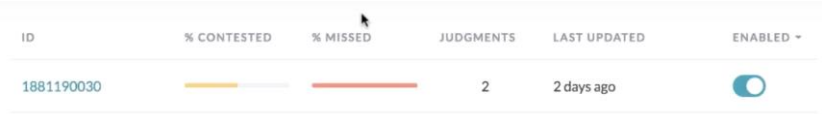
Project Proposal

Ahd Abdulrahman

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 objective is to expedite the diagnostic process for pneumonia by aiding physicians in rapidly detecting symptoms in medical images. By utilizing machine learning techniques, physicians can effectively filter out cases that do not exhibit any indications of pneumonia, enabling them to devote more attention to cases that require further scrutiny. Furthermore, the machine learning model's findings can offer physicians a fresh perspective, potentially prompting them to reconsider their initial diagnosis if it conflicts with the model's output.
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 classification task involves three possible labels for the given medical image: "yes", "no", and "unknown". The first two labels are selected to determine whether there are indications of pneumonia in the image, while the third label is included to accommodate cases where there is uncertainty. Alternatively, the labels could have been "pneumonia", "healthy", and "unknown". However, to make the classification process as simple as possible, the labels "yes" and "no" were used.

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?	Considering the size of the dataset provided in the CSV file, I provided 30 test questions. 63% was yes, 23% Unknown(N/A),13% was no.
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 a test question, clarify the language and remove ambiguity. Provide context and examples, align with learning objectives, review the rules, and provide a clear explanation. This ensures a quality question that accurately assesses the test-taker's knowledge and skills.
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.)	 Would be to provide additional examples for each category, ensure the guidelines are unambiguous, and provide further tips to enhance understanding.

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 dataset is extremely small. We have 101 unlabeled data points and 16 labeled data points. This means that the total number of data points is 117. This small dataset size could lead to significant sampling bias in our end predictions. From the overview of the project (and experience), we also know that the images are slightly different in size and taken under slightly different exposure times. This could lead to measurement bias in our end predictions. To improve the data, we must: -Increase the data size: This will help reduce the impact of sample bias. -Use a random sampling method: This will help to ensure that the data is representative of the population. -Use a reliable measurement method: This will help to reduce the impact of measurement bias. -Avoid data snooping: This will help to reduce the impact of data snooping bias.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long term?	Test questions can be improved by exposing yourself to new information and considering more complex scenarios. Additionally, it's critical to examine and update the advice regularly to make sure that it still applies and is valid considering any changes or new developments in the industry.