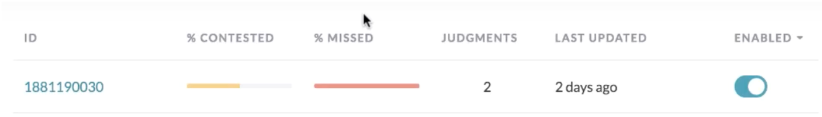
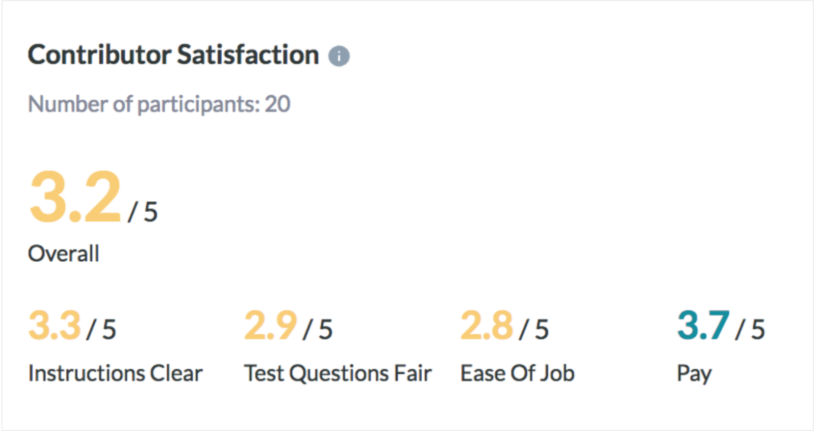
Project Proposal

Jessica Ofoh

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	My goal is to build a product that helps doctors quickly identify cases of pneumonia in children. My first task, as a product manager, is to build a labeled dataset that distinguishes between healthy and pneumonia x-ray images. Later, by using this dataset, ML can help doctors to flag serious cases, quickly identify healthy cases, and generally act as a diagnostic aid for doctors. Therefore, doctors can focus on treatment which is a more serious task.
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 basically 3 labels: • Yes, • No, • Ns (Not Sure), If yes is selected, types of symptoms are also asked by the annotators. The types of symptoms may help doctors in their treatments. Yes means the annotator is totally confident in his selection. If the annotator is not confident about his/her decision (i.e. he/she cannot either say yes or no) he/she may select Not Sure. If this option is selected the annotator also needs to select the likeliness of Pneumonia in a 1 (Not at all likely) to 5 (Extremely likely) scale. As a result with the help of all annotators and manual checks if necessary, the confidence level of the dataset may be increased and uncertainty may be decreased. One disadvantage of this labeling scheme is that we have 3 labels. However, we need binary classification (healthy or not). As a result, we need to find a method to decrease labels to 2 after the annotation is finished. If Not Sure answers are rare we can try the manual check. If not we can decide after calculating the mean of the scales. For example, if mean < 2.5 then no, else yes.

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?	We have 101 unlabeled and 16 labeled data. As a result, the total number of data is 117. As suggested by Appen, I have developed 8 test questions from labeled data which is more than 5% of unlabeled data.
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?	 ID · % CONTESTED · % MISSED · JUDGMENTS · LAST UPDATED · ENABLED ▾ 1881190030 · 2 · 2 days ago It seems that the instructions and/or example questions are not enough. First I need to reanalyze the missed question. Then: ● I can try to change the instructions and/or ● I can develop more examples to clarify the missed question. ● Obviously my test question is a tricky one. I may also make an example from this very question if I think that other test questions are enough for Quality Assurance.
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.)	 Contributor Satisfaction Number of participants: 20 3.2 / 5 Overall 3.3 / 5 Instructions Clear 2.9 / 5 Test Questions Fair 2.8 / 5 Ease Of Job 3.7 / 5 Pay From the results, it seems that the annotators didn't understand the instructions. They also seem to get difficulty answering the test questions and labeling jobs. Consequently, the first thing to do is updating the examples to clarify the labeling job. More examples may help them understand better. Next since the annotators find the instructions not clear I may try to improve the Steps and/or rules sections. No need to improve Test Questions for now. After I make the

	necessary changes I can launch a new set of data and check the feedback again. Then if necessary I can change 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?	The size of the dataset is extremely small. We have 101 unlabeled and 16 labeled data. As a result, the total number of data is 117. It is highly probable that we could get some 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. We could get some ***measurement bias*** because of this in our end predictions. In order to improve the data: ● We must get much more data, ● The size and exposure times of the new data should be the same.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	I think our data evolves in time because on new imaging technologies and new symptoms or diseases (Covid 19 for example) As a result we should use a dynamic model which is continuously trained for new data so that it can keep learning from new input. For this kind of data we may need to change our annotation job and update our data to include more relevant definitions , examples and/or test questions.