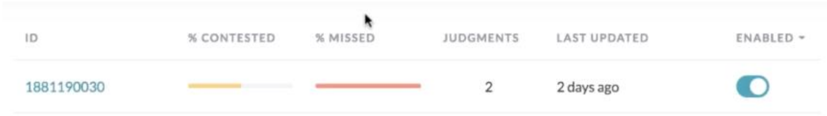
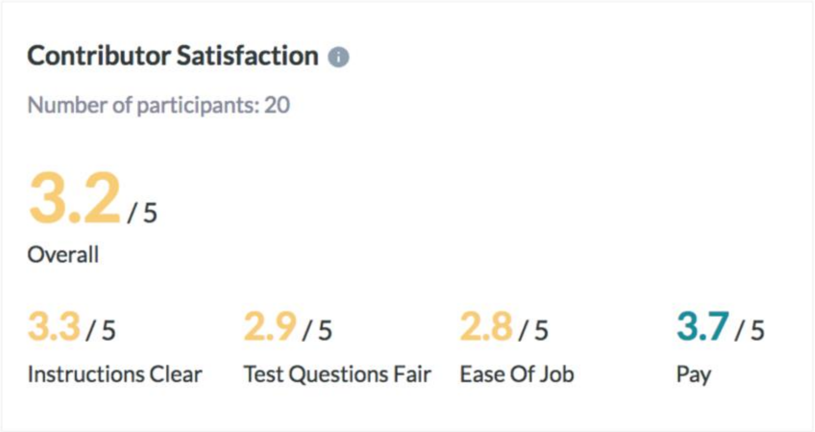
Identify Pneumonia Through Chest X-Ray Image

<Mariyyah Essam Samarin>

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	The healthcare sector has grown rapidly in recent years, owing partly to advances in artificial intelligence. Machine learning has proven to be an efficient diagnostic tool for a variety of medical disorders. To construct a model for detecting pneumonia using machine learning, labeled data is required.
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?	Because the annotators do not have a medical background, I chose simple labels. To address this issue, we used binary questions to simplify the situation. Annotators can choose 'yes' to indicate a sick patient, 'no' to indicate a healthy patient, and 'other' to indicate data confusion.

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?	Just one question” Does the X-Ray image contain pneumonia symptoms?”.
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 <the annotators missed because of a clutch in question or poor tips. However, we need to check tips and rules first and adjust them to be clear for annotators. Finally, we can change question format, we can involve the annotator in the process to get feedback.
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 <Questions for sure, questions must be related to examples meanwhile, the instructions should be construed >

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 data size is very important when it came to building an AI model. Data should be balanced to avoid biases in certain class, we can avoid these problems by collecting more data .
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	In my opinion, we can improve the data labeling process by asking more binary questions. The number of queries that can be asked is determined by the features in the dataset.