset amount. There are 5 questions label as pneumonia and 5 questions label as healthy, also 1 question label as Uncertainty. We highly encourage annotator label all image as pneumonia and not pneumonia. Uncertainty is only be chosen as a last resort.
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?	 Firstly, I will check this question to analyze the reason why there is high missed rate. If the test question was incorrect (I did a mistake), I will edit it. Otherwise, I will not do anything but wait for more annotators' results. If there is a very tricky question, then I would add it into my instruction as a sample question.
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.)	 Base on the result of the above image, the Ease Of Job got lowest mark, so I will redesign my question first, make the task much easier. Probably change the question type, or decrease the choose options, or improve the rubric of the task. I will also check my test question, make sure they are typical question with no bias. Probably add a bit more test questions. Finally, I will go throw my instruction to make sure it is clearly described and easy to read.

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 size of my data set for this task is fairly small. I may want to have much more data for the real world problem. Also, some of the X-ray image was cut by part (diaphragm area was not fully show). In addition, there are a bit more pneumonia image than healthy image in the data set, which may contain biases. A better data set will have better image quality and even distribution of both pneumonia and healthy cases.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Firstly, I may want to keep adding more data with good quality into the task, in order to increase accuracy of the training result. Secondly, if there are more symptoms that are able to help to indicate pneumonia, I will update my task instruction with more judgement conditions. Also, I will carefully pick both sample questions and test questions for the task, which will cover almost all situations for the whole dataset.