Specific Aims

A new source of biomedical big data is the mobile phone medical research application. On March 9, 2015, Apple launched its ResearchKit platform (apple.com/researchkit) for IRB-approved medical research applications to vastly scale up behavioral medical research data collection via iPhones, of which there are hundreds of millions. Our "mPower" Parkinson disease (PD) symptom tracking research study app (parkinsonmpower.org) was one of the five initial apps using ResearchKit, and in its first three months has already collected data from almost 70,000 participants. An example activity on the app is to say "Aaaaaah" into the microphone. This simple task, when performed in a laboratory, has already been shown to help classify people into those with and without PD and has shown promise at helping to estimate symptom severity of PD patients[1], as compared against a standard rating scale, the MDS-UPDRS[2]. The mPower app requests study participants to perform the voice activity three times per day. Over the course of the PD study, we will collect potentially hundreds of thousands of recordings.

However, audio quality recorded under non-laboratory conditions suffers from ambient noise and other problems that current computer algorithms are unable to account for. This means that we would need human operators to perform quality control of the mPower voice recordings for these recordings to generate useful features for analysis, and with this many recordings, we would need to enlist the help of many people. And to improve upon, and in the future automate assessment, we will need to gather information about the problems in these recordings. **Therefore, the primary objective of our project is to crowdsource the quality assessment and annotation of Parkinson disease voice recordings.**

Aim 1: Establish gold standard data to evaluate crowdsourced results.

To evaluate the voice annotations of non-experts, we will have an expert annotate over five hundred mPower voice recordings. These annotations will indicate what problems exist in each recording, and will serve as gold standards to help us determine which non-experts are performing well, and enable us to weight their contributions accordingly.

Aim 2: Crowdsource quality assessment and annotation of Parkinson voice recordings.

To crowdsource quality assessment and annotation of mPower voice recordings, we will build a Web application to collect annotations from thousands of Amazon's Mechanical Turk (AMT) "Workers." Each voice recording will be presented to multiple Workers to establish inter-rater reliability statistics. Finally, we will combine these annotations to guide automated editing or processing of the recordings in preparation for audio feature extraction and selection. As in prior research conducted by Dr. Little and colleagues[1], selected features will be compared against MDS-UPDRS scores, which we also collect from the mPower Parkinson app.

Exploratory Aim: Train a supervised learning algorithm on the crowdsourced results.

Crowdsourcing data quality control, such as proposed in this project, can solve the problem of ensuring good data quality at scales of data collection that have never before been attempted. We propose to scale this up even further by training a supervised learning algorithm on these human assessments, and evaluate how closely an automated approach matches these assessments.

This project will be the largest study ever conducted analyzing Parkinson voice data acquired over mobile phones. Such a study would simply be impossible without careful data curation, which is growing at a scale that necessitates the participation of many people. By using AMT, we will take an existing crowdsourcing solution and apply it to already gathered data to conduct our study.

Research Strategy

(A) SIGNIFICANCE

How mHealth technologies can revolutionize the clinical treatment and quality of life of Parkinson disease patients: Parkinson disease (PD) is a neurological disease with a complex constellation of symptoms that currently affects over one million Americans. In spite of the disease's complexity, the routine standard of care for monitoring the progression of PD, even in today's age of Big Data biomedical health technologies, is still a doctor's appointment, where the doctor administers a standard survey and assesses the patient's physical performance via a small number of tasks. Furthermore, due to cost and inconvenience to the patient, doctors' appointments are months apart, making it nearly impossible to objectively monitor patients' symptom progression with any meaningful time resolution. Perhaps as a result of this uninformative regimen, it is telling and not surprising that a recent study of PD patients showed that 42% of PD patients didn't see a neurologist once during the three-year study period[3].

In such a setting, we believe that distributed sensors and mobile health (mHealth) technologies such as those that PD patients could wear continuously, paired with advanced computational techniques to analyze tremendous quantities of longitudinal data, can take the neurologist's understanding and ability to treat PD to an unprecedented level of understanding. Sensor-based technologies have the potential to create new feedback loops for patients as well as clinicians that will allow better disease management through frequent, low-cost, longitudinal tracking. Ultimately this approach could scale to clinical management in PD, and then to other brain disorders, with an impact that would revolutionize quality of care, lower costs, and our understanding of the natural history of the disease.

mHealth apps: the birth of smartphone-enabled research studies with the potential for tremendous scale: The March 9, 2015 launch of Apple's open source ResearchKit platform for designing and conducting IRB-approved, electronically consented health research via a mobile phone heralds a new age of clinical research. Mobile phone-based clinical studies present an alternative to today's conventional clinical studies that rely on infrequent or short-lived patient visits, such as those currently in place for PD. mHealth studies such as those enabled by ResearchKit have the potential to alter the scope of data collection since an mHealth study can enroll anyone with a smartphone wherever they happen to be, and the frequency of data collection can happen both continuously and at designated times throughout the regular day of the study participants.

The launch of Apple's ResearchKit featured our "mPower" app (http://parkinsonmpower.org) as one of the first five research study apps to open for enrollment on ResearchKit. mPower is a PD symptom tracking research study app. In the first two months following its launch, over 5,000 individuals have enrolled in the mPower study and used the app to collect both self-reported survey and activity data. mPower uses a mix of standardized PD-specific surveys and tasks that activate phone sensors to collect and track health and symptoms of PD progression related to, for example, memory, dexterity, balance and gait. Our preliminary assessment of the mPower data collected during the memory, dexterity, balance and gait tasks indicates that they are of high quality and ready to serve as the basis for computational study.

In addition to these tasks, mPower also includes a task that prompts participants to say "Aaaaaah" into the phone's microphone. With respect to this voice phonation task, our preliminary work has already demonstrated that when performed in a controlled setting, the resulting recordings from this simple activity can accurately classify people as having or not having PD. Based on these preliminary findings, we believe mPower's voice recordings have the potential to help estimate symptom severity of a patient with PD[1].

However, in contrast to the high data quality of mPower's other tasks, the voice data not surprisingly suffers from ambient noise, interruptions and other artifacts that are currently best detected and assessed by humans. Coupled to this issue is that of scale: the number of voice recordings collected in just the first three months of the mPower study approached 70,000. Thus, in order to serve as the basis of robust computational study that is maximized for statistical power, we need an approach to data cleaning that can both meet our quality objectives as well as be implemented at the scale of this rapidly growing data set.

Crowdsourcing science with Amazon's Mechanical Turk (AMT): Amazon's Mechanical Turk (http://mturk.com) provides an online, on-demand, scalable workforce, where each "human intelligence task" (HIT) is submitted by a Requester and is performed by one of tens of thousands of paid Workers around the world. AMT is considered a convenient, affordable way to attract many workers to perform tasks that a computer is currently poor at executing. The median wage is approximately 1-2 dollars per hour, and short tasks (around 5 minutes) are awarded around 10 cents.

AMT has been evaluated for and used in many scientific studies, particularly in the social sciences[4-12] and in cognitive behavioral experiments[13-15], and is beginning to be used in clinical studies[16]. AMT has also been used to generate gold-standard scientific data, such as for the classification of medical images[17] and text annotation for clinical natural language processing[18]. Perhaps most relevant to the context of the mPower voice data set, AMT has been used in other language-related work, to crowdsource speech transcription and translation[19], evaluate spoken dialogue[20] and text quality[21], and for a variety of natural language processing tasks, including evaluation of NLP systems[22], paraphrase generation for machine translation[23], and evaluation of paraphrases[24]. *Therefore, to improve the quality of the mPower voice data recordings, the primary objective of this proposal is to crowdsource their quality assessment and annotation using Amazon's Mechanical Turk.*

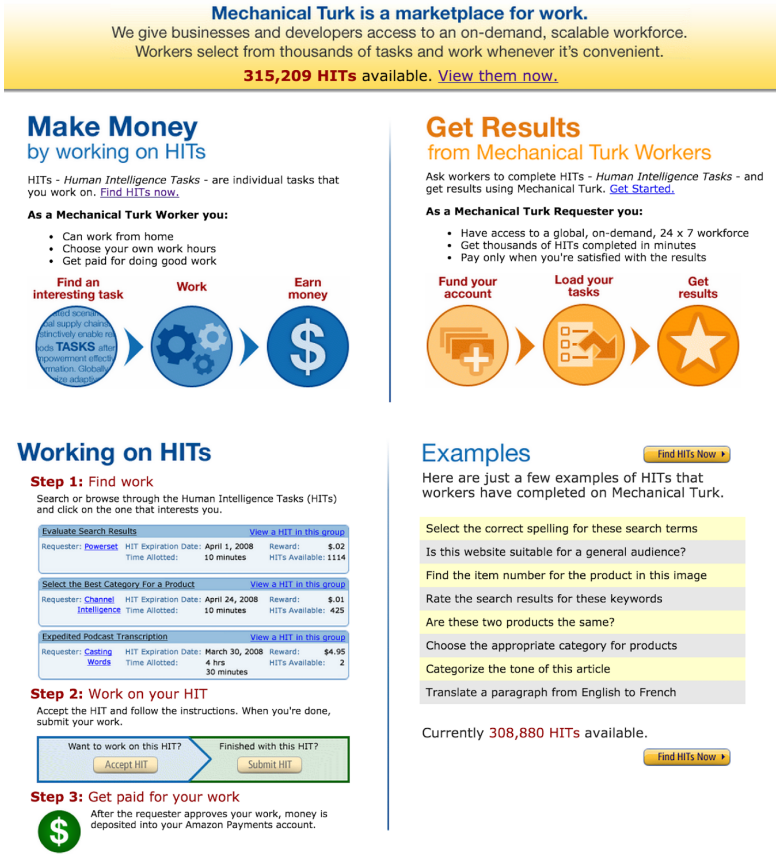

Fig.1. Amazon's Mechanical Turk Web site.

(B) INNOVATION

Aim 1 is innovative since it will introduce the world's first manually annotated set of phonations (audio recordings) acquired from individuals with and without PD. Such recordings (raw and processed) and meta-data (information about the quality of these recordings) can serve as a gold standard dataset in other areas of research, such as in the development of algorithms that automatically assess audio quality or edit audio data.

We also plan on using this dataset within an open challenge (see our DARPA seedling project description below). Since 2012, Sage Bionetworks has partnered with DREAM (Dialogue for Reverse Engineering Assessment and Methods) to run crowdsourced "Challenge" competitions. Founded in 2006 by IBM's Dr. Gustavo Stolovitzky (see his Letter of Support), DREAM Challenges engage diverse communities of experts and non-experts to competitively solve a specific problem in biomedicine in a given time period. Since 2006, DREAM has launched 34 successful Challenges, published over 60 DREAM Challenge-related papers, and aggregated a "crowd" of over 8,000 solvers. DREAM's track record of success relates to five key ingredients of its crowdsourcing model[25]:

1. Rapidly make new and often unpublished data sets available for crowd-based research.
2. Help determine if the toughest questions in science can be solved.
3. Provide an objective approach for evaluating different answers to a given scientific question.
4. Accelerate research by virtue of crowdsourcing the data.
5. Build a community of experts collaborating in real time.

Aim 2 is innovative since it will be the first instance of crowdsourcing PD data processing or analysis, and will be driven by the involvement of thousands of non-patients. Indeed, the results of this project will constitute the largest study ever conducted analyzing PD voice data acquired over mobile phones. Aim 2 is innovative also because it will create a program that will enable anyone to annotate audio files within AMT. We are aware of programs that allow a Worker to listen to audio files for transcription (e.g., http://thedesignspace.net/MT2archives/001038.html#more), but don't know of any for directly annotating or editing audio. Other researchers will be able to build on our audio annotation program once we share it through psiTurk's Experiment Exchange (see below).

The Exploratory Aim proposes to use crowdsourced edits and ratings of audio recordings to train a machine learning algorithm to perform the same task on unedited and unrated audio recordings. If this algorithm performs well, this will be an innovative and important contribution to voice data analysis and mobile phone voice quality assessment generally, and to future PD studies in particular.

(C) APPROACH

Aim 1: Establish gold standard data to evaluate crowdsourced results.

Dr. Max Little (Aston University) is a world expert in PD audio recording and analysis, having led the Parkinson's Voice Initiative (parkinsonsvoice.org) and having published research on estimating MDS-UPDRS[2] scores using audio recordings and other phone sensor measurements. The MDS-UPDRS is the most commonly used scale in the clinical study of PD, and consists of well validated, text-based questions covering different categories of symptoms.

Dr. Little collaborated with Dr. Klein on a previous project to assess the quality of audio recordings of PD patients gathered via PatientsLikeMe.org (see DARPA seedling project description below), and on the mPower Parkinson research app project. Dr. Little will himself rate and edit at least 500 of the recordings using our Web application (see below) for use as a gold standard to evaluate each Worker's accuracy and consistency. We will also use this information to weight Worker contributions accordingly. We will release these edited and annotated voice data alongside their original recordings as a project in Sage Bionetworks' Synapse platform (http://synapse.org).

Aim 2: Crowdsource quality assessment and annotation of Parkinson voice recordings. We will present 100,000 audio recordings to AMT Workers until each recording has been annotated by at least five Workers. At one cent per recording (a reasonable rate on AMT), one hundred thousand recordings annotated by five Workers each would result in a total cost of only $5,000.

Web application: We will create a Web application accessible through AMT that will enable a Worker to edit and annotate (rate the quality of) an audio recording. A Worker will listen to a voice recording, identify problematic segments of the recording (Figure 2), select categories describing why these segments are problematic (Figure 3), and rate the severity of the problems per category (Figure 4). Initial categories will include the amount of background noise, silent gaps, and presence of voiceless or rasping sounds. We will use an open source platform for setting up, testing code, posting HITs, and paying Workers on AMT called psiTurk (http://psiturk.org/, http://gureckislab.org/mtworkshop/), and will share our open source code under an Apache v2.0 license with anyone via GitHub (http://github.com) and through psiTurk's Experiment Exchange (http://psiturk.org/ee/). We have considerable experience designing and building Web apps for visualizing and interacting with data, and are poised to start building a browser-based tool with appropriate governance procedures that will enable AMT Workers to access and annotate the mPower voice recordings.

Training: The average person will not know how to edit a voice recording, so there will need to be some initial training and assessment. Training will be in two stages: tutorial, and evaluation. During the tutorial, a Worker will be presented with five or so recordings, each containing a problematic segment due to, for example, background noise. Each recording will have been previously edited and annotated by an expert (Dr. Little, Aim 1) but this information will not be shared with the Worker. The waveform of the recording will be visually displayed (see Figure 2), and when the playback reaches the problematic segment, the tutorial will highlight what steps the Worker is to take to select and annotate the segment. This could simply take the form of a screencast of an expert performing the task. During the evaluation stage of training, voice recordings containing various artifacts will once again be presented to the Worker, but the Worker is instructed to perform the selection and annotation steps. The Web app will measure the similarity between the Worker's annotations and the previously assigned expert annotations. Based on the dis/similarity, the Worker will receive feedback and further training. After training, during actual AMT sessions (HITs), every so often an expertly annotated recording will be presented to the Worker to evaluate how well the Worker is performing, and perhaps to infer effects of learning, attention, and drift. These intermittent evaluations can be used to retrain the Worker, or weight the confidence in their contributions.

Simplifying the task: Often in crowdsourcing, more complicated tasks have lower precision and inter/intra-annotator agreements. A way to handle this and collect cleaner data is to break the tasks down into smaller components[26]. To make each task easier and faster, we will test a pipeline approach that will break the all-in-one task shown in Figures 2-4 into separate tasks performed by different Workers. For example, some Workers would select problematic audio segments (Figure 2), while other Workers would annotate why these segments are problematic (Figure 3), and a third group would rate audio quality (Figure 4). This pipeline approach has other advantages, such as peer evaluation (Workers could grade/rate prior answers) and possibly directing Workers to the step in the pipeline that they are best at, based on evaluation during the initial training stage.

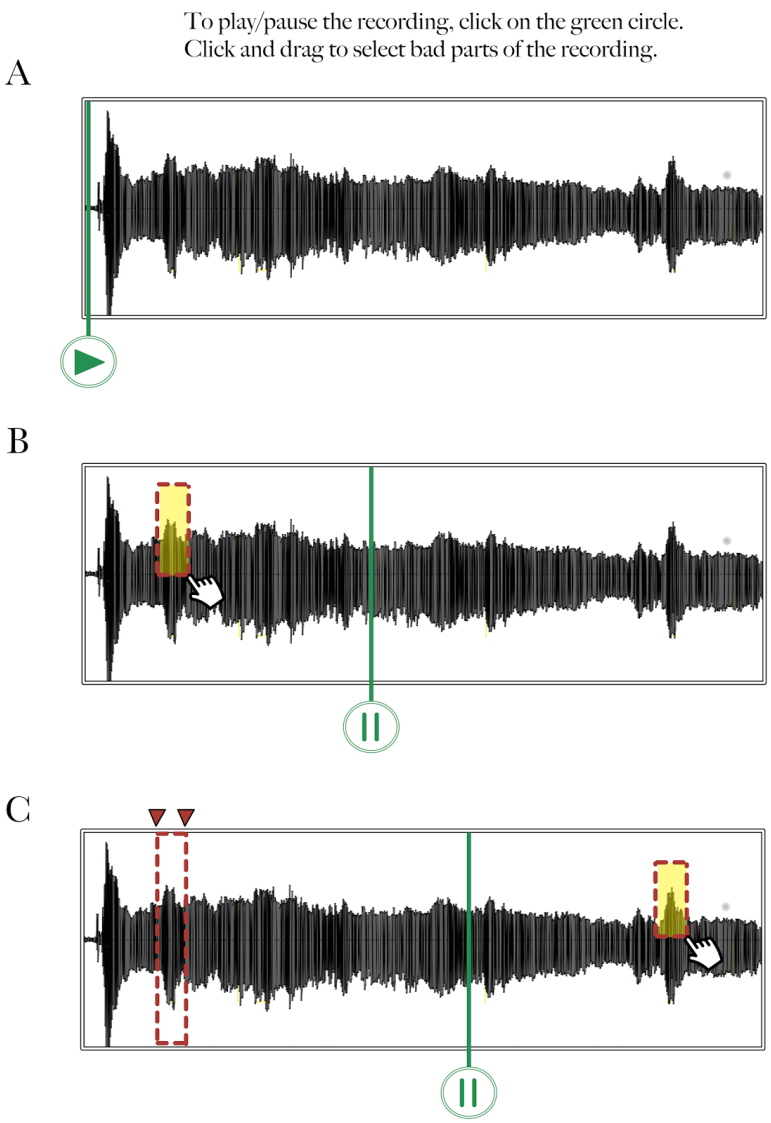

Fig.2. Mockup of audio recording annotation tool – Step 1: Selection. This figure shows a mockup of what an audio annotation Web application tool could look like. In this first step, (A) the Worker presses the Play icon to listen to the voice recording, (B) selects a problematic segment by clicking and dragging the mouse over the waveform, and (C) replays the recording if necessary and selects other problematic segments.

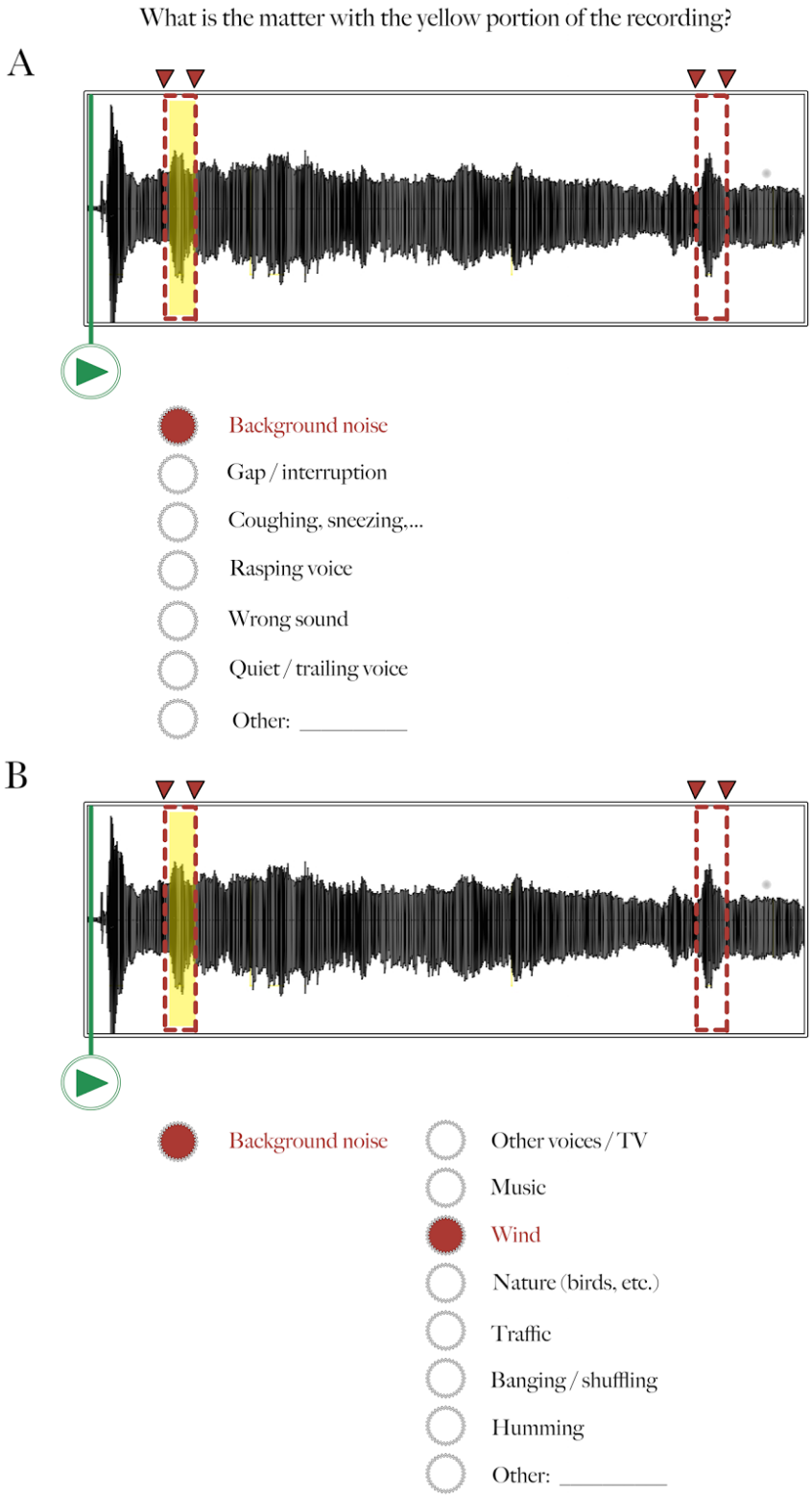

Fig.3. Audio recording annotation tool – Step 2: Annotation.
Following Figure 2, here the Worker selects one or more categories describing why the highlighted segment in the audio waveform is problematic. In this example, there was a lot of background noise (wind).

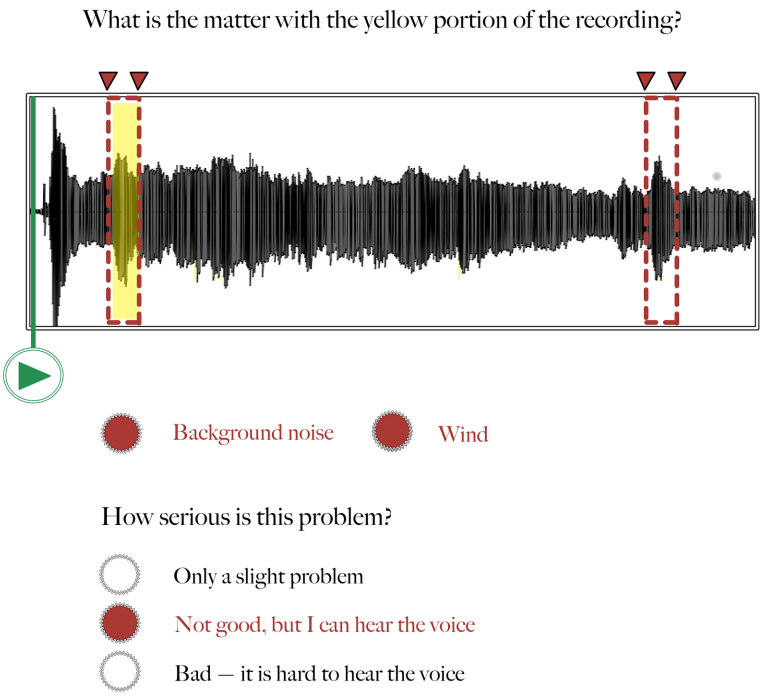

Fig.4. Audio recording annotation tool – Step 3: Rating.
Following Figures 2 and 3, here the Worker rates how serious the problem is that is affecting the highlighted segment of the recording. In this example, the Worker indicates that the background noise (wind) is not good, but that it doesn't interfere with his/her ability to hear the voice in the recording.

Aggregating the data: From at least five different edits and ratings per recording, we will assess inter-rater reliability, and be able to create a consensus edit of the audio so that only the highest quality portions of each audio recording can be selected for further analysis.

Exploratory Aim: Train a supervised learning algorithm on the crowdsourced results. We have considerable experience estimating MDS-UPDRS scores from voice data by training classifiers and regression models (see preliminary studies), such as linear regression, ridge regression, lasso, elastic net, KNN regression, random forest, boosted regression trees, etc. For this Exploratory Aim, we will use supervised machine learning methods to train on the crowdsourced annotations (start and stop points for each problematic segment of each recording, the problem category, and the severity rating). We will then evaluate how well the methods estimate MDS-UPDRS scores for a test set of voice recordings. We will use the following open source software packages: the scikit-learn (http://scikit-learn.org/) and Dato (https://dato.com) Python packages, and our own software in the R programming environment (http://www.r-project.org/).

We will design a proper semi-supervised learning model to:

* Synthesize our gold standard annotations with the AMT Workers' annotations.
* Do supervised classification of annotated audio recordings.
* Perform unsupervised classification of audio recordings which have not been annotated.

This approach would need a custom learning algorithm. One benefit of this modeling approach is that it can be transferred to other situations with a variable amount of annotation information.

Preliminary studies:

DARPA seedling project

Last year, we concluded a project in collaboration with PatientsLikeMe to acquire over 600 phone recordings of phonation in PD patients to test the feasibility of conducting a voice analysis challenge (competition). Like some of our other challenges (see Innovation section), this challenge would make biomedical data (voice recordings) available to anyone in the world to challenge them to estimate a behavioral measure (MDS-UPDRS score). After crowdsourcing data collection, as in the proposed project, the challenge would constitute a second stage of crowdsourcing, of data analysis.

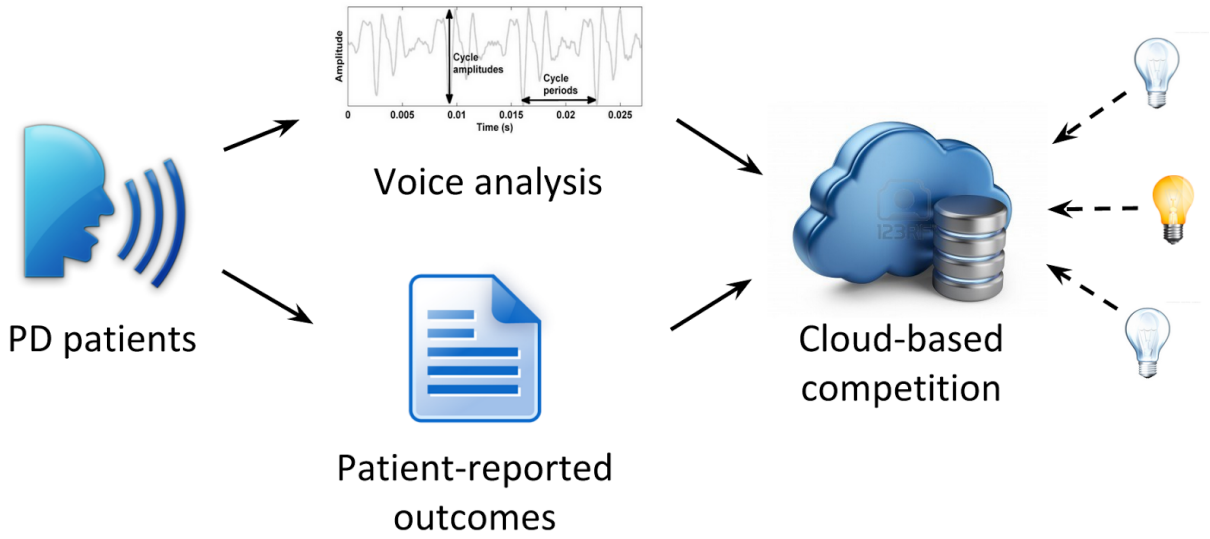

Fig.5. DARPA-funded seedling project.
This schematic represents our DARPA-funded seedling project to assess the feasibility of collecting phone voice recordings from PD patients for use in a competition.

The overall deliverables for the seedling project included:

1. The development of IRB-approved governance procedures for challenge-based data analysis of voice data and patient-reported outcomes gathered online,
2. A software-based online method to address data collection challenges,
3. A dataset of over 600 voice recordings and outcome data from people with PD integrated into Sage Bionetworks' Synapse data analysis platform (http://synapse.org), and
4. A completed dry-run to demonstrate the feasibility of conducting a crowd-based analysis challenge, accompanied by a study report addressing the lessons learned.

It was this project that brought Dr. Klein and Dr. Little together as collaborators, and clearly exposed us to problems of gathering audio data "in the wild" (landlines from English-speaking countries in that project) and the importance of very clear instructions, a uniform platform for collecting voice data, and recording on the phone itself versus transmitting over a phone line or network.
We concluded that without proper editing and quality control of audio recordings, a crowd-based analysis challenge was not feasible. It is for this reason that we propose the current project.

Android Parkinson app

With our collaborator Dr. Ray Dorsey (PD expert at the University of Rochester), Dr. Little helped to develop and test an Android app for people with PD that records activities (phonation, finger tapping, reaction time, gait, balance) to infer PD symptom severity. In a pilot study[1], the mean error predicting UPDRS (range 11-34) was 1.26 UPDRS points (SD 0.16), demonstrating that combining sensor data has the potential for estimating questionnaires. *This Android app inspired the mPower iPhone app and its treatment of the phonation task / voice activity (see Figure 6).*

mPower Parkinson iOS app

Sage Bionetworks, the University of Rochester, and Apple recently launched an iOS app that includes activities in the Android Parkinson app, passive sensor feeds, and select UPDRS survey questions. We are continuing to collect data from thousands of participants. *The mPower voice recordings and UPDRS questionnaire results provide data for crowdsourcing analysis in the current project.*

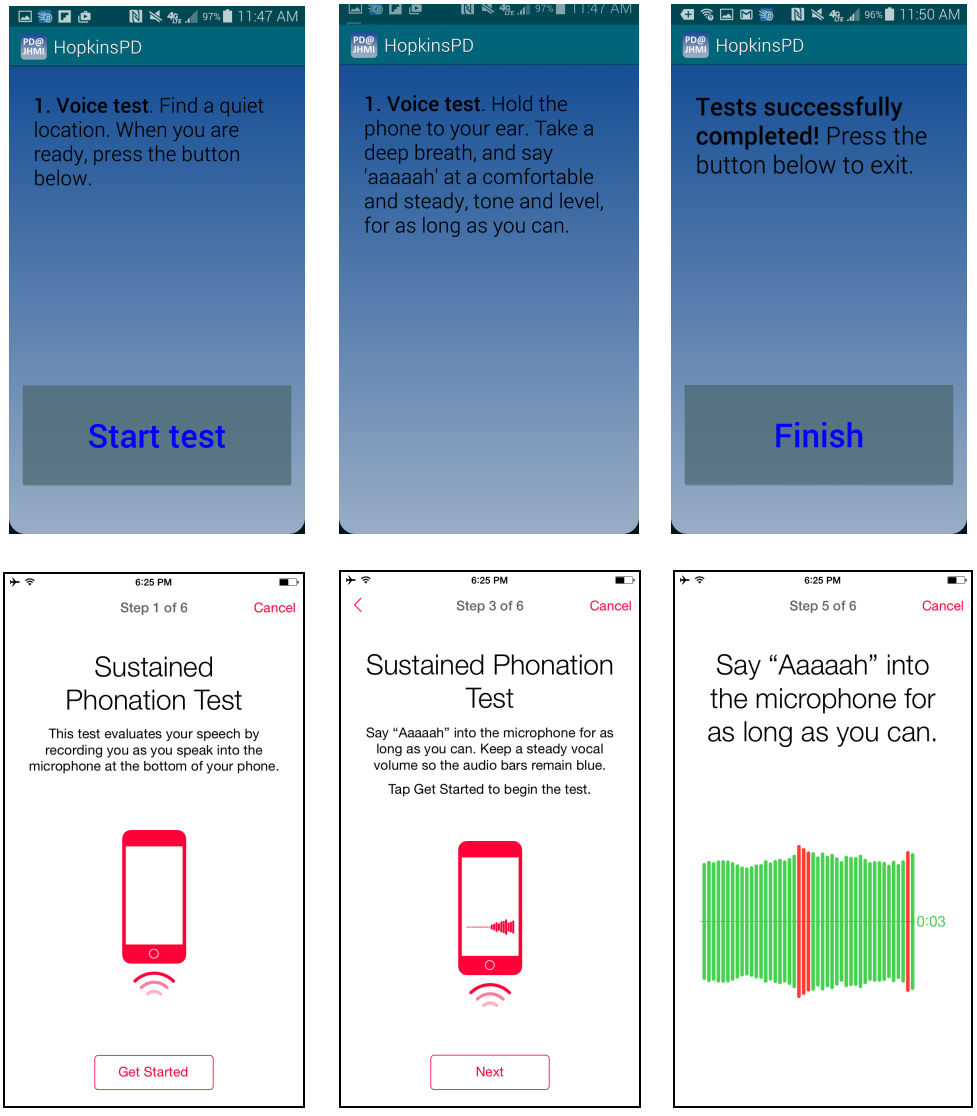

Fig.6. Android and iOS Parkinson app screenshots.

Top: Android PD app screenshots showing instructions for the phonation (voice) task.

Bottom: mPower PD app screenshots. Each participant in the mPower study is prompted to perform a voice activity three times a day. The rightmost screenshot demonstrates the visual feedback that is provided during audio recording, to try to keep the voice at the best amplitude for recording.

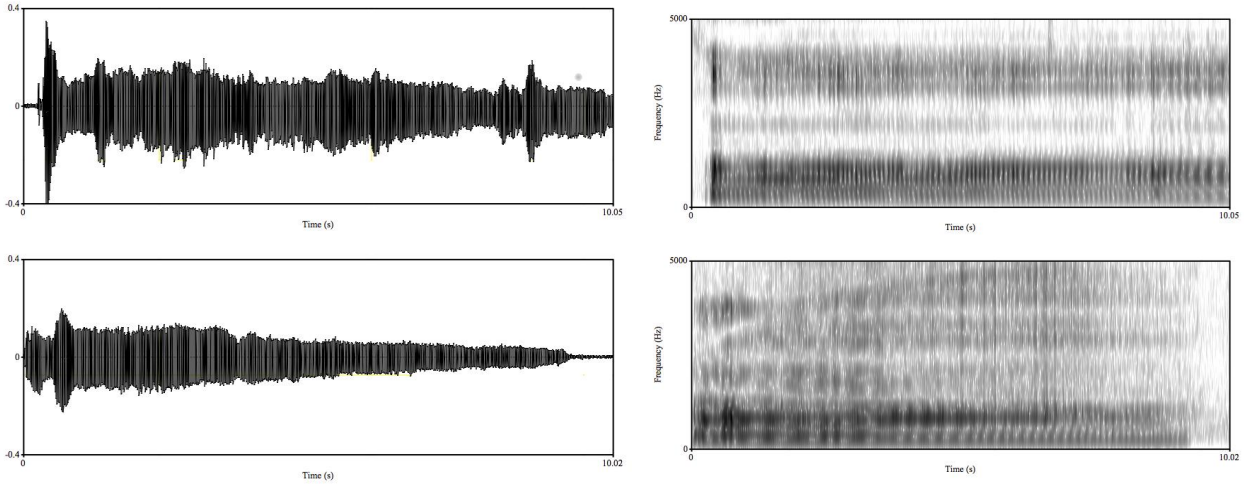

Fig.7. Example mPower patient voice data.
In the mPower app, PD patients are prompted to perform the voice activity three times per day: once before taking their medication, a second time when they feel they are at their best after taking their medication, and a third "random" time. This figure shows example voice data for a single patient on medication (top) and at a "random" time, very likely off medication (bottom). On the left are waveforms, showing the acoustic voice signal over time (0-10 seconds), from which one can clearly see that the patient's voice trailed off to a minimum (bottom left) compared to after medication (top left). On the right are spectrograms, representing signal amplitude at different frequencies (0-5 kHz) over time (0-10 seconds). The spectrogram after medication (top right) has more uniform frequency bands across the recording compared to the rather "muddled" spectrogram recorded at the random time (bottom right).

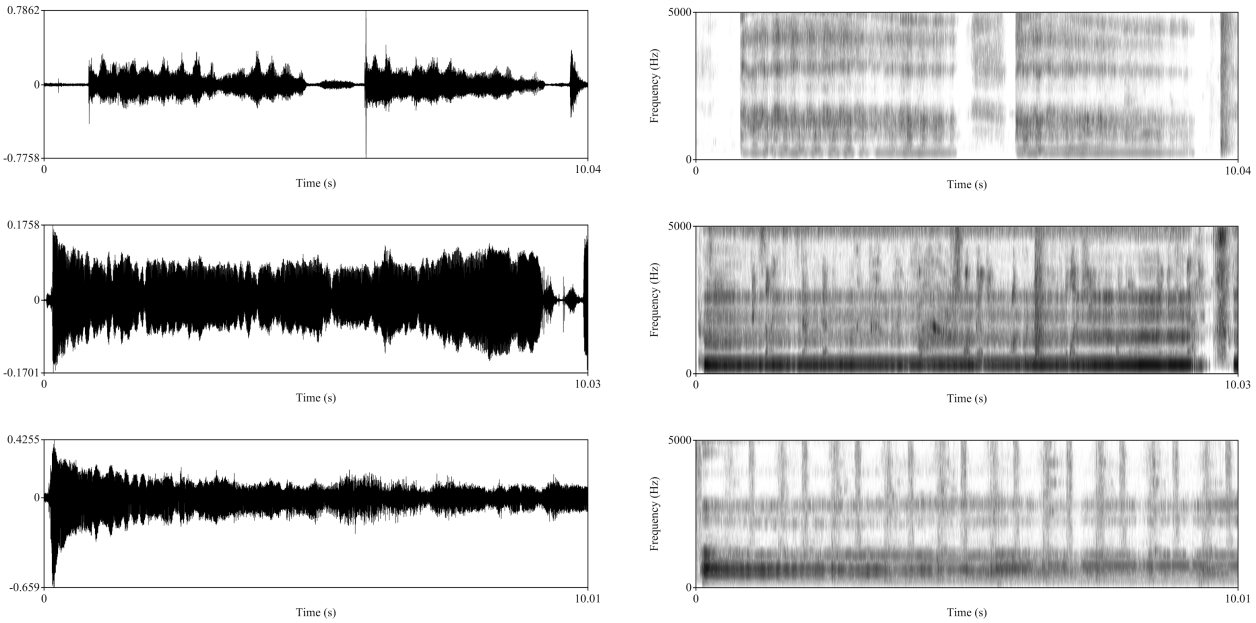

Fig.8. Example artifacts in mPower voice recordings.
These example waveforms (left) and spectrograms (right) represent the voice recordings (as in Figure 7) of three different mPower study participants. At top is a recording where the participant's voice fluctuates in intensity and stops three times as she catches her breath, visible as zero amplitude on the left and empty frequency band on the right. At the end of the recording there is a loud noise. Such a recording is easily edited and useable, as with most of the recordings, unlike the following two examples. In the middle recording, the participant did not follow the instructions, and rather than say "Aaaaaah," he said "Mmmmm," resulting in a different frequency spectrum (middle right). The different spectrum could potentially be detected by a computer algorithm, but an algorithm is more likely than a human to mistake frequencies introduced by ambient noise as the participant's voice; indeed, in the recording one can also hear different voices coming from a television program. At the bottom is a recording that suffered from a periodic, background banging noise, which is clearly visible as twenty vertical stripes that span the frequency range (bottom right).

Contingencies and Timeline:

Given our expertise in PD voice analysis and Web app development, we don't anticipate a problem creating the browser-based audio annotation tool (Aim 2) or using this tool to annotate voice recordings ourselves (Aim 1). We have listened to and reviewed hundreds of recordings, and we believe that the majority of recordings will be useable after annotation and processing.

As for the crowdsourcing itself (Aim 2), many data quality projects and scientific studies have been run using AMT, but in case we have difficulties, there are a considerable number of alternatives to AMT for crowdsourcing[27]. We should be able to attract Workers, given that we will be posting frequently and will have a large number of tasks, and it is known that Workers look for tasks that are recently posted and which have the largest number of tasks available[28]. We are fully prepared to aggregate the data, as this is one of the goals of the mPower study that is gathering the data for this project, and indeed we have begun group analyses on other mPower sensor-based data (from the tapping, gait, and balance activities).

Timeline

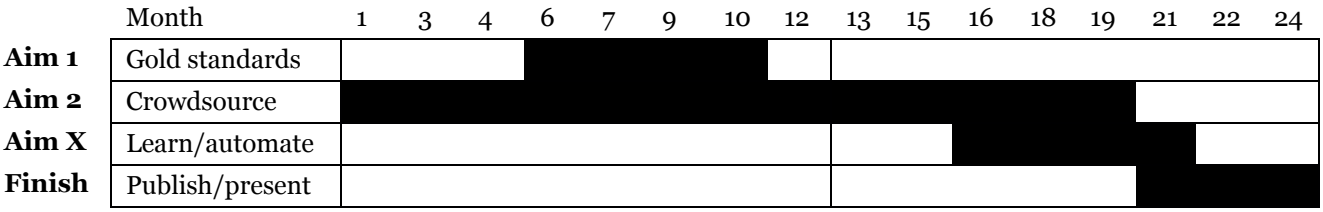

Fig.9. Timeline.
We will prepare and update the Web app (Aim 2) as we develop it for use in annotating gold standard audio data (Aim 1) and as we get feedback on its use in connection with Amazon's Mechanical Turk (Aim 2). Year 2 will consist primarily of testing the aggregation of annotated audio data for further analysis (Aim 2), to train an automated approach (Exploratory Aim), and to publish and present our findings.

Bibliography & References Cited

[1] Arora, S, Little, MA, Venkataraman, V, Donohue, S, Biglan, K, Dorsey, ER. 2013.
High-accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones. *Movement Disorders* 28(10):e12

[2] Ramaker, C, Marinus, J, Stiggelbout, AM, van Hilten, BJ. 2002.
Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Movement Disorders* 17(5):867–876. doi:10.1002/mds.10248. PMID 12360535.

[3] Willis AW, Schootman M, Evanoff BA, Perlmutter JS, Racette BA. 2011.
Neurologist care in Parkinson disease: a utilization, outcomes, and survival study.
Neurology 77(9):851–7.

[4] Paolacci, G, Chandler, J, Ipeirotis, PG. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5:411-419.

[5] Horton, JJ, Rand, DG, Zeckhauser, RJ. 2011. The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics* 4:399-42

[6] Suri, S, Watts, DJ. 2011. A Study of Cooperation and Contagion in Web-Based, Networked Public Goods Experiments, *PLoS ONE* 6(3).

[7] Amir, O, Rand, D, Gal, YK. 2012. Economic Games on the Internet: The Effect of $1 Stakes. *PLoS ONE* 7(2)

[8] Berinsky, AJ, Huber, GA, Lenz, GS. 2012. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis* 20(3):351-368.

[9] Mason, W, Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1):1-23.

[10] Goodman, JK, Cryder, CE, Cheema, A. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26(3):213-224.

[11] Crump, MJC, McDonnell, JV, Gureckis, TM. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8(3).

[12] Chandler, J, Mueller, P, Paolacci, G. 2014. Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers. *Behavior Research Methods* 46(1):112-130.

[13] Crump, MJC, McDonnell, JV, Gureckis, TM. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS ONE 8(3):e57410. doi: 10.1371/journal.pone.0057410

[14] CogSci2011 workshop: "How to use Mechanical Turk for Cognitive Science Research". 2011.

http://cognitivesciencesociety.org/uploads/2011-t4.pdf

[15] Mason, W, Suri, S. 2012. A Guide to Behavioral Experiments on Mechanical Turk. *Behavior Research Methods* 44(1):1-23.

[16] Shapiro, DN, Chandler, J, Mueller, PA. 2013. Using Mechanical Turk to study clinical populations. *Clinical Psychological Science* 1(2):213-220. doi:10.1177/2167702612469015

[17] Rodríguez, AF, Müller, H. 2012. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia* (CrowdMM '12). ACM, New York, NY, USA, 9-14. DOI=10.1145/2390803.2390808

[18] Zhai, H, Lingren, T, Deleger, L, Li, Q, Kaiser, M, Stoutenborough, L. 2013. Web 2.0-Based Crowdsourcing for High-Quality Gold Standard Development in Clinical Natural Language Processing. Eysenbach G, ed. *Journal of Medical Internet Research* 15(4):e73. doi:10.2196/jmir.2426.

[19] Callison-Burch, C. (2009) Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp: 286–295).

[20] Filip Jurcicek, Simon Keizer, Milica Gasic, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2011. Real user evaluation of spoken dialogue systems using amazon mechanical turk. Proceedings of INTERSPEECH, 11.

[21] Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp: 453-456). ACM.

[22] Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk

[23] Olivia Buzek, Philip Resnik, and Benjamin B. Bederson. 2010. Error driven paraphrase annotation using mechanical turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

[24] Michael Denkowski and Alon Lavie. 2010. Exploring normalization techniques for human judgments of machine translation adequacy collected using amazon mechanical turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

[25] Stolovitzky, G, Prill, RJ, Califano, A. 2009. Lessons from the DREAM2 Challenges. Ann N Y Acad Sci. 1158:159-95. doi: 10.1111/j.1749-6632.2009.04497.x. PubMed PMID: 19348640.

[26] Sabou, M, Scharl, A, and Föls, M. 2013.
Crowdsourced Knowledge Acquisition: Towards Hybrid-Genre Workflows.
Int. J. Semant. Web Inf. Syst. 9(3):14-41.
DOI=10.4018/ijswis.2013070102 http://dx.doi.org/10.4018/ijswis.2013070102

[27] http://www.quora.com/What-are-some-crowdsourcing-services-similar-to-Amazon-Mechanical-Turk

[28] Chilton, LB, Horton, JJ, Miller, RC, Azenkot, S. 2010. Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*(HCOMP '10). ACM, New York, NY, USA, 1-9. DOI=10.1145/1837885.1837889