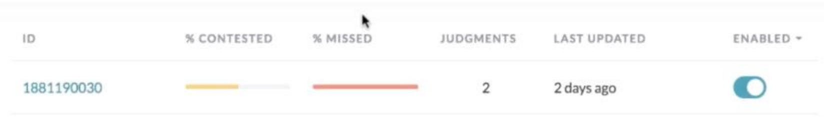
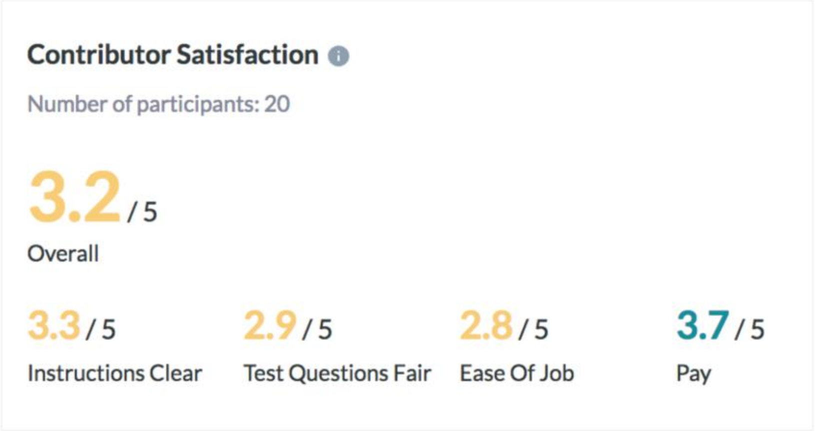
Project Proposal

Magido Mascate

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	We aim at helping Doctors to quickly identify suspicious cases of pneumonia through X-Ray Images. Pneumonia diagnosis process is suitable to apply Computer Vision Solution as there are significant stock x-ray images either as proprietary and public repositories to train and test Machine Learning Models.
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?	We decided to implement a Classification Systems, which can support mostly binary independent variable; either a person is healthy (**NORMAL**) or Infected (**PNEUMONIC**). All uncertainties are strictly flagged **SUSPICIOUS**.

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 is NOT DEFAULT maximum number of Test Questions. However, for this specific case We start with a minimum of Eight Test Questions to ensure **Cost Overrun Control** with our Minimum Viable Product (MVP). Jobs shall ensure a minimum of **20% Test Questions Balance** between Classes: **NORMAL**, **PNEUMONIC**, **SUSPICIOUS**.
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 NOT sure. It depends on the Use Case; however, I would review the Annotation Ontology to improve the Classes.
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 • Improve the Instructions and Test Questions

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?	Considering the source of the data we could **Omitted Variable Bias.** Factors such as **Smoker** or **Not Smoker Person**, or etc. can Impact on X-Ray Images Quality and so on the classes (**NORMAL, PNEUMONIA, FLU, etc.**).
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	To ensure a better data labeling, test questions We shall apply different approaches for Data Collection, Annotations. For test quality, its import to recruit field specialists to supervise the outcome.