Citizen Science for Mining the Biomedical Literature

Ginger Tsueng*[,1], Steven M. Nanis*[,1], Jennifer Fouquier[1], Benjamin M Good[1], Andrew I Su[1]

1 Department of Molecular and Experimental Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

(*) Equal contribution

gtsueng@scripps.edu

I. Abstract

Biomedical literature represents one of the largest and fastest growing collections of unstructured biomedical knowledge. Finding critical information buried in the literature can be challenging. In order to extract information from free-flowing text, researchers need to: 1. identify the entities in the text (named entity recognition), 2. apply a standardized vocabulary to these entities (normalization), and 3. identify how entities in the text are related to one another (relationship extraction.) Researchers have primarily approached these information extraction tasks through manual expert curation, and computational methods. We have previously demonstrated that named entity recognition (NER) tasks can be crowdsourced to a group of non-experts via the paid microtask platform, Amazon Mechanical Turk (AMT); and can dramatically reduce the cost and increase the throughput of biocuration efforts. However, given the size of the biomedical literature even information extraction via paid microtask platforms is not scalable. With our web-based application Mark2Cure (http://mark2cure.org), we demonstrate that NER tasks can also be performed by volunteer citizen scientists with high accuracy. We apply metrics from the Zooniverse Matrices of Citizen Science Success and provide the results here to serve as a basis of comparison for other citizen science projects. Further, we discuss design considerations, issues, and the application of analytics for successfully moving a crowdsourcing workflow from a paid microtask platform to a citizen science platform. To our knowledge, this study is the first application of citizen science to a natural language processing task.

key words: information extraction, citizen science, microtask, biocuration, natural language processing, biomedical literature

II. Background

Biomedical research is progressing at a rapid rate. The primary mechanism for disseminating knowledge is publication in peer-reviewed journal articles. Currently there are over 25 million citations indexed in PubMed, the primary bibliographic index for the life and health sciences. PubMed is growing by over one million new articles every year.

Despite the explosive growth of the biomedical literature, accessing the accumulated knowledge that it contains remains a difficult problem. Journal publications are primarily in the form of free text, a format that is difficult to query and access. This problem is especially pronounced for biomedical research articles because of the imprecise way that language is used to refer to important biomedical concepts. For example, the acronym PSA has been used to refer to a number of different human genes, including “protein S (alpha)”, “aminopeptidase puromycin sensitive”, “phosphoserine aminotransferase 1”, and most frequently “kallikrein related peptidase 3” (which often goes by its more common protein name “prostate specific antigen”). There are also many other uses of “PSA” outside of the context of human genes.

The challenge of structuring the knowledge represented in free text is often referred to as “information extraction”, which in turn can be divided into three subtasks. First, “named entity recognition” (NER) is the process of identifying the key concepts that are mentioned in the text. For example, named entities in biomedical texts might include genes, proteins, diseases, and drugs. Second, “normalization” is the application of standardized vocabularies to deal with synonymous concept terms. For example, mentions of “MNAR”, “PELP1, “proline, glutamate, and leucine rich protein 1” would be mapped to the same gene or protein rather than be treated as different entities. Lastly, “relationship extraction”, in which the relations between entities are characterized.

Currently, the gold standard for information extraction in biomedical research is manual review by professional scientists, a process referred to as biocuration (Krallinger et al. 2015). Although there is an active natural language processing community devoted to computational methods for information extraction (Campos, Matos & Oliviera 2012) (Torii et al. 2015) (Usie et al. 2015), the output of these methods are generally not of sufficient quality to be widely used without subsequent expert review.

Previously, we demonstrated that crowdsourcing among non-experts could be an effective tool for NER. We used Amazon Mechanical Turk (AMT) to recruit and pay non-scientists to identify disease concepts in biomedical article abstracts (Good et al. 2015). As a gold standard, we compared AMT workers to professional biocurators who performed the same task (Doğan, Leaman & Lu 2014). We found that, following statistical aggregation, crowd annotations were of very high accuracy (F-score = 0.872, precision = 0.862, recall = 0.883), comparable to professional biocurators.

Although this previous study demonstrated that non-scientists are capable of performing biocuration tasks at a high level, exhaustively curating the biomedical literature using a paid system like AMT is still cost-prohibitive. Here, we explored the use of volunteer-based citizen science as a scalable method to perform NER in the biomedical literature. We developed a web-based application called Mark2Cure (http://mark2cure.org) to recruit volunteers and guide them through a biomedical NER task.

In this paper we will address the challenges in moving from a paid microtask platform to a citizen science platform. We will assess the accuracy of citizen scientists in the same disease NER task used in our AMT experiment. We will also describe the use of site analytics for project improvement, and report the 'elements of citizen science success matrix' developed by Cox et al as one comparison point for other citizen science efforts.

III. Methods

A. Document selection

In total, this experiment describes the annotation of 588 documents. These documents were drawn from the training set of the NCBI Disease corpus, a collection of expert-annotated research abstracts for disease mentions (Doğan, Leaman & Lu 2014). Gold standard annotations for 10% of the document set (gold standard documents) were used to provide feedback and were randomly interspersed.

B. Mark2Cure Design

Mark2Cure was designed to provide a user-friendly interface for engaging members of the public to perform the biocuration task of NER. The goal of this experiment was to have all 588

documents annotated for the NER task by at least 15 volunteers. This study was reviewed by and approved by the Scripps Health Institutional Review Board.

Mark2Cure was composed of (1) a training module, (2) a feedback interface, (3) a practice module, and (4) a central 'dashboard' that organizes volunteer work into a series of 'quests'.

Training

In Mark2Cure, training was a series of four short, interactive tutorials. Training 0 introduced the basic web interface for highlighting concepts in text. Training Steps 1 - 3 introduced the annotation rules distilled from the NCBI disease corpus annotation instructions (Doğan, Leaman & Lu 2014), gradually increased the complexity of the text, and introduced a feedback mechanism to inform the user of their performance (figure S1). The tutorials were designed to provide enough guidance for the participant to perform the task well, with the constraint that overly lengthy tutorials would likely discourage participants. In total there were four tutorials.

Feedback

After a user submitted their annotations, feedback was provided by pairing the user with a partner. For visual comparison, Mark2Cure showed the user's own markings as highlights and their partner's markings as underlines (figure S2). A score was also calculated and shown based on the F-score (see data and analysis) multiplied by 1000. If the document was designated a gold standard document, then the user's partner was the gold standard annotations. If the document was not a gold standard and no other user had previously annotated the document, then the user was not shown any feedback and was give the full allotment of 1000 points. In all other cases, the user was randomly paired with a user who had previously annotated the document.

Practice

Once users completed the tutorials, they were required to work on the practice quest consisting of four abstracts in order to unlock the remaining quests. Users completing the practice doc were always paired with the 'expert user' for each of these documents.

Quests

For the purpose of organizing documents into manageable units of work, the full set of 588 documents was binned into 118 quests of five abstracts each. In addition to the per-document

point scoring system (described previously), a quest completion bonus of 5000 points was awarded upon completion of all five abstracts. For the sake of usability, users who started a quest were allowed to finish it even if other volunteers reached number of completions for that quest. Hence it is possible for abstracts to be completed by over 15 different users.

Mark2Cure is an open source project: https://bitbucket.org/sulab/mark2cure.

C. Data and Analysis

Mark2Cure was set up with Google Analytics for site traffic analysis. During the experiment period, emails were sent weekly to the mailing list via Mailchimp, and Mailchimp analytics provided open and click rate information. Mark2Cure also logged information regarding user sign ups, training and submissions. Precision, recall, F-scores, data quality and cost metrics were calculated as previously described (Good et al. 2015). A survey was sent via email to the subscribers on the Mark2Cure mailing list at the end of the experiment (379 subscribers), and 78 of the subscribers (20.5%) responded to the survey. The metrics developed by Zooniverse were loosely applied to study the success of the project, though these metrics are better suited for project suites with multiple longer-running projects. The active period for this experiment started on January 19th and ended on February 16th 2015. An export of the data generated in this experiment (which was used for the analysis) can be found at: https://figshare.com/articles/Mark2Curator_annotation_submissions_for_NCBI_disease_corpus/2062554

IV. Results

A. Training and Retention

Training 0 introduced the highlight mechanism underlying Mark2Cure. Training 1 introduced rules for highlighting disease mentions in a sentence-by-sentence manner. Training 2 allowed users to practice the rules they had learned while acclimating the user to longer spans of texts. Training 3 introduced the feedback and scoring mechanisms. User tests prior to the launch of this experiment indicated that all training steps could be completed in under 20 minutes, but user drop off had not been determined at that stage.

Based on Mark2Cure's log files, 331 unique users completed training 1 (figure 1A), 254 unique users completed training 2, and 234 unique users completed training 3. User drop off was

highest between Training 1 and Training 2, but 92% of users that completed Training 2 went on to complete Training 3. Of the 234 unique users that completed the training, over 90% of the users (212) contributed annotations for at least one document. To better understand the user drop off within the different training pages, we obtained the unique page views and average page view time for each training page using Google Analytics (figure S3). Problem pages within the tutorials identified with Google Analytics were confirmed by emails received from users having trouble on those pages.

In addition to responding to each email received, Mark2Cure published 204 tweets, nine blog posts, and sent eight newsletters during the 28 days that this experiment was running. We estimate having 466 email interactions with the users and 221 other communications during this time yielding values of 0.282 and 0.594 for our communications and interactions metrics respectively (table 1B). Post-survey results indicated that user interface issues were a common (but generally surmountable) obstacle to participation (figure 2).

B. Recruitment and Analytics

Five months prior to the launch of this experiment, we began to blog periodically about Mark2Cure and engage with the rare disease community on Twitter. We focused on the rare disease community because many members of this community are highly motivated to read scientific literature, engaged in research, active on social media channels, and experienced with outreach. By the time this experiment was launched, we had a mailing list of 100 interested potential users and 75 followers on Twitter; however, only about 40 users signed up after the first 10 days of the experiment's launch.

Sessions from new users peaked the day after the article on Mark2Cure was published in the San Diego Union Tribune (figure 3). A second, smaller surge in sessions from new users was observed the day after a blurb on Mark2Cure was published on California Healthline (on February 05); however, total sessions (from both new and returning users) rivaled that of the one seen from the San Diego Union Tribune article.

New users from social media peaked around Feb. 13th and 14th, as existing users were actively marking documents, tweeting about their activities, encouraging other users to do the same, and luring new users to try Mark2Cure. Over the course of this experiment, one user posted over 50 tweets and generated Mark2Cure specific hashtags to encourage others to join the effort. New users from social media peaked again on Feb. 16th with the release of Global Gene's

RareCast podcast interview featuring Mark2Cure. As with many large rare disease communities and organizations, Global Genes has a strong social media presence.

Survey results at the end of the experiment indicated that 67% of respondents learned about Mark2Cure from the newspaper, 24% learned about Mark2Cure from social media channels, while 15% learned about Mark2Cure from a friend or a google search (figure 5C). In addition to recruiting via traditional press and social media channels, recruitment was also driven by members of the NGLY1-deficiency community. Participants from this rare disease community made a concerted effort to ensure that this experiment would be completed successfully, with the hopes that Mark2Cure could be applied to accelerate research on NGLY1-deficiency. One participant from this community approached the Missouri Military Academy (MMA) and recruited both instructors and students to participate. Over 24 participants from MMA contributed about 10% of the task completions.

C. Distribution and Performance

Over the 28-day experiment, 212 users submitted 10278 annotated abstracts with over half of the annotated abstracts submitted in the final week of this experiment. As with all crowdsourcing systems, the distribution of task completions among users was highly skewed. The top 24% of users completed 80% of the tasks, and the top 6.1% of contributors completed over 50% of tasks (figure 4A).

Overall, the accuracy of contributions relative to the gold standard was quite high. The average F-score per user across all of their annotations was 0.761 with a standard deviation of 0.143 (figure 4B), which was on par with that of our previous AMT results (Good et al. 2015).

To assess the aggregate accuracy across all users, we computed the “minimum percent agreement” for each annotation, defined as the number of users marking a given annotation divided by the total number of users to process the document (usually 15), and computed accuracy statistics at multiple thresholds (Figure 4D). At 0% minimum agreement (taking the union of all user annotations), we observed precision of 0.274 and recall 0.976. At 100% minimum agreement (the intersection of all user annotations), we observed precision of 0.992 and recall of 0.122. The maximum F-score of 0.836 was reached at 40% minimum agreement.

These accuracy results were also very similar to our AMT experiments (figure 4C), in which precision ranged from 0.444 to 0.983 and recall varied from 0.980 to 0.321 (at a 0% and 100% agreement threshold, respectively). The maximum F-score of 0.875 was also reached at a

minimum agreement threshold of 40%. For comparison, these aggregate maximum F-scores were also on par with the individual expert annotators who performed the initial phase of annotations for the disease corpus (Doğan & Lu 2012).

The difference in maximal F-score could be attributed to the differences in the way the rules were presented in AMT vs Mark2Cure, or due to the difference in incentive structures and goals. In AMT, poor performers can be blocked or may self-select out of a task so as to not affect their ability to qualify for future tasks (poor performance reports can affect an AMT workers ability to qualify for jobs). In contrast Mark2Cure contributors are never blocked and are encouraged to continue contributing as data quality is only one of several metrics by which success is determined. Overall, the performance by citizen scientists was on par with that of AMT workers on this task, even though the demographics of the citizen scientists were quite different from those of our AMT workers.

To assess how increasing the number contributors affected the quality of the aggregate annotations, we simulated smaller numbers of annotators per document by randomly sampling from the full dataset. We found that the greatest increase in F-score was observed when the number of annotators was increased from two (F=0.690) to three (F=0.783) (figure 4E).

D. Demographics

The majority (65%) of survey respondents cited the ‘desire to help science’ as their motivation for participating (figure 5A). Though not necessarily representative, the results of our survey suggest that the participants in this experiment were demographically quite different from our AMT experiments. Women were more likely to participate (or report their participation) in our survey than men (figure 5D). On average, our participants were older than the participants from the AMT experiments (figure 5E), in part due to the readership/recruitment from the San Diego Union Tribune article (figure 5C). After the publication of that article, we received many inquiries about participating in the project from citizen scientists who volunteered information about their employment status and age (particularly from retirees). In pooling the contributions from just 14 of the 212 participants (6.6%) who volunteered this information (without ever being asked), we found that seniors and/or retirees contributed at least 26% of total document annotations in this experiment. In our demographic survey, 18 respondents (25%) reported being 66 years of age or greater (figure 5E), and 25 respondents (35%) reported being retired (figure 5F); hence, the actual contributions from this demographic group are likely to be higher. Student participation

was underreported (3%) in our survey and does not reflect the concerted effort of students from the Missouri Military Academy.

E. Elements of Successful Citizen Science Matrix

Data quality has been the focus of much academic research on citizen science, but it is not the only measure of success. Many citizen science projects have additional goals of engaging people in science and motivating them to incorporate scientific thought, hence the process of engaging citizen scientists can in itself also be a measure of success (Freitag & Pfeffer 2012). As one of the most established metrics, data quality is the most uniformly examined. Recent efforts have been made to define, expand, and apply additional metrics of citizen science success. The Zooniverse project created a set of useful metrics that could be applied across many citizen science projects. These metrics were normalized internally so that different projects by the Zooniverse team can be compared with one another. Although these metrics can be very useful for internally evaluating different projects, the normalization used in the Zooniverse paper means that that the reported results cannot be used for comparison purposed by researchers with only one citizen science project. Hence, we calculated and reported all the Zooniverse metrics as one data point to which other citizen science projects can be compared (table 1).

V. Discussion

In the 1980's, an information scientist named Don Swanson found that several abstracts about dietary fish oil contained mentions of blood viscosity, platelet function, and vascular reactivity. Swanson also found the same terms in abstracts from a disparate body of literature surrounding Raynaud's syndrome, allowing him to uncover the relationship between Raynaud's syndrome and fish oil—an undiscovered relationship even though all the information to establish the link was already publicly available (Swanson 1986). This hidden knowledge was uncovered by Swanson at a time when the biomedical literature was growing at an annual rate of about 10,000-15,000 articles. The rate of biomedical literature publication now exceeds 1 million articles per year and represents a body of knowledge that is increasingly difficult to harness. Information extraction is a necessary step towards harnessing the undiscovered but already available knowledge; however, it encapsulates some of the most time-consuming tasks in biocuration.

As the number of professional biocurators shrinks relative to the volume of literature to be curated, alternative strategies for keeping pace need to be explored (Howe et al. 2008). The entrance of citizen science into this domain opens up many new opportunities. Most immediately, citizen scientists can help to generate new annotated corpora for training and evaluating computational methods for information extraction (Good et al. 2015). Following Galaxy Zoo's example (Richards & Lintott 2012), we can set up computational systems that learn to perform the current tasks of the citizen scientists. Once these methods reach acceptable levels of performance, the citizens can be directed towards other areas still in need of human input.

Apart from helping to develop computational methods for information extraction, citizen scientists can help at much higher levels than machines are likely to reach. A motivated community of citizen scientists can accomplish nearly any goal, including the development of their own computational methods for solving complex tasks (Khatib 2011). Within the domain of biocuration, citizen scientists could help with challenges like: prioritizing the most interesting documents for curation, developing controlled vocabularies, annotating images in documents, creating summaries, or even soliciting funding for research. The scientific possibilities are limitless. To achieve them will require the development of strong synergistic cooperative relationships between professional biomedical scientists who can turn the scattered information in the public domain into new knowledge, experts in information extraction who can help dramatically increase the efficiency of processes for integrating that knowledge, and the large communities of citizens whose health and well-being will ultimately benefit from this work and who can become fundamentally important contributors to the process. The production of such a 'problem solving ecosystem' (Michelucci 2016), mediated through software such as http://mark2cure.org, represents the ultimate aim of the citizen science movement and is our primary objective moving forward.

VI. Conclusion

As a citizen science project, Mark2Cure would be classified as a science-oriented virtual project (Wiggins & Crowston 2011) subject to the issue of ensuring valid scientific results while designing for online participation/interest. We addressed data quality issues as other citizen science opportunities have, by using replication across multiple participants, having participants evaluate established control items, and using a corpus of text that had already been expertly

reviewed as a benchmark (Wiggins & Crowston 2014). As Wiggins (2014) pointed out, data quality issues in citizen science are often project design issues; hence data quality can often be improved by analyzing participant interaction and adjusting the design. By using the gold-standard NCBI disease corpus (as we did in the AMT experiments) and formulating our tutorials around the annotation rules set forth in the development of that corpus, we demonstrate that citizen scientists are willing to perform biocuration and (in aggregate) can perform comparably with expert curators. Furthermore, we demonstrate how researchers might utilize site traffic information and logged data in order to improve aspects of the design in order to achieve quality data.

This experiment successfully demonstrates the application of citizen science to NER and illustrates how citizen scientists might contribute to the ongoing efforts towards extracting information from biomedical literature.

VI. Acknowledgements

This paper would not be possible without the citizen scientists who participated in this experiment, and continue to contribute to our current efforts at http://mark2cure.org. We thank the members of the NGLY1-deficiency disease community who helped with our recruitment efforts and the faculty, staff, and student participants from the Missouri Military Academy. We especially thank Bradley Fikes of the San Diego Union Tribune for his excellent piece on Mark2Cure, which helped immensely in our recruiting efforts.
Contributors who gave us permission to publish their names can be found here: http://mark2cure.org/authors/beta_experiment

This work was supported by the US National Institute of Health (U54GM114833 to A.I.S.). This work was also supported by the Scripps Translational Science Institute, an NIH-NCATS Clinical and Translational Science Award (CTSA; 5 UL1 RR025774).

Tables and Figures

Figure 1

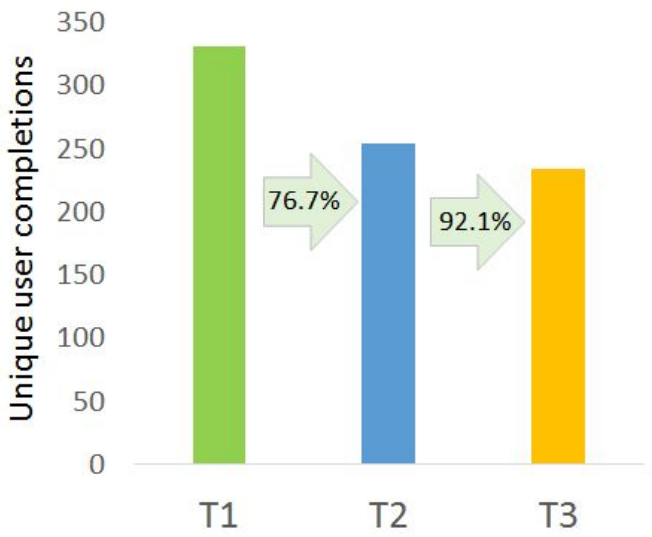


Figure 1: Training drop off and retention based on Mark2Cure training completion logs.

Figure 2:

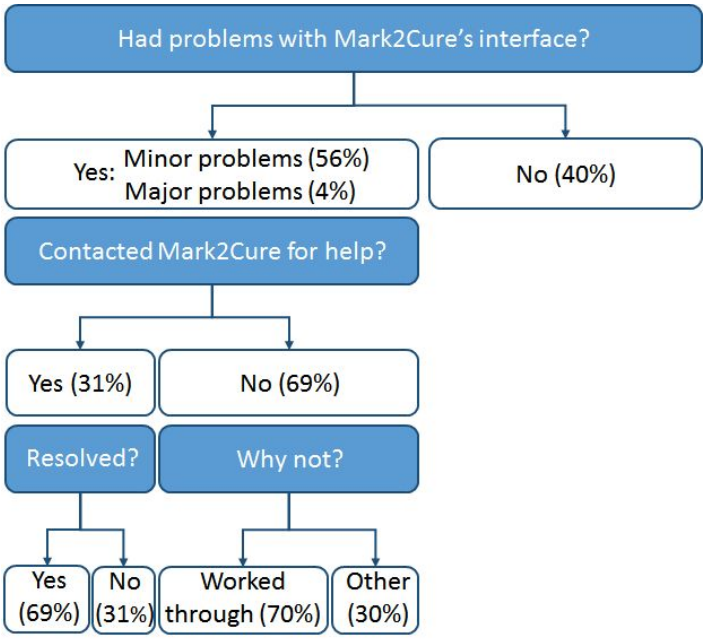


Figure 2 - Participant difficulties with the interface, attempts at resolution, and satisfaction with assistance provided.

Figure 3

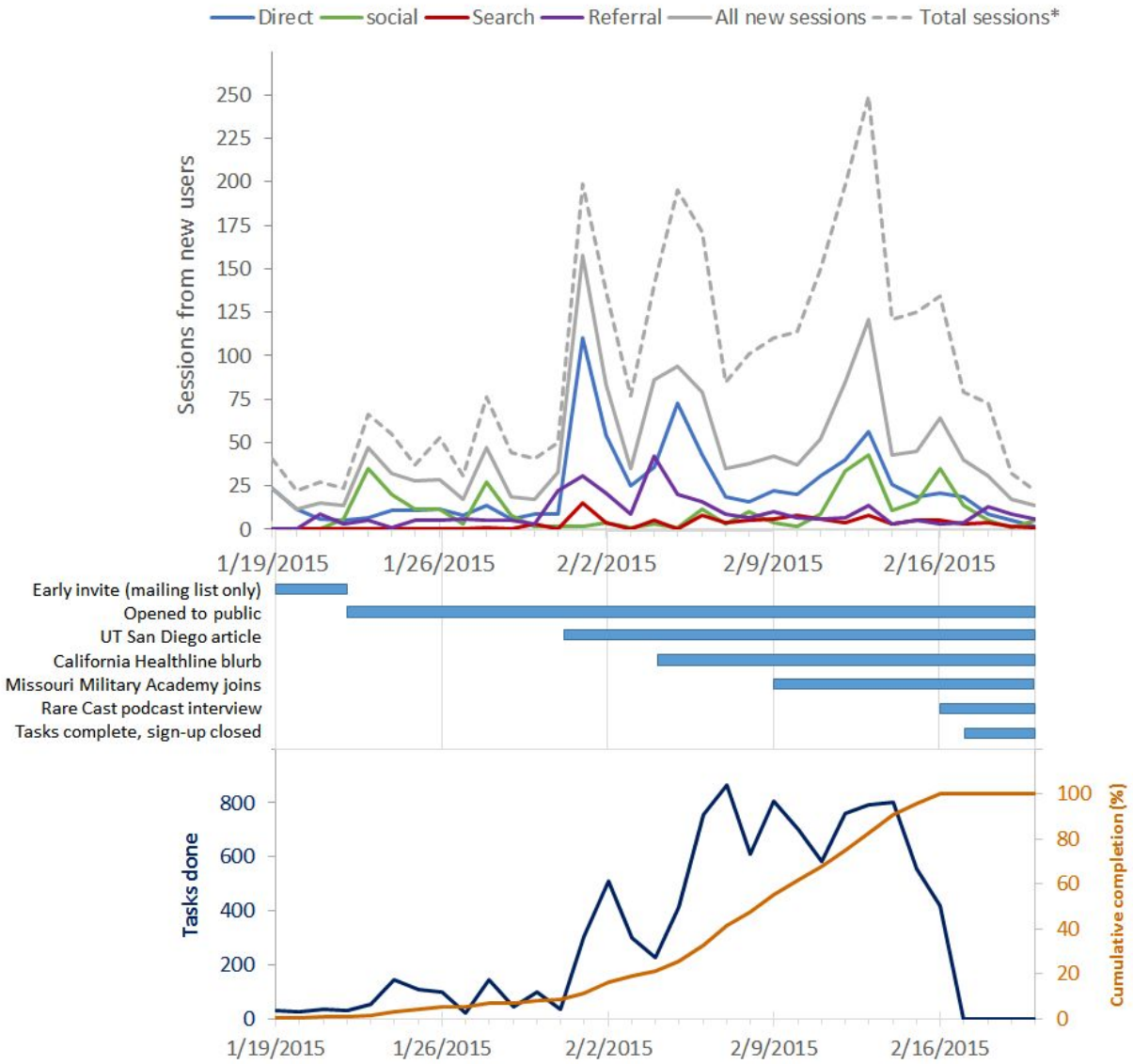
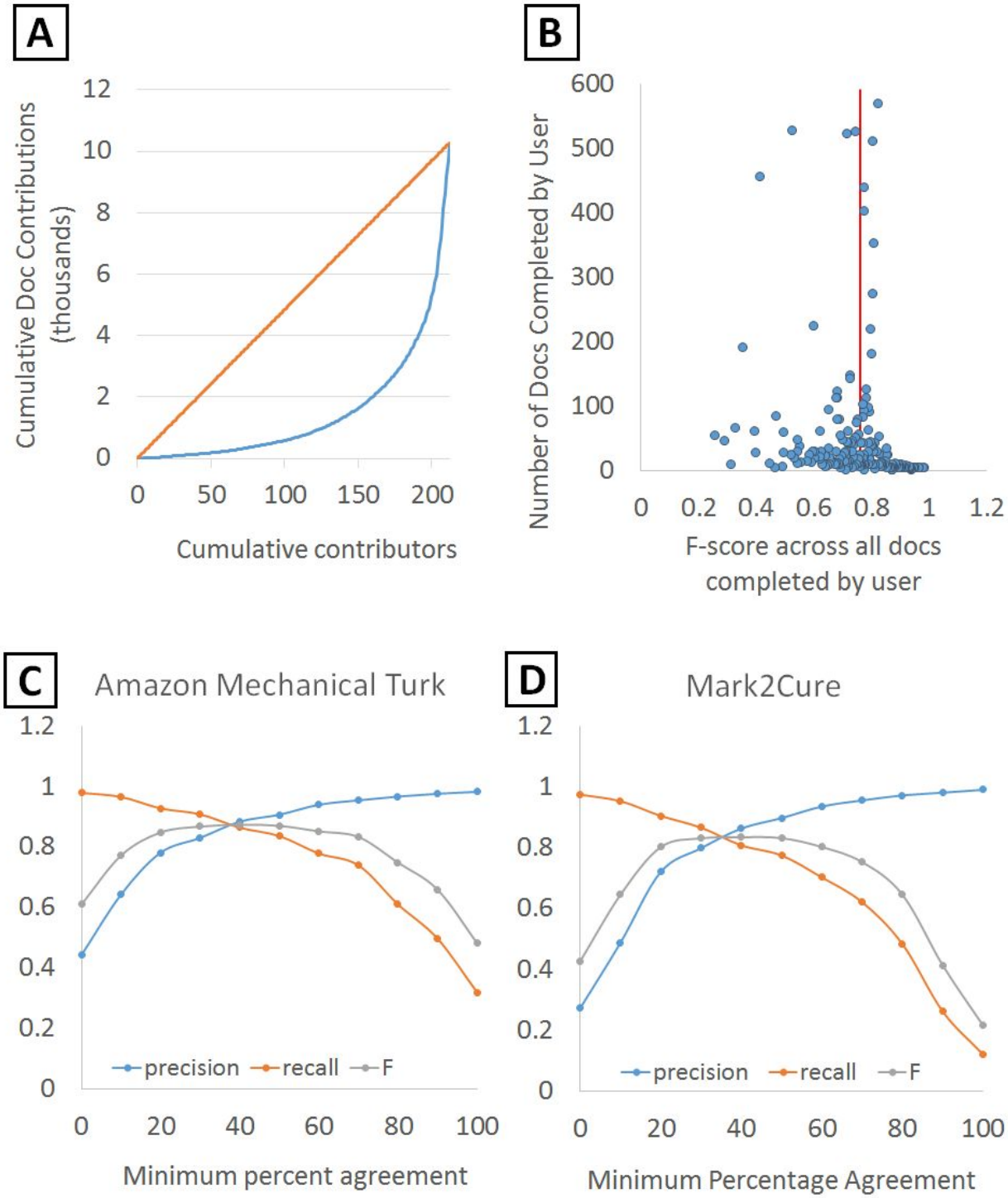


Figure 3: Top: Google Analytics of Mark2Cure's new-user sessions broken down by the source of the sessions. *Total sessions include sessions from both new and returning users. Middle: Timeline of significant events throughout this experiment. Bottom: The number of tasks done on a daily basis (dark blue) along with the cumulative tasks done as a percent of total completion (orange).

Figure 4


A
Cumulative Doc Contributions (thousands)
0
2
4
6
8
10
12
0
50
100
150
200
Cumulative contributors
B
Number of Docs Completed by User
0
100
200
300
400
500
600
0
0.2
0.4
0.6
0.8
1
1.2
F-score across all docs completed by user
C
Amazon Mechanical Turk
0
0.2
0.4
0.6
0.8
1
1.2
precision
recall
F
0
20
40
60
80
100
Minimum percent agreement
D
Mark2Cure
0
0.2
0.4
0.6
0.8
1
1.2
precision
recall
F
0
20
40
60
80
100
Minimum Percentage Agreement

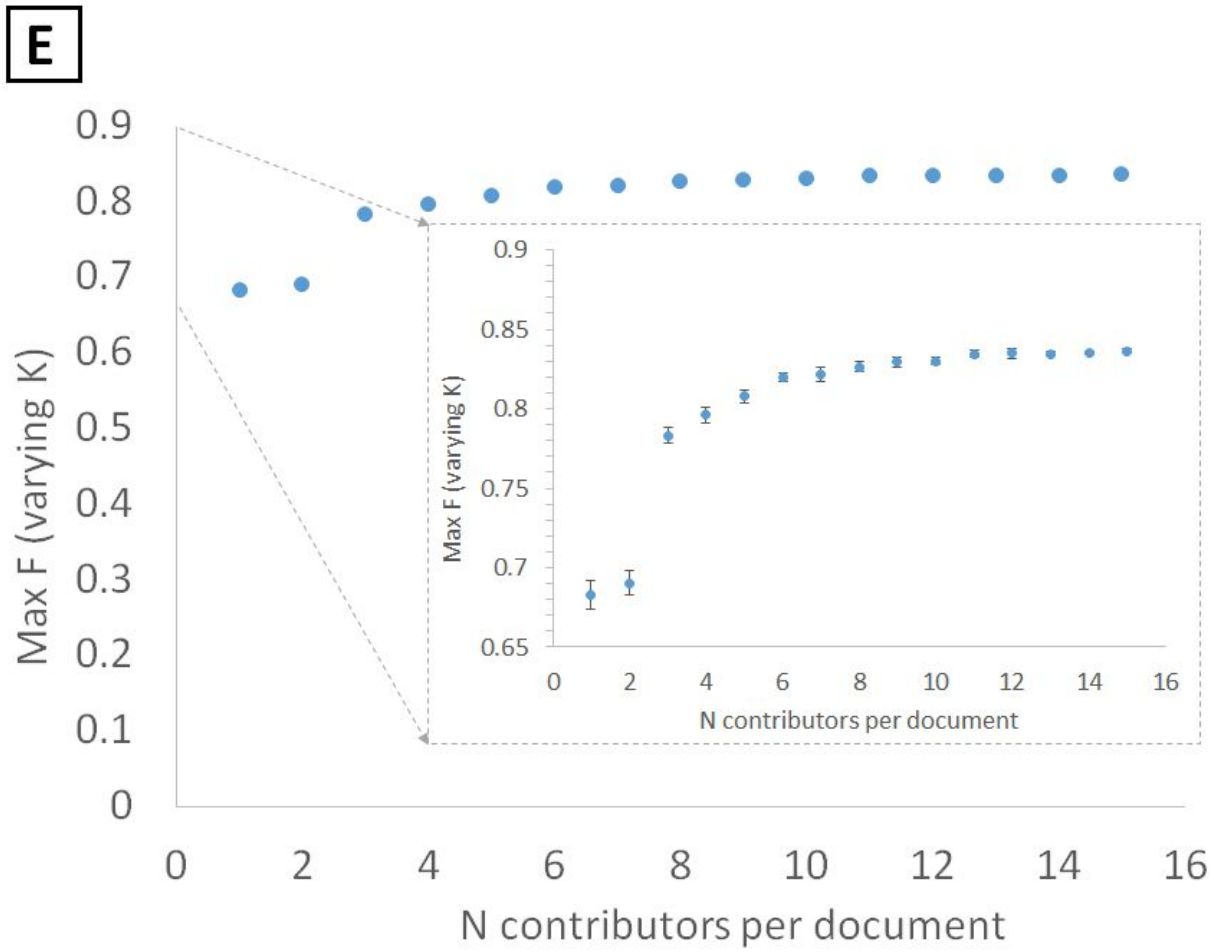
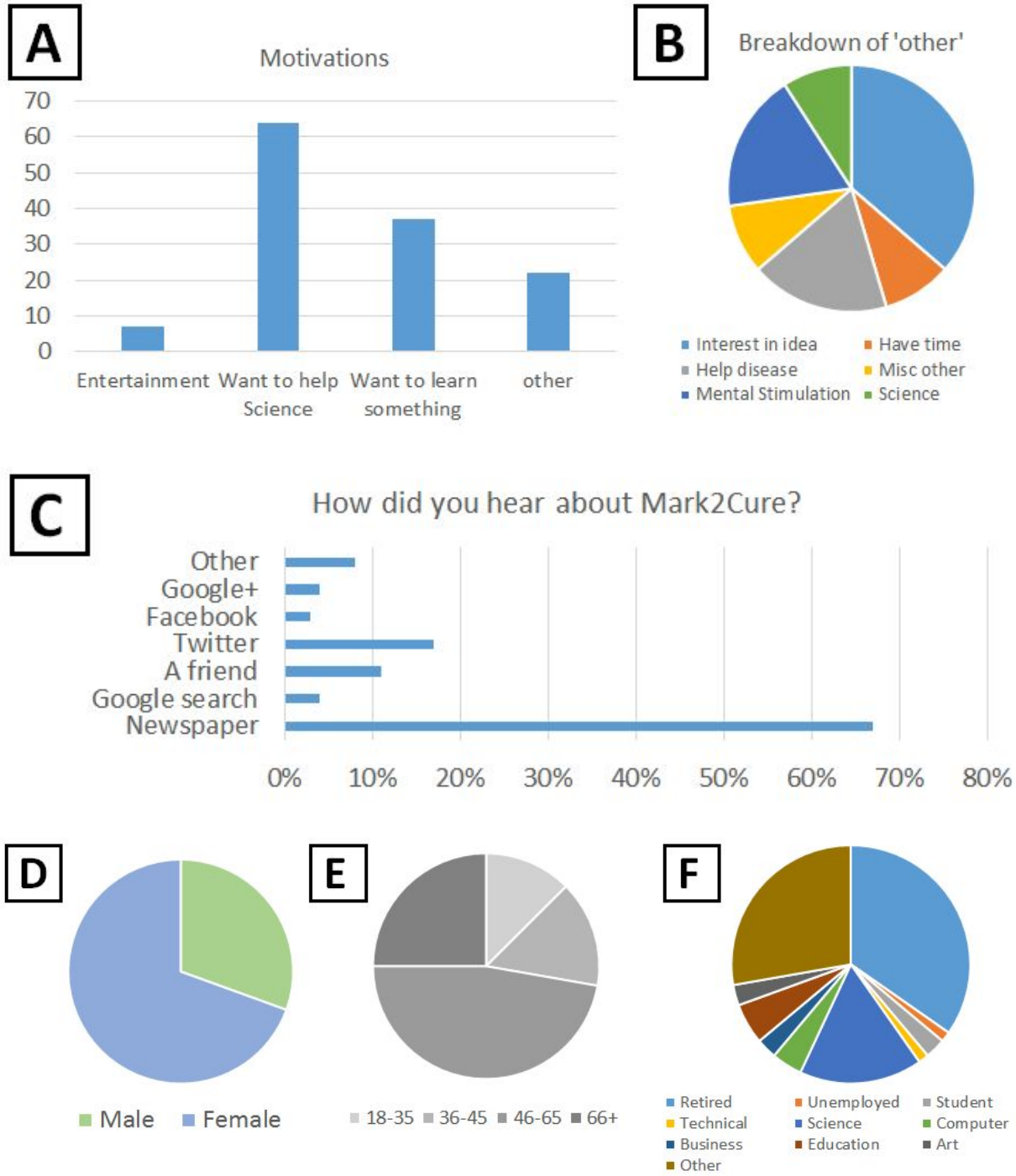


Figure 4 - A. Lorenz chart illustrating cumulative docs contributed relative to the cumulative number of contributors (blue) and expected plot if participants contributed equally (orange). B. Quality of each user's contributions based on the number of docs that user completed. The average F-score per user across all of their contributions (doc completions) is indicated by the red line. C. Effects of the minimum percentage agreement between annotators on the level of agreement with the gold standard in the Amazon Mechanical Turk experiments and in this experiment for Mark2Cure (D). E. The impact of increasing the number of contributors per abstract on the quality of the annotations.

Figure 5


A
Motivations
70
60
50
40
30
20
10
0
Entertainment
Want to help Science
Want to learn something
other
B
Breakdown of 'other'
Interest in idea
Have time
Help disease
Misc other
Mental Stimulation
Science
C
How did you hear about Mark2Cure?
Other
Google+
Facebook
Twitter
A friend
Google search
Newspaper
0%
10%
20%
30%
40%
50%
60%
70%
80%
D
Male
Female
E
18-35
36-45
46-65
66+
F
Retired
Unemployed
Student
Technical
Science
Computer
Business
Education
Art
Other

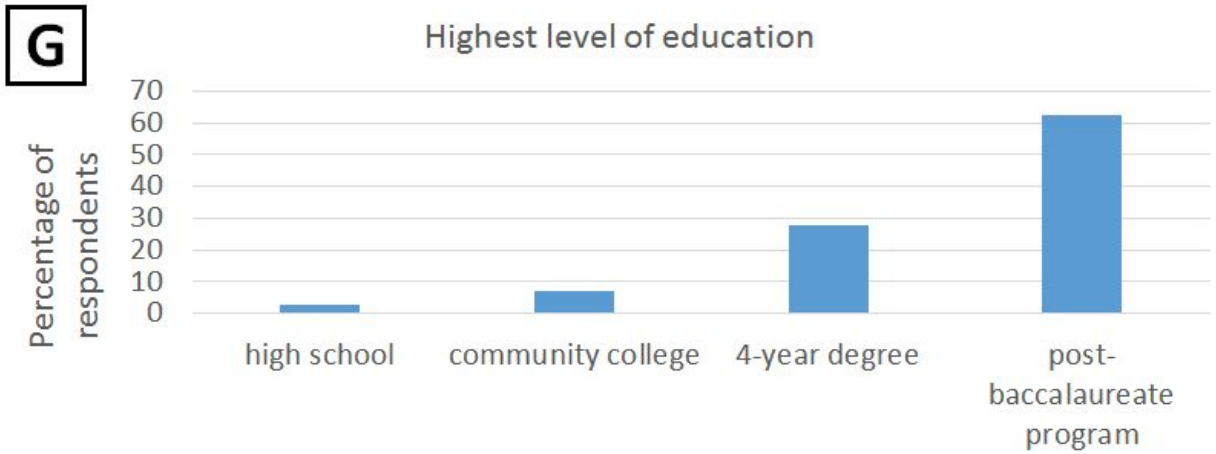


Figure 5 - User survey results. A. Non-exclusive motivations for participating. 89% of the end survey respondents wanted to help science, 51% wanted to learn something, and 10% were looking for entertainment. B. Further analysis of the 'other' motivations for participating. C. How participants discovered Mark2Cure. D. Ratio of female to male survey respondents was 69% to 31% respectively. E. Age demographics of the survey respondents. 28% of respondents were 18-45 years old, and 72% were 46 years of age or older. F. Occupational fields of the survey respondents. In terms of occupations, 35% of respondents were retired, 22% worked in a science, computer or technical field; 28% were care providers, science communicators, or journalists. Only 4% of respondents were students or unemployed, and the remaining 11% of respondents were employed in business, education, or art. G. Educational distribution of survey respondents. 83% of contributors completed a four-year college degree or higher.

Table 1

A. Contribution to Science Matrix			
Performance indicator	Measurement	Proxy	Results
Data value	Publication rate	$\frac{\text{Number of published papers}}{(\text{project age})^2}$	$\frac{0}{(295)^2} = 0$
	Completeness of analysis	$\frac{\text{Number of classifications}}{\text{Target \# of classifications}}$	$\frac{10278}{(586\times15)} = 1.17$
	Academic impact	$\frac{\text{Number of citations}}{(\text{project age})^2}$	$\frac{0}{(295)^2} = 0$
Project design and resource allocation	Resource savings	$1-(\frac{\text{Active project duration}}{(\text{one person workload}})$	$573.60*
	Distribution of effort	$1-(\text{Gini Coefficient})$	1-0.716 = 0.283
	Effective training	$1-(\frac{\text{Tutorial-only completions}}{\text{Total contributions}})$	1-(22/212) = 0.896

B. Public Engagement Matrix			
Performance Indicator	Measurement	Proxy	Results
Dissemination and feedback	Collaboration	$\frac{\text{\# papers with citsci co-authors}}{(\text{project age})^2}$	$\frac{0}{(295)^2} = 0$
	Communication	$\frac{\text{Project communication activity}}{(\text{project active period})^2}$	$\frac{221}{(28)^2} = 0.282$
	Interaction	$\frac{\text{Volunteer\&science team interaction}}{(\text{project active period})^2}$	$\frac{466}{(28)^2} = 0.594$
Participation and opportunities for learning	Project appeal	$\frac{\text{Number of volunteers}}{(\text{project active period})^2}$	$\frac{234}{(28)^2} = 0.298$
	Sustained engagement	$\frac{\text{Median volunteer active period}}{(\text{project active period})^2}$	$\frac{1}{(28)^2} = 0.001$
	Public contribution	$\frac{\text{Median classification/volunteer}}{(\text{project active period})^2}$	$\frac{13}{(28)^2} = 0.017$

Table 1: Computed metrics of “Contributions to Science” and “Public Engagement”, as defined by the Zooniverse projects (Cox et al. 2015). As a single citizen science project, we cannot perform the internal comparisons used in the Zooniverse paper and thus have no basis for comparing our results (Zooniverse results are normalized). Hence, we provide non-normalized results which may serve as a basis for comparison for other research teams who also only have one citizen science project. The resource savings value was not calculated based on traditional workload, as we already had Amazon Mechanical Turk experiments demonstrating the cost to complete the experiment (though the original cost of generating the NCBI Disease Corpus via professional biocuration is not known). In terms of Sustained Engagement, the “median volunteer active period” was based on dates of account creation and time stamp of their last submission. Because many contributors participated just once, the median active period was one day for this experiment, resulting in a sustained engagement of 0.001.

References

Campos, D., Matos, S. and Luis, J. (2012). Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. *Theory and Applications for Advanced Text Mining*. DOI: 10.5772/51066

Cox, J., Oh, E., Simmons, B., Lintott, C., Masters, K., Greenhill, A., Graham, G. and Holmes, K. (2015). Defining and Measuring Success in Online Citizen Science: A Case Study of Zooniverse Projects. *Comput. Sci. Eng.*, 17(4), pp.28-41. DOI: 10.1109/MCSE.2015.65

Doğan, R., Leaman, R. and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47, pp.1-10. DOI: 10.1016/j.jbi.2013.12.006.

Doğan, R., and Lu, Z. (2012). An improved corpus of disease mentions in PubMed citations. *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp.91-99.

Freitag, A. and Pfeffer, M. (2013). Process, Not Product: Investigating Recommendations for Improving Citizen Science “Success”. *PLoS ONE*, 8(5), p.e64079. DOI: 10.1371/journal.pone.0064079

Good, B., Nanis, M., Wu, C., and Su, A.I. (2014). Microtask Crowdsourcing for Disease Mention and Annotation in PubMed Abstracts *Biocomputing 2015*. DOI: 10.1142/9789814644730_0028

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O. and Yon Rhee, S. (2008). Big data: The future of biocuration. *Nature*, 455(7209), pp.47-50. DOI: 10.1038/455047a

Khatib, F., Cooper, S., Tyka, M., Xu, K., Makedon, I., Popovic, Z., Baker, D. and Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47), pp.18949-18953. DOI: 10.1073/pnas.1115898108

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D., Sayle, R., Batista-Navarro, R., Rak, R., Huber, T., Rocktäschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K., Ramanan, S., Nathan, S., Žitnik, S., Bajec, M., Weber, L., Irmer, M., Akhondi, S., Kors, J., Xu, S., An, X., Sikdar, U., Ekbal, A., Yoshioka, M., Dieb, T., Choi, M., Verspoor, K., Khabsa, M., Giles, C., Liu, H., Ravikumar, K., Lamurias, A., Couto, F., Dai, H., Tsai, R., Ata, C., Can, T., Usié, A., Alves, R., Segura-Bedmar, I., Martínez, P., Oyarzabal, J. and Valencia, A. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(Suppl 1), p.S2. DOI: 10.1186/1758-2946-7-S1-S2

Michelucci, P. and Dickinson, J. (2015). The power of crowds. *Science*, 351(6268), pp.32-33. DOI: 10.1126/science.aad6499

Richards, J. and Lintott, C. (2012). *Machine Learning & Supernovae*. [online] Galaxy Zoo. Available at: http://blog.galaxyzoo.org/2012/08/03/machine-learning-supernovae/ [Accessed 26 Jan. 2016].

Swanson, D. (1986). Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1), pp.7-18.

Torii, M., Arighi, C., Li, G., Wang, Q., Wu, C. and Vijay-Shanker, K. (2015). RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information. *IEEE/ACM Trans. Comput. Biol. and Bioinf.*, 12(1), pp.17-29. DOI: 10.1109/TCBB.2014.2372765

Usié, A., Cruz, J., Comas, J., Solsona, F. and Alves, R. (2015). CheNER: a tool for the identification of chemical entities and their classes in biomedical literature. *Journal of Cheminformatics*, 7(Suppl 1), p.S15. DOI: 10.1186/1758-2946-7-S1-S15

Wiggins, A. (2014). *Citizen Science Data Quality is a Design Problem | Andrea Wiggins*. [online] Andreawiggins.com. Available at: http://andreawiggins.com/citizen-science-data-quality-is-a-design-problem/ [Accessed 26 Jan. 2016].

Wiggins, A. and Crowston, K. (2014). Surveying the citizen science landscape. *First Monday*, 20(1). DOI: 10.5210/fm.v20i1.5520

Wiggins, A. and Crowston, K. (2011). From Conservation to Crowdsourcing: A Typology of Citizen Science. *2011 44th Hawaii International Conference on System Sciences*. DOI: 10.1109/HICSS.2011.207