# Machine-Crowd-Expert Model for Increasing User Engagement and Annotation Quality

**Ana Elisa Méndez Méndez**
**Mark Cartwright**
**Juan Pablo Bello**
New York University
New York, NY, USA
anaelisamendez@nyu.edu
mark.cartwright@nyu.edu
jpbello@nyu.edu

## ABSTRACT

Crowdsourcing and active learning have been combined in the past with the goal of reducing annotation costs. However, two issues arise with using AL and crowdsourcing: quality of the labels and user engagement. In this work, we propose a novel machine ↔ crowd ↔ expert loop model where the forward connections of the loop aim to improve the quality of the labels and the backward connections aim to increase user engagement. In addition, we propose a research agenda for evaluating the model.

## KEYWORDS

Crowdsourcing; active learning; sound classification



*CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland Uk*



ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3313054

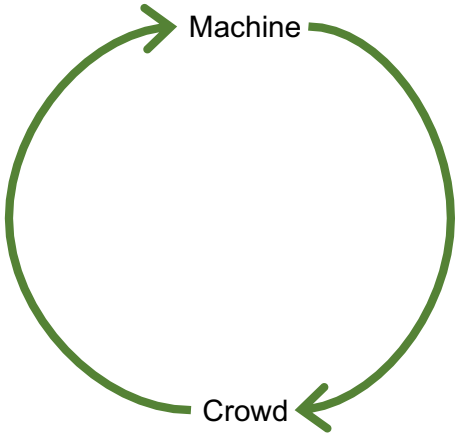


**Figure 1: Active learning loop.**

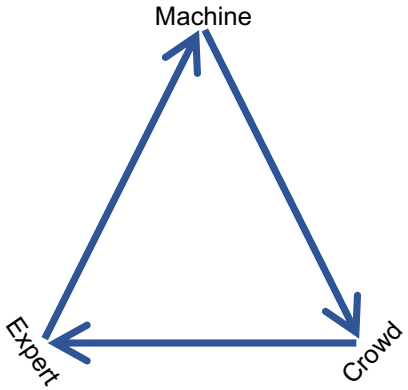


**Figure 2: Active learning with experts in the loop.**

## INTRODUCTION

The collection and annotation of data are essential components of machine learning that dictate its success. However, modern machine learning models are data hungry, and it is non-trivial and costly to collect the large datasets that have transformed fields of research, requiring thousands of annotators. To support more varied and niche applications, we need methods for high-quality annotation that are more efficient to reduce costs, and more engaging to motivate and retain annotators. In addition, datasets are ideally annotated by domain experts, but this approach is costly and time consuming due to the scarcity and high-value of experts.

One method for increasing annotation efficiency is by parallelizing annotation with crowdsourcing. Crowdsourcing is a tool for distributing tasks through open calls [10], e.g. on paid platforms like Amazon Mechanical Turk or citizen science platforms like Zooniverse. Snow et al. [15] showed that crowdsourcing annotation tasks can achieve similar results as domain experts, and in the past decade, crowdsourcing has been used for many different annotation tasks and has driven much of the recent success in machine learning [1]. However, for tasks requiring specific domain knowledge, experts add important value to the annotations and in consequence to model performance [5, 9]. Another method for increasing annotation efficiency is to collect fewer, more informative annotations using active learning (AL) [14]. AL is a machine learning method where the machine itself chooses the data from which it will learn, from an unlabeled data pool, based on different sampling strategies [14]. In AL only a small amount of labeled data is needed to train an initial model. This model is used to make predictions on an unlabeled data pool and select the most uncertain examples for labeling. These examples are labeled by an "oracle" (e.g., a human annotator) and then added to the labeled data pool for retraining the model (see Figure 1).

AL and crowdsourcing are not exclusive to each other and can be combined for greater efficiency [1, 7], but two issues arise in both crowdsourcing and AL: annotation quality [16] and user engagement [2]. Annotation quality is often addressed by aggregating multiple novice annotations with different strategies [1, 7, 16]. Most recently, it was approached by Callaghan et al. [5] by incorporating experts into the AL loop (see Figure 2). In their framework, crowdworkers are initially queried to provide a label by majority vote, but if the percentage of agreement is below a threshold, a domain expert is then queried to provide an overriding label that is added to the labeled data pool.

User engagement must be addressed by studying the annotators, not just their responses, and Amershi et al. [2] stress the importance of studying the users of interactive machine learning systems such as AL. An important attribute of user engagement is task interactivity [11]. Cakmak et al. [4] show that tasks are tedious for users when users are asked to provide only binary responses. In their work, they tested various passive and active learning approaches and found participants preferred an interactive method where the users could also query the learner, while also had comparable model

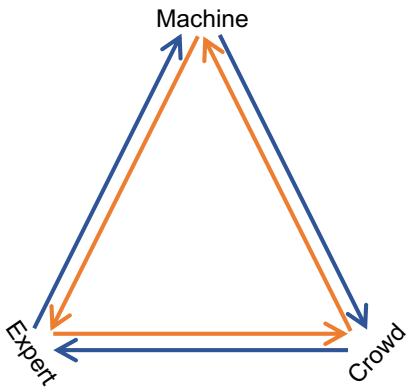


**Figure 3: Proposed model. The blue arrows represent the forward connections and the orange arrows represent the backward connections.**

performance. Dow et al. [8] show that incorporating learning into crowdsourcing can also improve user engagement and performance.

In this work, we propose a novel annotation model that combines domain experts, crowdsourced novice annotators, and active learning with the goal of improving both annotation quality and annotator engagement. Additionally, we propose a research agenda to evaluate the model. While we present the model in the context of annotation for sound event detection, we believe the framework of this model could be used in a variety of machine learning applications that require domain knowledge for annotation.

## PROPOSED MODEL

Our proposed annotation model combines domain expert annotators, novice crowdsourced annotators, and active learning in a way that seeks to improve both annotation quality and annotator engagement. Expert annotators are included to improve quality and task definition; and novice crowdsourced annotators and active learning are included for efficiency. The model consists of three vertices: *machine*, *crowd*, and *expert*; which are connected together in a complete, directed (i.e., K3) graph and can be decomposed into a *forward loop* (machine → crowd → experts) and a *backward loop* (machine → expert → crowd). The goal of the forward loop is to increase label quality and model performance, while the goal of the backward loop is to increase user engagement through interactivity and learning, which we hope will indirectly, also lead to better-quality labels and consequently better models.

### Forward connections

*Machine → crowd.* In this connection the system queries the crowd asking for labels. These queries can be requested using different sampling strategies (e.g., uncertainty sampling), depending on the needs of the system. For example, the machine may ask users to identify the presence of construction equipment in a sound recording for which the AL model is uncertain. This class might be hard to disambiguate from other busy urban scenes where loud engines and impact sounds are also common.

*Crowd → expert.* The expert receives from the crowd those examples with an uncertainty above a certain threshold, similar to [5]. It is important to note that hiring experts has a higher cost than hiring crowdworkers, which is why we need to set up a threshold that is not only cost efficient but also effective in capturing uncertainties so they can be resolved and added to the model. Continuing with the previous example, if agreement between users on the presence of construction equipment in a recording is lower than a set threshold, this recording will be passed to the expert for labeling.

*Expert → machine.* When experts receive the queries from the crowd, they verify the examples and label them accordingly. Continuing with the example, the expert would listen to the recording to search for a construction equipment and assign the label that will be used as ground truth for training.

**Backward connections**

*Crowd → machine.* In the crowd-machine connection, the crowd queries the system. Studies have shown this to be helpful in user engagement [4], but it is not common practice in AL. For example, a user could have access to a pool of pieces (e.g., sound recordings of urban sounds and construction sounds) that they can construct into an example whose label (e.g., "construction equipment present") is predicted by the machine—probing the limits of the machine's understanding. The user-labeled examples would be added to a separate pool of synthesized data for training the machine.

*Expert → crowd.* This connection is one of the key components in our proposal, as it is novel in AL. Pan, Larson and Law [12] and Dow et al. [8] demonstrated that users learn from receiving feedback and [8] also proved that it motivates users and improves their production. We propose that once experts are required, they also provide feedback to novices. Besides improving motivation and engagement, we aim to reduce costs as well, since having more knowledgeable workers will decrease the amount of times the expert is needed. To continue with our example, when experts identify the presence/absence of construction equipment in the recordings, users receive feedback on their performance and can revisit the recordings to understand and learn from their mistakes.

*Machine → expert.* In the last part of the loop, the experts receive from the machine sets of examples that are clustered together using semi-supervised learning. Experts analyze these examples and determine if those examples correspond together or not. This also allows for the possibility of creating new sets of labels that might correspond better to what is present in the examples. Although studies have previously worked on generating new taxonomies [3, 6], it is not common practice in crowdsourcing, and its incorporation in AL is novel. In the construction equipment example, it could be that the machine identifies two or three different clusters within construction equipment and suggests that these should be separated. Experts may listen to these examples and determine that the category should be divided into jackhammers, hoe rams and pile drivers, generating more specific sound categories.

**Challenges**

Experts' time is costly and scarce. Therefore, we need to set a threshold that minimizes queries to experts without sacrificing performance—a balance between cost/efficiency and quality. Another challenge is the synchrony of tasks and feedback. Schaekermann et al. [13] found that synchronous communication between users is problematic because it requires users to log in at the same time and in multiple stages, which requires workers to be motivated and incentivized to come back to continue with the task. In our case, synchronous feedback from experts to crowdworkers would require experts to be available to provide feedback as users submit their responses. Although this is not ideal, since

(1) The machine will query the crowd based on the AL least confident sampling strategy, presenting the users batches of 10 recordings at a time.
(2) The crowd will make annotations on the recordings and these are aggregated to determine the certainty probability. Based on a fixed threshold, some recordings will be considered as having high certainty and those will be feedback to the machine, and those with low certainty are passed to the experts for further annotation.
(3) Ground truth annotations, from the synthesized dataset, will be used as experts to submit annotations to the machine when needed.

**Sidebar 1: Forward loop**

experts are often not easily available, it would not require users to come back to perform the next set of annotations. On the other hand, if expert feedback is provided asynchronously, users would have to log back in to find their feedback and continue with the task, which would require good motivation and incentivization.

## FUTURE WORK

The research agenda will consist of three separate initial experiments for integrating the backward connections with the forward loop (machine → crowd → expert) in the context of audio annotation for sound event detection. In the cases where it is possible, we will use synthetic data with ground truth annotations to provide the correct feedback and be able to evaluate the quality of the model and annotations. See Sidebar 1 for a description of the forward loop.

*Integration of expert → crowd.* When experts are required, the ground truth annotations will be used instead of experts to give feedback to the crowd. We will test two different participant-specific feedback scenarios: (1) where feedback is reported synchronously as responses from the crowd come in, (2) where feedback is reported after a waiting period. We will also experiment with feedback that is not participant-specific but general to all annotators.

*Integration of crowd → machine.* In this part of the experiment, crowdworkers have the option to decide to query the machine with synthesized examples. The participants will have access to a pool of sounds that they can mix together. After they have a complete example, they ask the machine to make a prediction. If the example is correct, then the participant has the option of building a new example or continuing with labeling.

*Integration of machine → expert.* When testing the integration of the machine → expert portion, we will not use synthesized data as ground truth, since creating a new taxonomy requires a human to listen to the examples and create the new labels. After the expert has done the first iteration of labeling the examples from the crowd, they can decide to query the machine for clusters. Once the machine has grouped the recordings into these clusters, the expert listens to the examples and decide if there are new categories of sound that can be added to the existing taxonomy.

Our goal is to show how this new framework affects performance, by measuring annotation quality, and user engagement, by applying surveys and looking at the number of tasks completed and the number of annotation sessions performed. Once each of these components has been experimented with and refined, we will experiment with the whole model. By doing an ablation study, we will turn off feedback connections to understand how each of them affects performance in the whole model.

## ACKNOWLEDGMENTS

This work was supported by National Science Foundation award 1544753.

## REFERENCES

[1] Vamshi Ambati, Stephan Vogel, and Jaime G Carbonell. 2010. Active Learning and Crowd-Sourcing for Machine Translation.. In *Proceedings of the International Conference on Language Resources and Evaluation*, Vol. 1. 17–23.

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120.

[3] Jonathan Bragg, Daniel S Weld, et al. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *Proceedings of the First AAAI conference on human computation and crowdsourcing*. 25–33.

[4] M Cakmak, C Chao, and A L Thomaz. 2010. Designing Interactions for Robot Active Learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (2010), 108–118. https://doi.org/10.1109/TAMD.2010.2051030

[5] William Callaghan, Joslin Goh, Michael Mohareb, Andrew Lim, and Edith Law. 2018. MechanicalHeart: A Human-Machine Framework for the Classification of Phonocardiograms. *In Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 28:1–28:17. https://doi.org/10.1145/3274297

[6] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008.

[7] Joana Costa, Catarina Silva, Mário Antunes, and Bernardete Ribeiro. 2011. On using crowdsourcing and active learning to improve classification performance. In *Proceedings of the 11th International Conference on Intelligent Systems Design and Applications*. 469–474.

[8] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[9] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. 2012. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 871–880.

[10] Edith Law and Luis von Ahn. 2011. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 3 (2011), 1–121.

[11] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.

[12] Shengying Pan, Kate Larson, Josh Bradshaw, and Edith Law. 2016. Dynamic Task Allocation Algorithm for Hiring Workers that Learn.. In *Proceedings of the International Joint Conferences on Artificial Intelligence*. 3825–3831.

[13] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *In Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 154:1–154:19.

[14] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

[15] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.

[16] Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems* 159 (2018), 244 – 258. https://doi.org/10.1016/j.knosys.2018.07.010