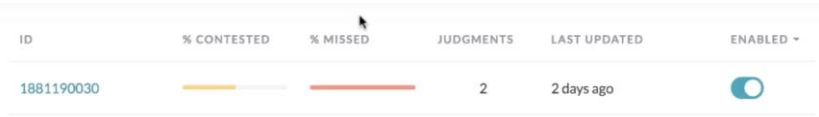
Identifying Pneumonia in Children Image Annotation Project Proposal

Dixon Liang

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 we are trying to solve with this project is misclassifications or delays of identification of pneumonia in children. Using ML, we can quickly identify clear cases which otherwise could take longer than necessary. In most obvious cases, the signs in images are clear and obvious.
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?	The choice of labels I decided to add were, “normal (no pneumonia)”, “pneumonia”, and “unclear”. In terms of our objectives, it makes most sense to categorize very simply, whether the lungs are normal/healthy or looks like they have pneumonia. A third option should also be added if it is unclear, these are the cases that need to be investigated further which are not obvious.

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 gave three test questions which cover the three potential results. One test question was a clear case of pneumonia, another test question was a clear case of no pneumonia, and finally the last test question was one that could have been as ambiguous.
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?	 If it were 100% missed, it is highly like the image in question is very misleading (excluding testing error like mislabeling). The best decision at this point unless it was our error is to throw out the image and use a different one.
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.)	 The areas I would look to improve are to find more obvious test questions. I would also make it clearer that choosing "unclear" is an acceptable answer. Perhaps, should make it clearer that the decision should be able to be made very quickly, any further thought should be classified as "unclear".

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 biases do seem to be more towards cases of pneumonia than not. It makes sense as the reason for scans is if there is suspected cases of pneumonia. The data could be improved by evening out the labeling (throwing out some of the more unclear cases that were labeled pneumonia as a suggestion).
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	A way we could improve the data labeling job for the future is continuing to refine that data set as we get better images (ex. more clear cases). We can replace older images or images that might be confusing with better ones that come.