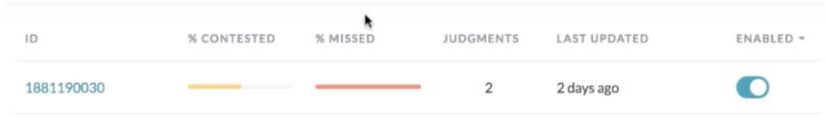
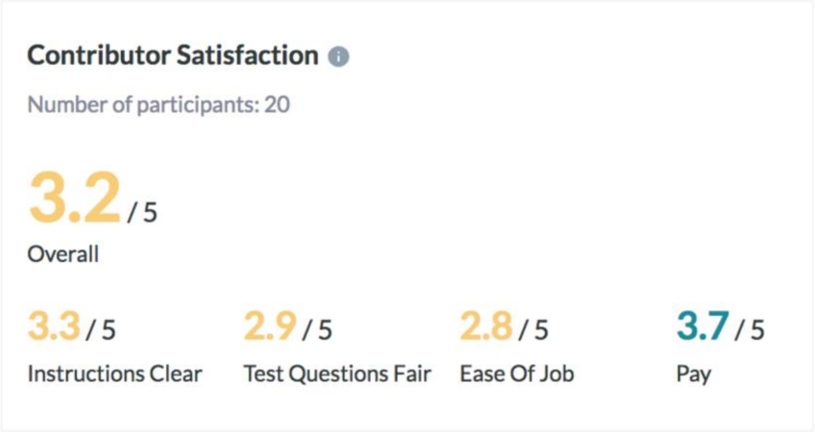
Project Proposal

Md. Abdullah Al Mamun

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	We are solving healthcare industry problem. Build an algorithm to automatically identify whether a patient is suffering from pneumonia or not by looking at chest X-ray images.
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?	• Pneumonia • Normal • Unknown We are trying to identify Pneumonia or normal by looking at chest X-ray images. But sometimes it's hard to identify whether a patient is suffering from pneumonia or not. That's why we added Unknown.

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 this dataset, we developed 8 test questions to prepare for launching a data annotation job.
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 · I would like to eliminate this question and add another question which is not very hard to answer.
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 I would like to focus on examples. Because it will clear the instructions of the job.

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?	Considering the size and source of the data, there are few x-ray images contain symptoms of pneumonia which is a limitation of data. We need to balance our dataset that's why we need to collect more x-ray images which is from a pneumonia suffering patient.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	The labels used to identify data features must be informative, discriminating and independent to produce a quality algorithm. Test questions are also very important to focus on the contributors. When added new data we will improve labeling job and increase the number of test questions and change some of them.