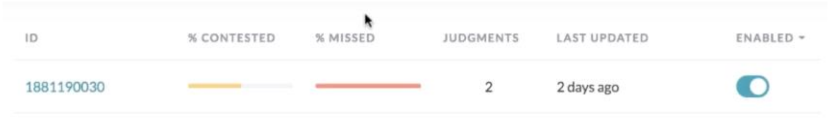
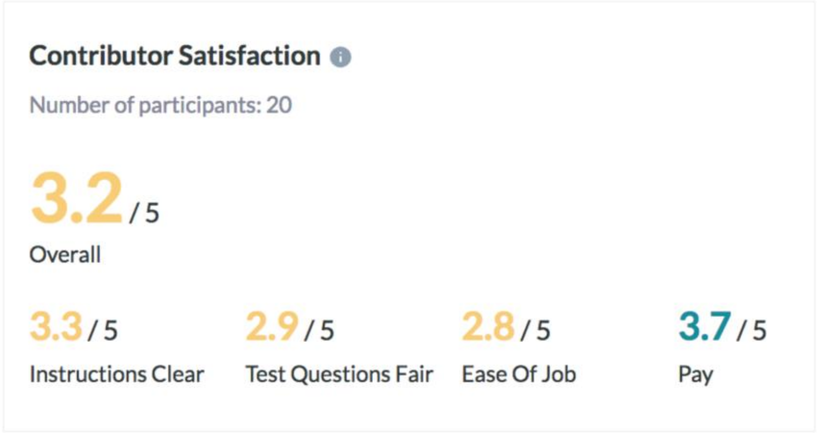
Project Proposal

Abdulelah Al-Dossari

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	determine whether or not the X-ray of a child had pneumonia. We can use it for the first diagnosis, give priority to who has the worst condition. Also, we can reduce the time taken by the doctor to give the results, and the only job of the doctor is to make sure the diagnosis is true or not.
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?	I added "yes" option if the x-rays indicated that pneumonia is present (cloudy areas), and "no" option if the x-rays indicated no pneumonia (clear and no cloudy areas). I added another question which is "What is the percentage of your certainty of the answer?" to determine the degree of confidence that the x-ray indicates whether the patient has pneumonia or not. If the x-ray is not clear, you can choose "0% - 20%" which includes uncertainty. I preferred to choose this labelling because I built this project to find out if the x-rays indicated the presence of pneumonia or not (binary classification), and to make the patient's condition as clear as possible to the doctor.

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?	10 test questions.
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 · Paraphrase the question to be more clear.
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 - Add more labels to make the answer specific. - Paraphrase Rules and Tips to make the user understand the questions and answer easily. - Paraphrase the questions to be more clear.

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 level and strength of the cloudy areas determine whether the x-ray indicates the presence of pneumonia or not. - The data can be improved by adding more cases of pneumonia. - We can ask doctors or experts to suggest questions or ideas that can make the results better.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	- Increase the number of data and add all cases, in order to make the data comprehensive and varied. - Update the data every period, and add new cases of pneumonia (if any). - We can add a question "Is there one of the patient's family has pneumonia?", because research has proved that the genetic factor plays a role in the incidence of pneumonia. - We can add to the questions "Does he feel pain in his chest when breathing?", "Does he have nausea?", "Does he have tremors?". These questions are symptoms of pneumonia, and we add them to make sure that the disease that the patient has is pneumonia, and not asthma or any other disease. - We can also make the project inclusive of all age groups, and add to the questions "how old is the patient?", and the answers are "0-17", "18-40", "41- older", because the symptoms of pneumonia in children differ from that of the elderly.