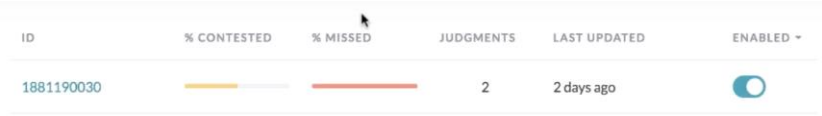
Project Proposal

<Hamad Sami AlAssafi>

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 concentrate in health industry, in this project we try to label dataset that contains X-Ray for multiple patient, then we can use it in Machine Learning to train a model to decide whether the patient is a normal or have pneumonia
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 have used three type of labels: 1. Pneumonia 2. Normal 3. Unknown I have used these labels because instead of putting (other) as an option, that will make the labeling process hardier and not beneficial.

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?	The dataset contains (118) row, and I have created (9) test questions from the dataset.
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?	 Try to explain the test questions in more understandable manner for the annotators
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.)	 Explain in detail about each test question, also if there are misunderstanding in the labels, I will try to explain it more and more, which will reflect in the contributor satisfaction rating.

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 classification is little bit hard by using the X-Ray images, especially that the most of the annotators are not doctors or have any medical background • It is best to make the data divide to (50 & 50) between (normal, Pneumonia)
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	We could uses a static model because the images most of the times have the same size and properties.