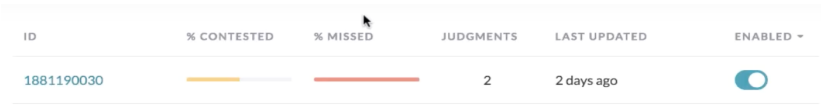
Project Proposal

Anastasia Chueva

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Pneumonia is one of the top diseases that cause mortality in children under the age of 5. Depending on the type of pneumonia, the immediate admission of antibiotics could be required. The goal of this annotation job is to create a labeled dataset that can be leveraged to train an ML model to automatically distinguish between healthy and pneumonia cases. By using ML the task of manual image observation can be automated, saving time and recourses. This is especially important for the 3rd world countries where child mortality from pneumonia is very high due to the inability to timely and correctly detect pneumonia cases because of a lack of specialists.
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?	Annotators have to answer one question: “Does the image contain any opacity or cloudy area that obscures the lungs?” And there are 3 possible answers (labels) : yes, no, unclear The task of identifying whether the image contains pneumonia symptoms may not be easy. To minimize abstractness, the annotators asked to evaluate whether the obvious symptom (cloudiness in the lungs) is present and limiting their answers to three labels: yes, no, unclear. Each label will have associated value such as following: <cml:radio value="1" label="Yes"/> <cml:radio value="0" label="No"/> <cml:radio value="2" label="Unclear"/> When the full dataset is annotated, developers can develop a model based on labels with values 1 and 0. Images that were labeled Unclear (value=2), may need to be either further clarified with subject matter experts to help understand whether they belong to healthy.

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?	The size of the dataset is 117, where 16 cases pre-labeled (1- pneumonia case, 0- healthy case). Annotators have to answer one question: “Does the image contain any opacity or cloudy area that obscures the lungs?”
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?	 Currently, there are 3 possible answers: yes, no, unclear. In case that annotators chose “Unclear” 100% of the time, adding a comment box to capture annotators feedback why it is difficult to deterministically answer whether the image contains opaqueness or not.
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.)	 I would first target test questions to ensure there is equal distribution of use cases annotated with “Yes”, “No”, “Uncertain”. Secondly, I would improve the instructions to include more details and tips for evaluating the images.

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?	Potential biases in the data: • The current size of the dataset is 117, where test (labeled) dataset is 16. Assuming that all images are labeled 0 or 1 as a result of the annotation job, it might not be enough data to train the model. • Quality of pictures (some pictures have higher resolution than others) • Equipment (equipment used for X-rays may differ) • Gender of subjects (potential biased based on organ size/deposition) • Population (all X-rays are of children, so the dataset is not representative of all age categories) • The orientation of the X-ray (some X-rays are tilted or taken under certain angel, that might create a bias for the model)
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Assuming that the goal is to create a tool that quickly identifies cases with pneumonia, I would add more specific questions like: ○ “Is diaphragm shadow visible?” Answers : Yes, No, Unclear ○ “Are there cloudy areas around the ribs?” Answers “Right side”, “Left side”, “Both”, “Unclear” In case we want to create an application that can help to differentiate between the bacterial and viral pneumonia, then we need to consult with the subject matter experts to understand how the images can be further categorized.