PROPOSAL

DataBrew

Madagascar cough data labelling project

Summary:

There are many potential applications for automated cough detection. But in order for these applications to have a public health impact, the algorithms used for detecting coughs from raw audio input must be very accurate. Accuracy requires “training” machine learning models to distinguish between “signal” (a cough) and “noise” (everything else). Efficient “training” requires large, meticulously labelled, accurate data. We propose to manually label 31,748 .wav files of which approximately half are coughs.

Time period:

The project will take 15 working days (3 weeks), to begin at the moment of contract execution.

Deliverables:

The product of this collaboration will be:

1. A dataset of consisting of (i) filename, (ii) cough status, (iii) confidence score for all 31,748 files
2. Instructions and code explaining both how data was created and how to use it.

Budget:

The project will be carried with an “at-cost” budget:

Activity	Details	Cost
Create command line labelling tool	2 hours of data scientist time at $75/hour	$150
Training of technicians	1 hour of project manager time ($40/hour) and 3 hours of technician time ($14/hour)	$82
Manual cough labelling	15 seconds per cough (132 hours total) x 3 technicians (for cross-validation) = 396 hours at $14/hour	$5,544*
Manual cough dispute “resolution”	10 hours of project manager time to review all “disputed” labels and provide a fourth opinion ($40/hour)	$400
Combine 3 datasets and generate confidence scores and labels	2 hours of data scientist time at $75/hour	$150
Documentation, data transfer, instructions for funder, etc.	1 hour of project manager time at $40/hour	$40
On-call time for answering questions, joining follow-up meetings	2 hours of data scientist time at $75/hour	$150
TOTAL		**$6,516**

*If necessary, this item can be reduced by $1,848 if only two technicians are used for cross-validation, which would bring the total budget to $4,668