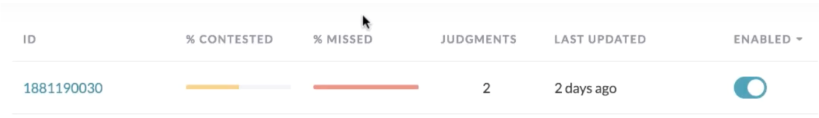
# Project Proposal

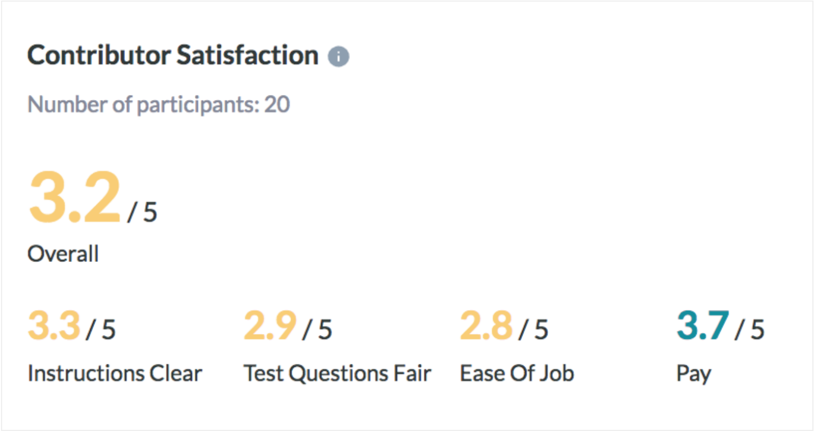
*Rishab Mudliar*

## Data Labeling Approach

| | |
|---|---|
| **Project Overview and Goal**<br>What is the industry problem you are trying to solve? Why use ML in solving this task? | The idea is to build a product that would help doctors in identifying cases of pneumonia in an easy way. There are a lot of X-ray images related to pneumonia that exist so using ML it will become easy to predict whether a patient has pneumonia or not based on the newer images that come. This will make the job easier for the doctors who can do this job very easily now. |
| **Choice of Data Labels**<br>What labels did you decide to add to your data? And why did you decide on these labels vs any other option? | I decided to add things like whether it is an x ray image because we can only determine pneumonia for x ray images. Apart from that of course determining whether a patient has pneumonia or not based on the information provided before. An option for "Others" has been provided as well for people not sure about "Yes" or "No". Finally the confidence about the answers has also been asked which will be a means to measure accuracy |

## Test Questions & Quality Assurance

| | |
|---|---|
| **Number of Test Questions**<br>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job? | 6 |
| **Improving a Test Question** | <br>ID · % CONTESTED · % MISSED · JUDGMENTS · LAST UPDATED · ENABLED<br>1881190030 · 2 · 2 days ago<br>Since most of the people missed this question. This could be a |

| | |
|---|---|
| Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question? | corner case so what we could do is specify this in examples which would be great for the annotators if they come across a similar question. We could also give explanation for why they are wrong by specifying in the explanation |
| **Contributor Satisfaction**<br>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.) | <br>Contributor Satisfaction<br>Number of participants: 20<br>3.2 / 5 Overall<br>3.3 / 5 Instructions Clear<br>2.9 / 5 Test Questions Fair<br>2.8 / 5 Ease Of Job<br>3.7 / 5 Pay<br>Here we can see that the test questions don't seem fair so we will try to include more test questions to even it out and make annotators introduced to more unique cases. Next the job doesn't seem easy so we could help make the rules and tips more clear. Adding more instructions and making it more clear about what they have to do is also a great addition. |

## Limitations & Improvements

| | |
|---|---|
| **Data Source**<br>Consider the size and source of your data; what biases are built into the data and how might the data be improved? | Many people are affected with diseases but getting affected by a particular disease is not that common. Same is the case with pneumonia. It could be that there could be more cases of people not getting affected by pneumonia. So a bias could be there. Since the dataset consists of X-Ray images and if people have not seen them, it could cause problems during annotation as they might not be able to follow the instructions. To solve the first problem we could make sure that the cases of patients having pneumonia is almost the same as that of the other case. For solving the second problem, some X Ray images could be shown to the annotators to get familiarized with the dataset. |

| | |
|---|---|
| **Designing for Longevity**<br>How might you improve your data labeling job, test questions, or product in the long-term? | As time goes on new data might come and we might want to integrate it as well. This new data might have some more corner cases that we might have to address so adding them to the examples would be a good thing to do. Apart from that we could also include questions where we ask the annotators about the quality of the questions provided which could help us improve our questions. Adding more test questions which would help annotators encounter different questions and understand the reasons as well. |