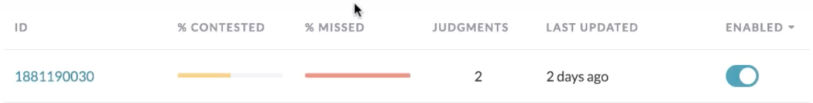
Project Proposal

Christina Loiacono

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Problem: This task is to create an aid for doctors to quickly identify cases of pneumonia in children. Machine Learning should be used in this task, to be able to provide enough data per image to sort them into 3 main categories. That way, a doctor can quickly make a decision after comparing against over a hundred images, and seeing what closely matches to the current patient. The categories are: 1. Yes, there is definitely pneumonia present in this image. 2. No, there is no pneumonia present in this image 3. Unsure – would recommend a doctor taking a closer look at this case
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?	Choice of Data Labels: By labeling each image with: Yes, No, or Other (I don't know/image invalid), that gives 3 main categories for the images to be sorted into. A clearly labeled "Yes" tells a doctor that this image is definitely pneumonia. A clearly labeled "No" tells a doctor that this image is definitely not pneumonia. An "Other" response to an image will alert a doctor that if they see a case similar to this, they should respond by not categorizing the case as being healthy, and to take a closer look at the patient for signs of pneumonia. I also gave the option to identify whether the annotator is a an expert or a non-expert at identifying pneumonia cases, which may add helpful bias if actual doctors are correctly labeling pneumonia 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?	This dataset has around 118 images. This means I should have around 5-10 test questions. I included 8 test questions, to cover all the possible answers, and avoid bias in the contributor's answers.
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?	 To improve the accuracy of a test question that almost 100% of annotators missed, I would include more test questions, to have at least 5% test questions in this job. That way the annotator will have

	about 1 test question for every 19 data points that are being labeled, out of the 118 images. In total, there should be about 5-10 test questions since this dataset is small. This will help the annotators complete the job faster. The test questions will be designed to be indistinguishable from the overall dataset, so that they follow similar instructions and the reasoning will explain how a test question answer was reached.
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.)	 These results are below 3.5, so the entire model needs improvements to become more accurate. I would go back through each question to eliminate bias. I would eliminate confusing phrases in questions to make the questions easier to answer. I would also modify the instructions based on the new changes to the questions, so there is less confusion, and help to walk the annotators through the questions as they answer them. I would repeat the important information in multiple places, that this model is meant to help identify cases of pneumonia in children. I would also add more images in the instructions to provide clarity to the annotators of what pneumonia x-ray looks like, and what a healthy x-ray looks like, showing them the cloudiness/opacity, what structures to look for such as the lungs, heart, and diaphragm.

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 bias built into this data is that there are only around 118 images, and these are all specifically x-rays of children. The data might be improved by adding another few hundred images, and having more responses clearly defined as "yes/no" rather than "other".
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	This data labeling job could be improved by adding more images, and seeing how the responses go to become more accurate, and to modify based on the results. This data would not be very reliable to be used for adults, unless they add adult x-rays, and add a question to identify whether the x-ray is a child or an adult. In the instructions, I would improve the clarity of what is pneumonia and what isn't pneumonia with more examples and images already labeled for annotators to see while they are trying to determine what they are looking for.