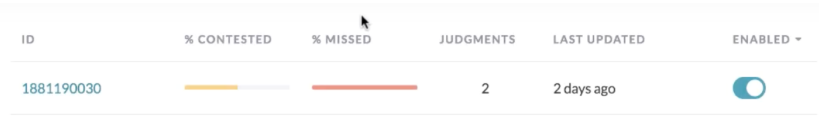
# Identifying Pneumonia Project Proposal

*Juan Martitegui*

## Data Labeling Approach

| | |
|---|---|
| **Project Overview and Goal**<br><br>What is the industry problem you are trying to solve? Why use ML in solving this task? | The goal of the project is to create a product that can help doctors identify quickly pneumonia cases in children. With ML we should be able to create a product that can go through lots of images and patients and identify which ones are likely to have pneumonia saving lots of hours of expensive doctor's time in diagnostic looking at x-ray's |
| **Choice of Data Labels**<br><br>What labels did you decide to add to your data? And why did you decide on these labels vs any other option? | I decided to label the data in 3 ways:<br>Yes, there are signs of pneumonia<br>No, there are no signs of pneumonia<br>Not sure<br><br>As the people labeling this data are not trained doctors or health professionals, I decided to make it easier for them with this labels that represent exactly what we are asking them to do.<br><br>I thought of a "scale" approach "chances of having pneumonia" but that would be too cumbersome to explain to someone. |

## Test Questions & Quality Assurance

| | |
|---|---|
| **Number of Test Questions**<br><br>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job? | As the dataset had only 117 images... I decided to keep the test questions to 10 which should be sufficient for this project. |
| **Improving a Test Question**<br><br>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question? |  <br><br>In this case I think we should rephrase the question and ad more clear examples, give more details, maybe annotating images or give more clear examples of “each label”. Also it might be a bug maybe something is wrong with the data and the option the person picks does not match in the HTML with the “correct” label. |
| **Contributor Satisfaction**<br><br>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.) |  <br><br>I believe if this is the case we should make the instructions a bit more clear and give more details and examples to the test questions seem more fair. As ease of the job is also rated very poorly maybe we might be trying to do too much for one job and we might need to split it into 2 jobs. |

# Limitations & Improvements

| | |
|---|---|
| **Data Source**<br><br>Consider the size and source of your data; what biases are built into the data and how might the data be improved? | I'm not a doctor but it seems the dataset is too small to account for all type of pneumonias in children. Also the labeled data maybe marked as pneumonia when it's too late, maybe there are different stages of pneumonia. We don't know the distribution of "ages" or "genders" of this x-rays so there might be a bias there. We should also integrate different timeframes, looks like the images are all from April 2019 maybe how pneumonia displays over time differs. |
| **Designing for Longevity**<br><br>How might you improve your data labeling job, test questions, or product in the long-term? | As we get results I'd probably would add more clear examples.<br>I would also add more test questions that are more "on the edge" to make sure the concepts are well undertood.<br>Also depending on how many "Not sure" we get we probably can eliminate that label if it is not meaningful making the job a lot easier and faster. |