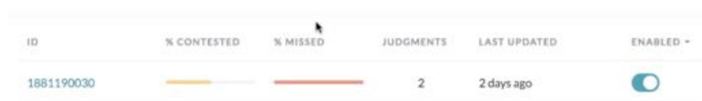
Project Proposal

Eric Armstrong

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 helping doctors to correctly identify cases of pneumonia in children based on x-ray images. ML may be well-suited for solving this task. Looking at x-rays can be monotonous, human errors in pneumonia identification may be likely. ML could be implemented in an attempt to identify the most likely cases of pneumonia in a consistent and scalable fashion. This could give doctors a starting point for diagnosis, or could serve as a 'second opinion.'
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 labels I chose to assign to the data are "CLOUDY," "CLEAR," and "UNSURE." Annotators should not feel the pressure to be making any sort of diagnosis as to pneumonia or not. What we care most about are instances of particular cloudiness in the x-ray images, particularly in the lungs and the diaphragm. The two primary labels are clear, simple, and aligned with the task. I wanted annotators to have another clear option when they were unsure. These ambiguous instances, if they can be correctly diagnosed by a human expert and reincorporated, will be very valuable for future ML model and annotator trainings. There is a possible weakness to including UNSURE as an option. As an alternative, we could simply look at contested questions.

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?**	10 Test Questions Considering the size and variation of the dataset, ten seemed to be a more than satisfactory number of test questions for getting annotators calibrated on what to look for in the images. No formula or hard calculation was used in deriving this number. It was a naive estimate. About 1 question for every 19 data points has been recommended (5%); this was surpassed. Most cases seemed to distinctly fall into either pneumonia or healthy, with a few somewhat ambiguous cases. Therefore, most of the test questions reflect either pneumonia or healthy, but we also included and unclear case to reinforce that as a possible option for annotators.
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?**	 First, I would inspect test question. I would want to understand what may have been ambiguous about it to the annotators. Once I had an understanding there, I could point out what may have been overlooked or incorrectly flagged. Explain to testers that this may be one of the more difficult instances to keep an eye on. It may be a good idea to move this test question to the “examples” section to better familiarize annotators.

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.)

Overview, steps, rules & tips should all be simplified where possible.

The next obvious step could be to improve the visual examples as well as the test questions. Clearly, annotators need more or better information to get acquainted. And with poor examples and test questions for the annotators, it's no surprise that they did not find the job that easy.

I would find most obvious examples and test questions to show first before moving on to more difficult items. For the ambiguous cases, I would be sure to add sufficient written and highlighting /marking if needed.

If the problem persists, it may be worthwhile to consider changing or rewording the answer schema entirely to something more clear. .

Limitations & Improvements

Data Source

The data size is quite small, and there seemed to be a relatively number of high cases of pneumonia compared to healthy cases. I do not believe this is representative of a typical population.

Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The data was likely cleaned and selectively assembled for this specific assignment, with pneumonia overrepresented. The data may be improved by… 1) Increasing the number of images 2) Including more labeled ambiguous cases 3) Make it representative of the real population 4) Including any other meta-data for analysis Medical professionals may also be able to weigh in on possible false positives that comes up: images that appear to show pneumonia, but in fact do not, or display some other medical condition.
Designing for Longevity **How might you improve your data labeling job, test questions, or product in the long-term?**	This job in Figure 8 can be monitored after launching. Most future changes would likely start with a look at the data. As mentioned earlier, removing the UNSURE option for annotators could have benefits or disadvantages. It may increase false positives and negatives, but it would also drive annotators to be more decisive. ...However the “design for all scenarios” camp may not be such a bad idea. Perhaps having a rating scale of confidence (in pneumonia being present or absent) would be the ideal scenario. Maybe a scale of 1-5 would be fitting, with 3 being unsure.