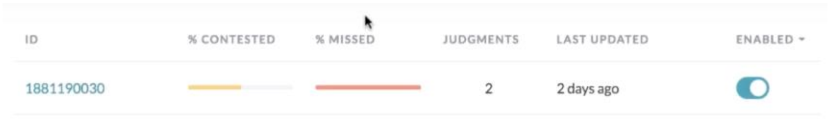
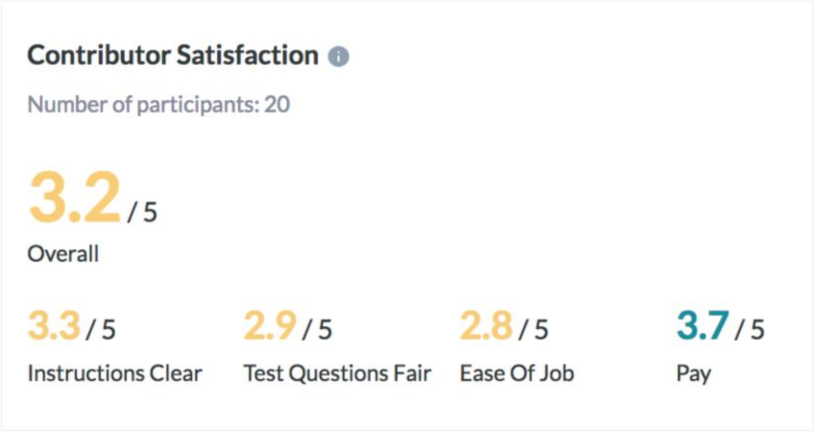
Project Proposal

Daniel Antonio Montilla Lopez

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 problem being solved is the slowness of the X-ray analysis to detect possible cases of pneumonia. Using machine learning we can streamline the analysis of these X-rays by supporting doctors with bots that help identify serious cases of pneumonia, as well as identify possible healthy cases.
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?	The labels of "yes", "no" and "unknown" were added to mark the X-ray images. These labels were chosen, as a third label "unknown" is necessary for cases where the person labeling the X-ray images. images may not be able to identify whether or not an x-ray image has pneumonia.

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?	I developed 14 test question for launching the 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 could improve the instructions, offering more details so contributors do not skip questions.
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 add more sample x-ray images so collaborators can better analyze the images, and I would swap some of the x-ray images from the tests for others.

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 biases that exist are that there are more images from one category than another, in addition, the selection of "unknown" in the test x-ray images can influence the bias of the data.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	I would improve the data labeling job by adding new test questions from time to time, with new x-ray images, as well as seek the advice of doctors to include more examples of x-ray images in the instructions.