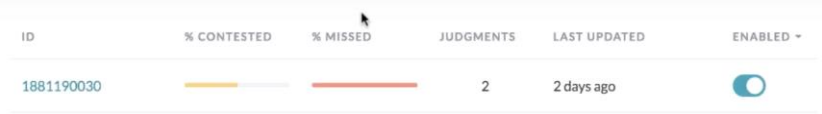
Project Proposal

Luluh Alyahya

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Using ML to analyze X-ray images to help doctors quickly identify pneumonia cases in children. Also, this model will allow doctors to swiftly rule out cases. It can also help prioritize the most severe cases and reduce the time doctors spend on diagnosis, notice that the doctor needs to confirm whether the diagnosis is correct.
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 used three labels for classifying X-rays: "yes," "no" The "yes" label indicates pneumonia (cloudy areas), "no" indicates no pneumonia (clear areas), I also added a question to rate the certainty of the answer from 1 to 5. I chose this labeling system to determine if X-rays reveal pneumonia (binary classification) and to provide clear insights into the patient's condition. This approach ensures that the diagnosis is straightforward and helps the doctor make informed decisions based on the X-ray results.

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?	9 test questions.
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?	 Confirm that the rules are easy to understand. Additionally, provide a full explanation to help the annotator with the reasoning behind the labeling and clarify the justification given. Also, in most cases you need to rephrase your 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.)	 It should add more labels and examples to make answers clearer. Check that the rules and tips are easy to understand and not confusing. Simplify the rules, tips, and questions to make them more straightforward.

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 density and extent of cloudy areas in the X-ray determine if pneumonia is present. - The data can be improved by adding more cases of pneumonia.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Augment the dataset to cover all pneumonia instances and implement routine updates. Integrate questions about family history of pneumonia and any chest pain experienced while breathing.