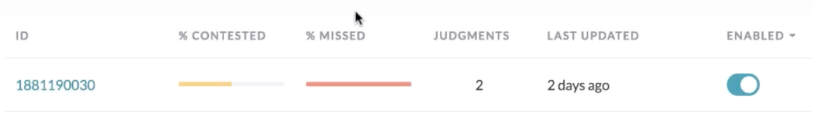
Project Proposal

Osama Alsubaie

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	This project helps doctors to quickly determine if there are pneumonia symptoms in the images provided in the data set. Using ML helps doctors to quickly eliminate cases that do not have any pneumonia symptoms and spend more time on the cases where symptoms appear. Moreover, using ML might help the doctor to rethink his decision if the model says otherwise than what the doctor thinks.
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?	There are three labels “Yes”, “No”, “Not Sure”. - **‘’Yes’’** label is chosen when there are pneumonia symptoms given the image. Chosen Yes will make new selection choices appear to select the types of pneumonias present in the image, the choices are: 1- Opaque area/s in the lungs. 2- No diaphragm shadow (diaphragm shadow is obscured) - **‘’No’’** label is chosen when there are no pneumonia symptoms given the image. - **‘’Not Sure’’** label is chosen to leave room for uncertainty. Chosen Not sure will make new selection choices appear to select level uncertainty [high, low]

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?	8 test questions were developed. The answer to 3 of the questions were yes, 3 were no and 2 were Not sure so that there is no bias towards any specific label.
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?	 By rephrasing the question to remove any ambiguities that might be present. Moreover, check that the rules and the tips specified are clear and unambiguous. Finally, improve the detailed description for why the question labeled the way it is.
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.)	 By adding more examples for each label, check that the rules and tips stated are clear and non-ambiguous and try to improve them.

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?	We might need some more data for the ML model to be robust enough to deal with all possible scenarios and train the machine learning model to learn patterns. There is no bias in the data set since all cases have the same number of labels approximately. However, if there are biases in the dataset, we need to address the issue either by augmenting the class that does not have more labels or excluding some data from the label that have more 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 if new data is added with more corner cases. Rules and tips also might need to be updated to reflect those new cases.