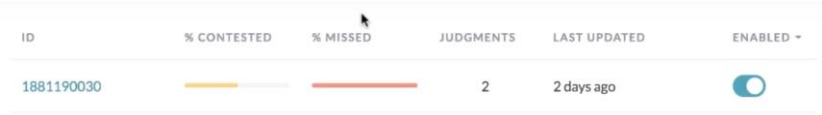
Project Proposal

Jorge Luiz Figueira da Silva Junior

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 a respiratory disease with a high mortality rate, especially in children and the elderly. A quick diagnosis would facilitate early treatment, thus resulting in better clinical results. In this project, data labeling job is designed to build a labeled dataset that distinguishes between healthy and pneumonia x-ray images. Rapid radiological interpretation is not always available, especially in resource-poor settings. Machine Learning Techniques have the potential to contribute to health professionals, assisting in the diagnosis of diseases. They allow quick classification of complex cases for huge amounts of images, with an accuracy superior to manual detection.
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?	Given that the labeled data will be used to develop a product capable of identifying the presence or not of pneumonia on x-ray images, this is clearly a binary classification problem. Then we defined the labels with yes as present and not as absent. An unknown label was included to account for uncertainty in the annotation. This method of data labeling was chosen because the objective of the problem is to find out whether pneumonia is indicated or not. On the other hand, a numerical scale approach (Likert Scale for example) could also be used, but taking into account non-experienced annotators, the first approach is simpler.

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 using 100 rows for a standard launch. I believe that 20 test questions are enough to help annotators understand what they are doing and thus maintain their accuracy.
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?	 These statistics reveal that we may have a problem with the instructions or the level of certainty of the annotators. Therefore, the steps to take are to detail the instructions or include more examples. Also in this specific case, we could take this question and bring it to the instructions as an example.
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.)	 Ease of job was the most critical point. This may indicate that the instructions provided were not enough for the annotators. To mitigate this effect, the questions would be revised, and further commented examples would be provided.

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?	As described by the data source, the images provided were taken with slightly different sizes and taken with slightly different exposure times. These effects on images can directly affect accuracy. In this scenario, different biases are being incorporated: Measurement bias: when incorrect measurements result in data distortion; Recall bias: Since some healthy lungs may appear cloudy, annotators may mislabel and data may be inconsistent. To improve the data, an image collection standardization process should be adopted, with future images being taken by the same devices and usage settings.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	It is an iterative process. Tracking the satisfaction feedback, identifying which questions are frequently missed is an opportunity to improve the data labeling job, redesign questions, rules, tips, increasing the number of examples. Additionally, the inclusion of authorities in the field in the project (such as doctors and other specialists) could enhance the quality of the instructions.