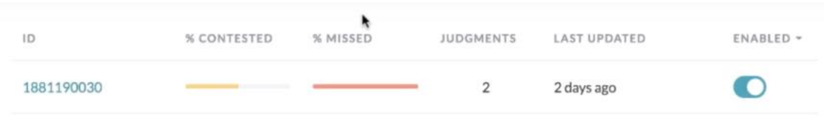
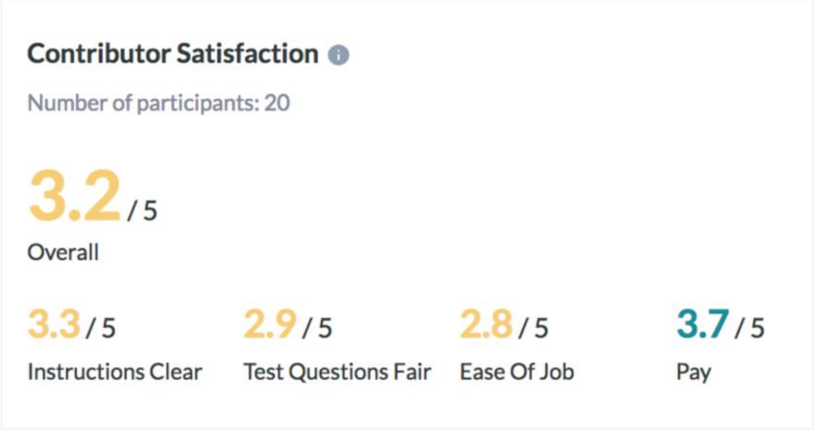
Project Proposal

<B.Akintade>

Data Labeling Approach

Project Overview and Goal • Industry problem we are trying to solve • Why use ML in solving this task?	Doctors have problems with identifying cases of pneumonia in children. The industry problem I am trying to solve is to build a product that helps doctors to help flag serious cases of pneumonia and quickly identify health cases in children. ML is used in solving this task to help build a classification system that is efficient and deliver faster result/outcome.
Choice of Data Labels • Labels added to data • Reason for choice of data label vs any other option	Choice of data labels was binary classification where the classification property is the presence of the disease - 'normal', 'pneumonia', and 'unknown; to make it easy for annotators to visually assess symptoms on presented data/images. This choice also helps to capture uncertainty in annotation and test questions. Any other choice of data label e.g measuring on a numerical scale may give room for low-confidence with annotators' decision making.

Test Questions & Quality Assurance

Number of Test Questions • Number of test questions developed to prepare for launching a data annotation job	Three test questions were prepared – Review the image, paying close attention to visual symptoms in the lungs and diaphragm. Classify this chest x-ray image • Normal • Pneumonia • Unknown
Improving a Test Question • Steps to improve or redesign test question	ID | % CONTESTED | % MISSED | JUDGMENTS | LAST UPDATED | ENABLED 1881190030 | | | 2 | 2 days ago | ● Go back into the 'quality' tab/page to either create a new test question or/and also include a reason for the answer to the question so that when the annotator misses a test question, they can see the correct answer after they submit, along with a reason for why they were incorrect. I will also ensure that I have a uniform answer distribution for all test questions.
Contributor Satisfaction • Areas of Instruction document improved (after running a test launch and gotten back results from your annotators; instructions and test questions were rated below 3.5)	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 Go into the 'quality' tab/page to check the highly missed questions, see where annotators are misunderstanding the job, and include more examples; possibly even take a particular image, and add it as an additional example in the instructions.
• Number of test questions intended	At least 5% test questions included into training set or ~1 test question for every 19 data points labelled. Eight (8) test questions created

Limitations & Improvements

Data Source • Biases built into data • Steps for data improvement	Size of data used has ~117 jpeg images, which may be subject to biases in inaccuracy, authenticity and precision because the data values might just be approximations and not be exact real values; also, data trustworthiness - largely based on the collection, processing methods, and origin of the data, data integrity might be impacted by the overall data quality. Data quality and accuracy can be improved by • getting rid of incomplete/inaccurate data; • establishing baseline approach and developing standard processes to ensure same level of image quality (e.g standardized process for chest x-ray image capture and documentation, reporting and submission of data, ensuring back-up routine for the electronic database etc.) Putting these steps in place will help to reduce likelihood of occurring errors and improve data quality
Designing for Longevity • Suggestions to improve data labeling job, test questions, and/or product in the long-term?	Changes can be made by • Setting up a dynamic model which is continuously trained on new data, so it can keep learning from new input • Updating the data to include more relevant examples, which might also result in modifying the test questions, and/or completely changing the annotation job

References:

1. Common quality issues facing Big Data
https://www.kdnuggets.com/2017/05/must-know-common-data-quality-issues-big-data.html

2. Improving data quality control in quality improvement projects
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2734082/