

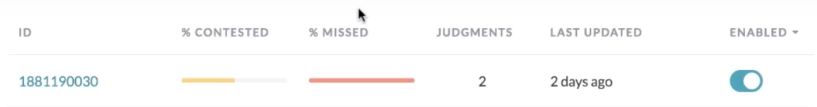

# Project Proposal

*Adalberto Gonzalez Ayala*

## Data Labeling Approach

|   |   |
|---|---|
| <b>Project Overview and Goal</b><br><br>What is the industry problem you are trying to solve? Why use ML in solving this task?                  | <b>Build a product that helps doctors quickly identify cases of pneumonia in children.</b><br>This classification system:<br>Can help flag serious cases.<br>Quickly identify healthy cases.<br>And, generally, act as a diagnostic aid for doctors.<br><br>ML has proven to outperform humans in computer vision tasks, and this can be used in benefit of the patients.   |
| <b>Choice of Data Labels</b><br><br>What labels did you decide to add to your data? And why did you decide on these labels vs any other option? | We've 3 main labels and sub sections for some of them<br>Yes                      How confident you are about your selection?<br>No<br>Not sure<br>Describe why<br><br>Yes: if you find visible symptoms of pneumonia<br>How confident you are about your selection? You can rate that serves as scale<br>No: if there's no signs of pneumonia.<br>Not sure<br>Describe why ( so we can identify unforeseen situations like x-ray contrast, shadows or any other situation) |
|   |   |

## Test Questions & Quality Assurance

|   |  |
|---|--|
| <p><b>Number of Test Questions</b></p> <p>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</p>   | <p>We've prepared 8 questions that represent a variety of situations representing the 7% of the dataset size.</p> <p>Also considering we included an option to manage uncertainty we can monitor this answer and improve the questions if necessary.</p>   |
| <p><b>Improving a Test Question</b></p> <p>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</p>  |  <p>First I would review the questions that were missed to try to identify why they were missed.</p> <p>Alternatives could be:</p> <ul style="list-style-type: none"> <li>-Cases not included in the instruction/examples</li> <li>-Mixed instructions</li> <li>-Instructions not detailed enough</li> </ul> <p>All of these depends on the cause.</p>                           |
| <p><b>Contributor Satisfaction</b></p> <p>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.)</p> |  <p>I would try to improve everything till we go as close as possible to 5, accuracy of labeling is as important as people willing to work in a comfortable environment, that means we can provide better instructions, better test Questions and look forward to build a better Ease of job. Or also could increase the pay since we know this is a very challenging task.</p> |

## Limitations & Improvements

|  |   |
|--|---|
| <b>Data Source</b><br><br>Consider the size and source of your data; what biases are built into the data and how might the data be improved? | Having more data helps, in the case of bias, social biases have to be addressed as gender or age of kids among others, in terms of data biases, control the contrast and consistency of the machines that are used to take the xrays. |
| <b>Designing for Longevity</b><br><br>How might you improve your data labeling job, test questions, or product in the long-term?             | Once the challenge of detecting presence of pneumonia gets solved then we could get more specific and illustrate where, this could help the doctors to provide an additional insight in the diagnosis and treatment.                  |