Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists

Mark Cartwright
New York University
New York, USA
mark.cartwright@nyu.edu

Graham Dove
New York University
New York, USA
grahamdove@nyu.edu

Ana Elisa Méndez Méndez
New York University
New York, USA
anaelisamendez@nyu.edu

Juan P. Bello
New York University
New York, USA
jpbello@nyu.edu

Oded Nov
New York University
New York, USA
onov@nyu.edu

ABSTRACT

Annotating rich audio data is an essential aspect of training and evaluating machine listening systems. We approach this task in the context of temporally-complex urban soundscapes, which require multiple labels to identify overlapping sound sources. Typically this work is crowdsourced, and previous studies have shown that workers can quickly label audio with binary annotation for single classes. However, this approach can be difficult to scale when multiple passes with different focus classes are required to annotate data with multiple labels. In citizen science, where tasks are often image-based, annotation efforts typically label multiple classes simultaneously in a single pass. This paper describes our data collection on the Zooniverse citizen science platform, comparing the efficiencies of different audio annotation strategies. We compared multiple-pass binary annotation, single-pass multi-label annotation, and a hybrid approach: hierarchical multi-pass multi-label annotation. We discuss our findings, which support using multi-label annotation, with reference to volunteer citizen scientists' motivations.

CCS CONCEPTS

• **Human-centered computing** → **Computer supported cooperative work**; **Empirical studies in collaborative and social computing**; • **Information systems** → *Speech / audio search*; • **Applied computing** → *Sound and music computing*.

KEYWORDS

Audio annotation, citizen science, crowdsourcing

ACM Reference Format:
Mark Cartwright, Graham Dove, Ana Elisa Méndez Méndez, Juan P. Bello, and Oded Nov. 2019. Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300522

1 INTRODUCTION

Annotating rich audio data is an essential aspect of training and evaluating machine listening models, which have the potential to enable powerful applications in diverse domains such as bioacoustic monitoring, urban noise monitoring, electric vehicle sensing, assistive technologies, and more. This is a difficult and time consuming task, due in part to audio's temporal dimension. We approach this problem from the context of a New York City-based project that is working closely with city agencies such as the Department of Environmental Protection; and which aims to monitor, analyze, and mitigate urban noise pollution using a smart sensor network powered by machine listening models that detect noise sources [5]. Noise pollution is a major concern to many urban residents and has many negative effects, e.g. on citizens' health [4, 21] and students' learning [4]. Because of these societal concerns, we are exploring citizen science-based inquiry, including our approach to training machine listening models, and a key task is to establish how best to design tasks to acquire multi-label audio annotations with volunteers. Since our aim is for high ecological validity, this study was not conducted as a controlled experiment, but rather we undertake an analysis of a real-world data collection solving a real-world audio annotation problem. This activity took place on Zooniverse, the most widely used citizen science platform, and as such,

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300522

our focus is on the authentic behaviors of volunteering citizen scientists, which may differ from those of paid workers on commercial platforms. Best practice for crowdsourced audio annotation remains understudied, with prior research typically focusing on paid crowdworkers rather than volunteer citizen scientists. In this paper we make some initial steps towards addressing this area of concern.

Previous research recommends breaking complex tasks into small units of work for paid crowdsourcing [31]. In large-scale audio annotation efforts, this has resulted in paid crowdworkers performing single-class binary-labeling tasks [20, 22]. But citizen scientists and crowdworkers have different motivations, and in online citizen science projects, where audio annotation is rare, image annotation typically adopts a multi-labeling approach [34, 51]. However, like video annotation, annotating audio is more complex than annotating images due to the addition of a temporal dimension; and while full multi-label annotations for N classes can be constructed from N binary class annotations, this does not scale well. Should researchers follow the norms of citizen science image annotation (multi-label) when collecting multi-label audio annotations with volunteers? Or those of crowdsourced audio annotation with paid crowdworkers (multiple binary-label)? What are the effects on annotation throughput and quality of adopting these contrasting approaches? Two previous studies point in contrasting directions. In [22], a pilot study indicated that crowdworkers found a multi-label audio annotation task difficult, were unhappy, and had low annotator agreement. However, in [50], a study on video annotation with crowdworkers found that multi-label annotation tasks resulted in higher quality annotations than those from binary annotation tasks.

This paper contributes to the literature on best practices for crowdsourced audio annotation, and seeks to answer these questions, by comparing different annotation task types on the Zooniverse online citizen science platform [1]. We compare multiple-pass binary annotation, single-pass multi-label annotation, and a hybrid approach: hierarchical multi-pass multi-label annotation; in work from the early stages of our urban soundscape annotation campaign. We present an analysis of the throughput of the three different annotation task types and a sensitivity analysis on the effect of aggregation variables on annotation quality. Our results suggest that practices developed for crowdsourced populations may not translate to volunteer-based populations.

2 RELATED WORK

In the past decade, there have been several large-scale, multi-label and multi-class paid-crowdsourcing efforts to annotate media for training machine learning models, e.g. ImageNet (14M images, 20k classes) [15], COCO (328k images, 91 classes)[33], Places (10M images, 434 classes) [56], AudioSet (1.8M audio recordings, 636 classes) [20], and OpenMIC-2018 (20k audio recordings, 20 classes) [22]. While each has taken a slightly different approach to suit their media, they all use weak search engines [15, 20, 33, 56] or classifiers [22] to generate candidate sets which are then verified by humans with binary [15, 20, 22, 56], limited multi-label [20], or hierarchical multi-label [33] annotation. Of these efforts, AudioSet [20] and OpenMIC-2018 [22] are the only audio annotation efforts. It is also worth noting that AudioSet annotators were exposed to both audio and accompanying video during annotation because they found it too difficult with audio alone. COCO [33] is the only dataset with full multi-label annotations, and required 70k worker hours to complete image segmentation and labeling.

Because of the immense human effort required, researchers have sought to optimize multi-class and multi-label crowdsourced image annotation [7, 12, 13, 16, 32], often with approaches that break larger multi-label tasks into small subtasks. In audio annotation, studies have typically focused on the effects of sound visualizations rather than trade-offs between binary and multi-label annotation when investigating rapid methods [53] and best practices [10]. However, a small-scale study using data labeled by experts rather than through crowdsourcing indicates that sound event detection accuracy is higher when using a single model trained on full multi-label data rather than aggregating multiple single-class models trained on binary data [8]. For video annotation, which like audio has the additional complexity of a temporal dimension, research has indicated both that breaking annotation into smaller subtasks is wasteful [50], and also conversely that performance and annotator satisfaction increase when annotating one object rather than multiple objects, (although error accumulated in sequences of micro-tasks largely negated this finding) [54].

All of the above annotation efforts and studies were performed with paid crowdworkers. While intrinsic motivations can affect the quantity [11] and quality [46] of paid crowdwork, the primary motivation of paid crowdworkers is still payment and therefore extrinsic [30]. It's therefore unclear if the findings of studies using paid crowdsourced annotation will translate to annotation with citizen scientists, who are not motivated by payment. Viewing the problem from the opposite direction, [36] show that while paid crowdworkers can complete labeling tasks with comparable quality to volunteer citizen scientists, they are also highly sensitive to changes in payment method.

Zooniverse [1], the largest online citizen science platform, has hosted several large image annotation projects, most notably Galaxy Zoo (900k images, 6 classes) [34] and Snapshot Serengeti (1.2M image sets, 48 animal presence classes) [51], but audio annotation projects on the platform are rarer and smaller in size [35, 48]. On the Zooniverse platform,

image-based projects typically use full multi-label annotation. Although the initial Galaxy Zoo project was limited to six classes, there are now several projects inspired by Snapshot Serengeti that have on the order of 50 classes [1].

Unlike paid crowdsourcing, citizen science relies on voluntary efforts of contributors. Volunteers typically express multiple motivations [42–44], which change dynamically through a project's temporal shifts [47]. Typically, online citizen science projects have skewed participation patterns, with a small number of highly motivated volunteers contributing the majority of work [19]. The quantity of online volunteer contributions correlates with collective-, norm-oriented-, reputation-, and intrinsic- motivation; while contribution quality responds positively to collective- and reputation-motives [40]. Volunteers' individual contributions may be increased through highlighting instances of novelty [25]. The importance of intrinsic motivations highlights a need for engaging participation mechanisms, and association with a project's aims helps motivate new volunteers [39]. Where volunteers are drawn from a particular community, motivations for engagement in citizen science closely reflect motivations for community membership of that community and its core activities [55].

The desire to learn about a particular topic is an important motivator in both online [42, 43] and in-real-life [18] citizen science, as is the desire to contribute to "authentic" scientific research [38]. In addition to knowledge acquisition, sharing new knowledge can motivate and sustain engagement [26]. Other methods of motivating sustained engagement include competition amongst volunteers and gamification of citizen science activities [6, 24, 41]. Such regular engagement with a project builds community membership, and leads to task aptitude and familiarity; while task difficulty and boredom, and competing priorities are barriers to contribution [28]. Other factors that impact on volunteers' engagement, both positively and negatively, include a project's coordination practices and the volunteer's previous domain experience [17]. In addition, scientists' concerns about data quality and interpretation, and wider ethical concerns about visibility, authorship and attribution [14, 45], can lead to tension between the motivations and epistemic contributions of citizen science participants as individuals and their status within a collective distributed effort [29]. We build on this literature by investigating best practices for crowdsourced audio annotation with volunteer citizen scientists.

3 DATA COLLECTION

Platform

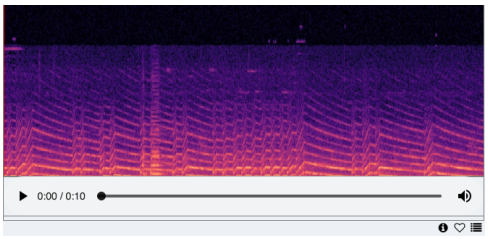
Zooniverse is the largest platform for online citizen science projects [1], offering a ready-made community of motivated volunteers from which to recruit. While this facilitates our aim of studying the authentic behaviors of volunteering citizen scientists, the Zooniverse platform is designed for real-world data collection and not for controlled experiments, and researchers therefore have limited facility for controlling variations over the presentation of their tasks. Also, during the design stage of this research, we were in close contact with expert Zooniverse moderators who provided informed advice, suggesting practical compromises between ensuring that primary data were collected and enriching these primary annotation data with user-focused data. For example, this expert advice included removing links to external questionnaires in order to "lower the barrier to entry as much as possible". However as an alternative, the Zooniverse platform does enable researchers to view more qualitative participant responses through its 'Talk' boards. Suggestions such as these highlight tensions between eliciting real-world annotation with citizen science volunteers and gathering rich participant data. Because of factors such as this, our comparison of multiple annotation task types in the early stages of our data collection should be considered more akin to A/B testing than strictly controlled experimentation.

Audio data

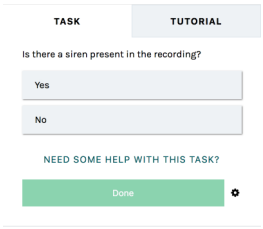
Our audio data consist of 10 second clips from two distinct sources, and represent 22 classes of sound sources. The first dataset consists of audio taken from selected YouTube video clips, where the sound source was identified, and additionally confirmed by members of our research team. These provided ground truth for our analysis. The second were recordings selected from amongst the 30 years' worth of audio data collected by 50+ sensors installed in busy locations around New York City [5]. The 22 classes of sound sources that make up our labeling taxonomy are derived from requirements requested by the city's department of environmental protection, and based on its legally enforceable noise code. They include examples of engines of different sizes, construction machinery and tools, human and animal vocalizations, music, and vehicle alert signals and sirens. We selected 2 examples for each class from the YouTube dataset, and 3 examples from the dataset of sensor recordings, creating a total of 110 10 second audio clips. The 3 examples selected from the sensor data were curated by members of our research team from an initial sample of 30 examples automatically selected for each class using "VGGish" embedding-derived audio features [27].

Tasks

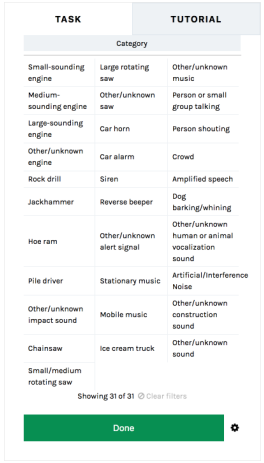
Volunteers were randomly presented with 1 of 3 task types: 1) *binary-labeling*, 2) *one-stage multi-labeling*, and 3) *two-stage hierarchical multi-labeling*. All volunteers were presented with the same tutorial and field guide to familiarize themselves with examples of each sound-source class. In all three



(a) Accompanying sound visualization



(b) Binary



(c) Multi-label

Figure 1: Screenshots of the binary and multi-label annotation tasks on Zooniverse along with the spectrogram sound visualization shown to annotators.

task types, the audio was presented both aurally and visually (using a spectrogram representation; see Figure 1a).

In the *binary labeling* task (see Figure 1b), volunteers were asked to decide whether a single suggested sound-source class was present or not in the recording. This task type provided both positive and negative labels explicitly.

In the *one-stage multi-labeling* task (see Figure 1c), volunteers were presented with a list of 30 class labels and an audio clip, and were asked to select all the sound-source classes present in the audio. The list of label options included our 22 sound-source classes plus labels for unknown or uncertain examples of the "superclasses"; e.g. *engines*, *construction machinery*, or *alert signals*. This task type provided positive labels explicitly and negatively labels implicitly. Previous studies [50] indicate that requesting explicit negative labels reduces both precision and recall, and increases task completion time in a multi-label task.

In the *two-stage hierarchical multi-label* task, each stage was undertaken separately and by a different volunteer. In stage 1, the audio was presented to a volunteer alongside a list of 9 superclass labels; e.g. *engines*, or *powered sawing tools*. Identification of sounds in this stage provided a filter for possible class labels. For each selected superclass in stage 1, we posted a stage 2 task in which the same audio clip was presented alongside the sublist of our 22 class labels that correspond to the superclass. For example, if the audio had been identified as containing engine sounds in stage 1, the list of possible labels shown in stage 2 would include: *large-sounding engine*, *medium-sounding engine*, *small-sounding engine*, *other/unknown engine*, *artificial interference noise*, and *other/unknown sound*. Stage 2 tasks were undertaken at a later date, and the audio was presented to a different volunteer. This task type provided positive labels explicitly and negative labels implicitly. We included this task as a compromise between the multi-labeling and binary labeling tasks — it requires fewer sound-source classes to be annotated simultaneously and possibly fewer sound-source classes to be annotated overall for full multi-label annotation.

Task Type	Unanimous Agreement Pct.
Binary	81%
Multi-label	91%
Hrchl. Multi-label Stg 1	79%
Hrchl. Multi-label Stg 2	65%
Task Type	**Krippendorff's α (95% CI)**
Binary	0.52 [0.46, 0.58]
Multi-label	0.53 [0.44, 0.60]
Hrchl. Multi-label Stg 1	0.45 [0.37, 0.52]
Hrchl. Multi-label Stg 2	0.45 [0.35, 0.54]

Table 1: Annotator agreement.

4 ANALYSIS

We collected at least 5 annotations per recording for the multi-label task and both stages of the hierarchical multi-label task and at least 3 annotations per recording for the binary annotation task, from a total of 339 unique volunteers. We limited our analysis to these minimums. It was impractical to collect more than 3 annotations for the binary task, due to it's poor scalability. The annotator agreement for all three task types is presented in Table 1. We also downloaded the comments and messages that volunteer citizen scientists left on the Zooniverse 'Talk' boards as a way of approaching user-focused data. Because there were too few of these messages to undertake a detailed qualitative analysis, we have instead included example comments in our Discussion, as a way to illustrate particular points regarding Zooniverse volunteers' responses to the different annotation approaches, and not as study findings in their own right.

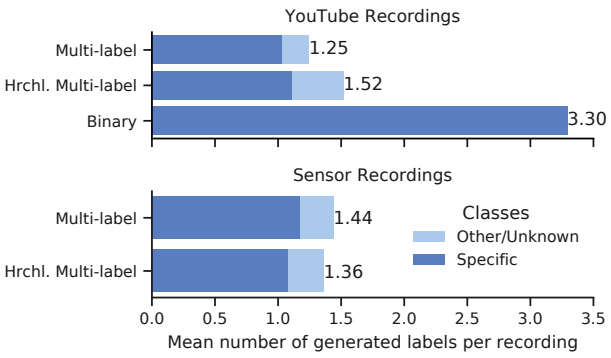


Figure 2: The mean number of positive labels generated in a full multi-label annotation. *Other/Unknown* indicates the catch-all classes in the multi-label annotation tasks.

Task Type	22-class Ann. Time (s) (95% CI)
Binary	308.64 [301.01, 316.09]
Multi-label	33.54 [31.48, 35.58]
Hrchl. Multi-label	50.66 [47.52, 53.88]

Table 2: Mean time (in s) to complete a full 22-class annotation for each type of annotation task.

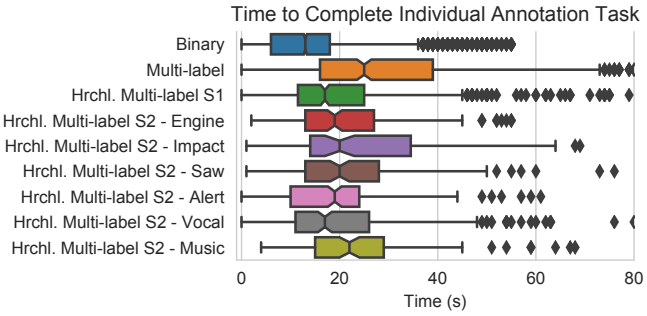


Figure 3: The time to complete individual annotation task for each annotation task type. *S1* and *S2* indicate stages 1 and 2 respectively.

Annotation throughput

We define the throughput of an annotation task as the rate of label generation. This is a function of the number of volunteers, the quantity of labels that they generate in a single task, the speed at which they complete a task, and the number of annotation tasks they are motivated to complete. Maximizing this measure helps collect data quickly and progress research. It also respects the time volunteer annotators freely give to the project as productive contributors. Our analysis focuses on the quantity of positive labels generated and the speed of task completion in response to the annotation task. To calculate task completion times, we computed the time difference between annotation tasks submitted by the same volunteer and removed outliers in the top 5th percentile, which may represent different annotation sessions.

The binary annotation task generated over twice as many positive labels per full multi-label annotation as the multi-label annotation tasks, (see Figure 2). In addition, as shown in Figure 3, it took about twice as long to complete an individual multi-label annotation task (32.81 s, 95% CI [30.80, 34.86]) as it did an individual binary annotation task (14.06 s, 95% CI [13.74, 14.38]). When scaled up for 22 classes, it took more than 9 times as long to annotate a full multi-label annotation using binary labeling tasks (see Table 2). Therefore, if full annotations are needed, multi-labeling tasks have higher throughput; but if only binary annotations are needed, the two task types have comparable throughput.

Annotation quality

In addition to maximizing throughput, we also want high quality labels. To measure quality, we calculated the precision, recall, and F-measure on aggregated annotations of the YouTube recordings, since these have positive ground-truth labels. However, we also need negative ground-truth labels to compute our metrics. These we obtained by labeling a limited amount of data that we had high confidence about ourselves. Our ground-truth contained one positive and one negative label for each of the 44 YouTube recording and was balanced to have an equal number of positives and negatives for each class.

With this ground-truth data, we varied both the number of annotators and the voting threshold required for a positive label, and measured their effects on annotation quality. To account for the many possible combinations of annotators, we estimated the true positives, false positives, and false negatives using a sample of 1000 random combinations of annotators for each task type / recording pair. For example, in one sample of the multi-labeling task aggregated with three annotators and a voting threshold of two, we randomly chose three of the five annotations and labeled a class as positive if at least two of the annotators labeled it positive. We then calculated the true positives, false positives, and false negatives using our ground truth data, and repeated the process 1000 times. For hierarchical multi-labeling, we performed a similar process but aggregated annotations at both stages, with the output of stage one informing the inclusion of subtasks in stage two. The annotations from the subtasks were combined together to form a full multi-label annotation. For binary labeling, we aggregated binary annotations for each class and then combined all 22 class annotations together to form a full multi-label annotation. We then summed the true positives, false positives, and false negatives over the recordings to compute the quality metrics for each combination of

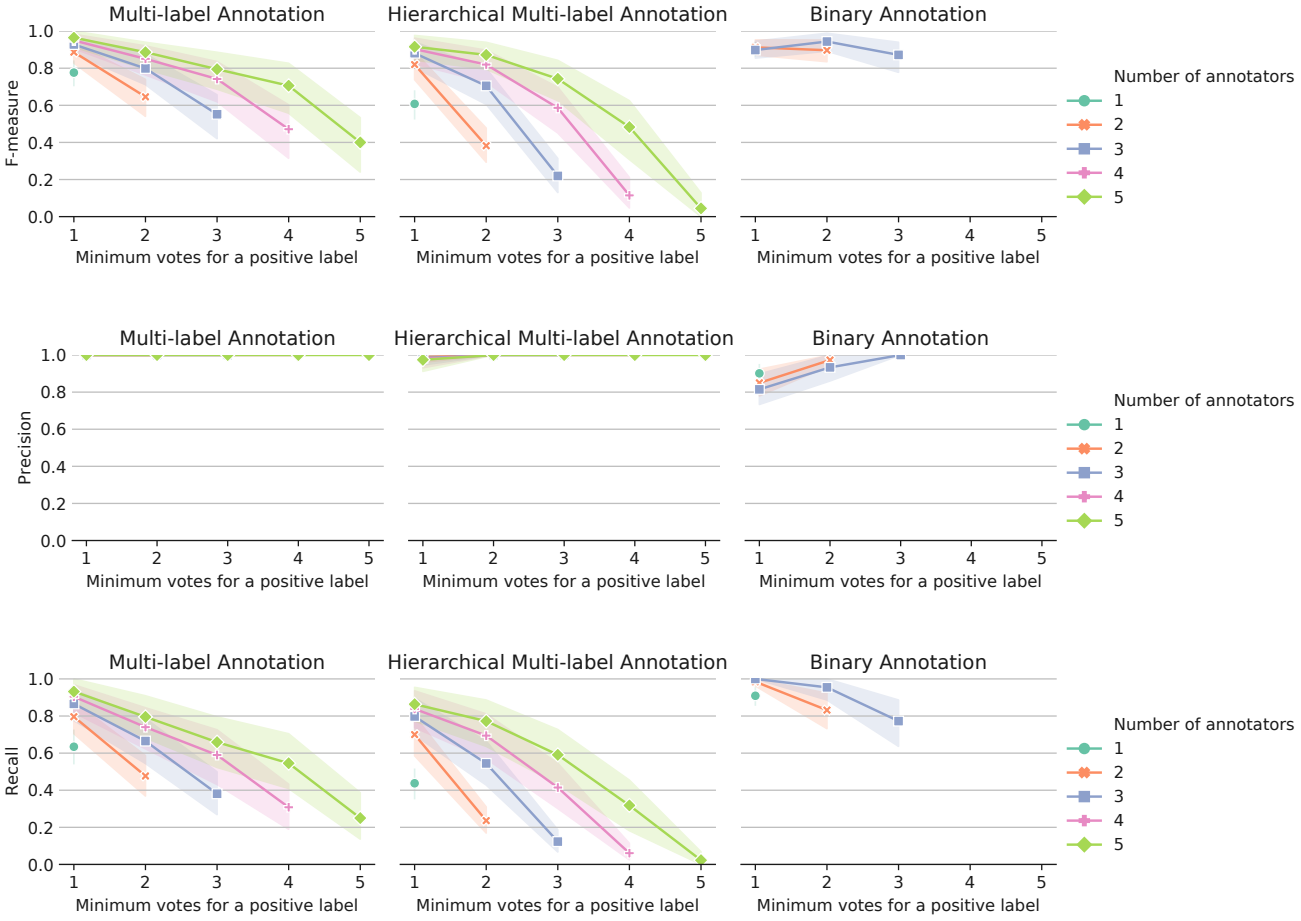


Figure 4: The quality metrics for each annotation task type while varying the the number of annotators per example and the minimum number of votes required for a positive label (i.e., voting threshold) during aggregation. The bands are the 95% CIs around the metrics, computed using 1000 bootstrap samples.

task type, number of annotators, and voting threshold. Figure 4 shows the results of varying the number of annotators and the minimum voting threshold for all three task types.

Using a three-way ANOVA, we investigated the effect of annotation task type, number of annotators, and voting threshold on the number of type I (false positive) and type II (false negative) errors for aggregate annotations. We computed the ANOVA directly on the type I and II errors rather than macro-averaged metrics because our limited ground-truth labels for each example and/or class could lead to ill-defined precision and recall metrics with macro-averaging. For the number of type I errors, we found that task type ($F(2, 1548) = 81.62$, $p < 0.001$) and voting threshold ($F(4, 1548) = 7.15$, $p < 0.001$) had significant effects, but number of annotators ($F(4, 1548) = 1.59$, $p = 0.17$) did not. There were also significant interactions between task type and voting threshold ($F(6, 1548) = 13.75$, $p < 0.001$) and task type and the number of annotators ($F(6, 1548) = 3.54$, $p < 0.01$). For the number of type II errors, we found that task type ($F(2, 1548) = 129.74$, $p < 0.001$), number of annotators ($F(4, 1548) = 41.94$, $p < 0.001$), and voting threshold ($F(4, 1548) = 103.52$, $p < 0.001$) all had significant effects. There were also significant interactions between voting threshold and the number of annotators ($F(6, 1548) = 6.76$, $p < 0.001$).

To test where these differences occurred, we also ran post hoc Tukey HSD tests ($\alpha = 0.05$), but we only report differences in main effects for simplicity. For type I errors, we found significant differences between the binary task and the two multi-label tasks but not between the two multi-label tasks themselves. In addition, we only found significant differences in voting thresholds pairs 1–2 and 2–3. For type

II errors, we found significant differences between all task types and all thresholds, but did not find significant differences in number of annotator pairs 1–2, 2–3, and 3–4.

Overall, we found that with low voting thresholds, the binary task produced aggregate annotations with more type I errors (lower precision), and with high voting thresholds, the multi-label tasks produced aggregate annotations with more type II errors (lower recall). We didn't find any significant differences between the two multi-label tasks in recall, but we found that the hierarchical multi-label task produced annotations with lower recall than the multi-label task as the voting threshold increased. At low voting thresholds for all tasks, we found minimal differences in performance when the number of annotators was 3 and above, but more annotators was always better. The aggregate annotations from the multi-labeling and hierarchical multi-labeling tasks both attained their highest F-measures with five annotators and a voting threshold of one annotator (0.96 and 0.92 respectively). Whereas, aggregate annotations from binary labeling achieved their highest F-measure with three annotators and a voting threshold of one annotator (0.94). For fair comparison, the max F-measure of multi-label annotations aggregated with three annotators was 0.93.

5 DISCUSSION

Multi-label audio annotation is a time consuming task for which few studies have investigated best practices. Our analysis suggests using a multi-label annotation task, collecting at least three annotations per example, and aggregating them with a low voting threshold will deliver results equal in quality to those collected using binary labeling, and will do so more quickly.

We found annotators tended to "over annotate" when attending to one class at a time and to "under annotate" when asked to attend to many classes. In light of these tradeoffs, binary audio annotation may be preferred when high recall is prioritized, for example when training a gun shot detection model for which the cost of a false negative may be someone's life. And multi-label audio annotation may be preferred when precision is prioritized, for example when training an urban noise pollution detection model for which the cost of a false positive is an unnecessary investigation by a city noise inspector. When aggregated however, these tendencies can be balanced by adjusting the number of annotators and the voting threshold for each task type, making peak performance comparable. Also, the throughput of multi-label annotation was higher, which is advantageous when training a single model with multi-label output rather than multiple binary-class models. Such a single model may have higher accuracy [8], and we suspect that this difference may be greater when the model must distinguish between several similar classes.

With only 30 classes, we did not find it advantageous to use the two-stage hierarchical multi-label model. At low voting thresholds, the hierarchical multi-label task produced labels of similar quality to the single-pass multi-label task, but at high voting thresholds, the downward trend on recall is more extreme than in the multi-label case, due to compounding type II errors at each annotation stage. While we suspect that an advantage for the hierarchical approach may appear as the number of classes is increased, additional experimentation is required to investigate this.

We found the unanimous vote measure, 91% for multi-labeling and 81% for binary labeling tasks, to be higher that the equivalent reported by the AudioSet data collection (76.2%), indicating greater consistency between citizen scientists than paid crowdworkers. However, we observed low scores when calculating Krippendorff α agreement, indicating both that annotation of urban sound is a difficult task, and that there is room to improve the design of our tasks.

Limitations

Because we are presenting a study of data collection with an existing community of volunteer citizen scientists on the most well-established platform for these activities, and because we are using an intentionally restricted audio dataset, the following limitations should be acknowledged with respect to our analysis and findings.

Using the Zooniverse platform. Zooniverse is a platform designed for real-world citizen science data collection, rather than for controlled experimentation, and because of this it provides researchers with only limited control over the variations with which tasks and recordings can be presented. Therefore this was not a controlled study, rather it was more akin to an A/B test of different designs and measuring their impact. However, it is also important to recognize that findings from this study should be considered subject to the potentially biasing impact of the particular norms associated with participation in Zooniverse projects. While it appears to us that our findings are at least in part a reflection of volunteer citizens scientists greater intrinsic motivation when compared to paid crowdworkers, it is also possible that they reflect particular norms, standards, and expectations cultivated through volunteer participation in multiple Zooniverse projects over the years. As the reflections of Zooniverse's UX team indicate [52], norms are emerging around the practices associated with large-scale citizen science participation on the platform, and while it is currently and quite significantly the largest instance of such a platform, Zooniverse is not the only option available. Future studies might compare the power of these norms by running simultaneous studies on different platforms. For example, do the findings of previous studies, which both note the differing motivations

of volunteer citizens sciences, e.g. [42–44] and also used data from Zooniverse volunteers, hold true for alternative platforms? Another possibility is that the kind of network affects seen with Google, Amazon and Facebook may apply in a citizen science context to Zooniverse. Either way, the choice of platform for similar studies in the future will remain an important consideration. These norms and practices also impacted our collection of user-focused data, as we followed the advice of expert moderators and chose to remove links to external questionnaires so as not to raise barriers to volunteers' participation. Instead we hoped that data downloaded from the comments and messages that volunteer citizen scientists left on the Zooniverse 'Talk' boards would be sufficient to provide similar insight. As it turned out, there were too few of these messages to undertake a detailed qualitative analysis, and so these merely provide illustration rather than findings in their own right. Data from additional questionnaires would have helped us assess the preferences and motivations of the annotators, and enriched our user data. However, in striving for ecological validity this was a tradeoff we chose to make, exemplifying tensions between eliciting real-world annotation with citizen science volunteers and gathering rich participant data.

Generalizability to other sources of audio data. There are also limitations of our study due to our data. Our ground truth data were limited to a small selection of urban soundscapes which may reduce the generalizability of our results to other audio recordings. However, soundscapes such as these, dominated by *technological* and *human* sounds, are typically evaluated to be in the *chaotic* quadrant of Swedish Soundscape-Quality Protocol [2, 3]. In a previous study [10], we found that as urban soundscapes become more complex, annotators' precision stays about the same but their recall goes up. Therefore, while future a study is necessary to assess how our current results will translate to other types of soundscapes or audio recordings, the results from our previous study provide some insight into how our current results may translate to other urban soundscapes. A general study of audio recording annotation would likely require a dataset several orders of magnitude greater in size and would not have been feasible in our study design. The ground-truth data were also limited due to the incompleteness of their labels. The effects of this can be seen in the surprisingly perfect precision results for the multi-label annotation task. With complete ground-truth labels, we suspect that the metrics would be lower but the trends would remain the same.

Comparison to previous studies

Limitations aside, our analysis both supports and contrasts prior studies on the annotation of temporal media. We see how annotators "over annotate" in binary labeling tasks and "under annotate" in multi-labeling tasks, similarly to [50]; and that errors often compound when an annotation task is broken down into a series of dependent sub-tasks, like [54]. However, our findings oppose [22] who suggested that a multi-labeling task resulted in lower annotator agreement and unhappier crowdworkers. In contrast, messages from our Zooniverse project's 'Talk' boards include comments from volunteers who found the binary task limiting; while no comments suggest that the multi-label task was too complex or time-consuming. There are too few messages to undertake a qualitative analysis, and so we include comments from three volunteers as illustration, and in order to spark ideas for future research (N.B. the animal diary and animal camera trap projects in the quotes are referring to Snapshot Serengeti [51] where volunteers are asked to identify wildlife from motion-triggered cameras left in the countryside):

> *"There might be a better way than is that X sound yes or no to classify quicker. People will get tired of listening to sound clips faster than other quick options, like the animal diaries. You want to squeeze as much data out of each audio clip."*

> *"I hear drums, observer/audience yelling applause, at least one large size dog that is very unhappy about the noise. This takes place outside. I have no way to label more than two features, so it will probably be more frustrating than I can deal with to participate."*

> *"In my opinion, this project should use the same model as the animal camera trap projects, that is, have a list of sound categories that one can click on for each clip, and give the opinion to choose more than one category."*

There are likely to be a range of other factors contributing to our findings, each of which merits inquiry beyond the scope of this particular discussion.

Recognition over recall. For example, having the full range of classification options in the multi-label task immediately visible and directly available may be significant, bringing to mind previous HCI discussion around the relative importance of recognition over recall with regards to direct manipulation interfaces (e.g., [9, 23, 49]).

Motivations of volunteer citizen scientists. Another important factor may be the differing motivations of volunteer citizen scientists who are likely to be driven by intrinsic motivations [19, 40, 42], and paid crowd-workers who are primarily motivated by financial incentives [30, 37, 46]. This may explain why binary labeling tasks, in which individual instances can be completed quickly, have appeared to be more effective in previous audio annotation undertaken by crowd-workers;

and why, in contrast to this, our findings suggest that multi-labeling tasks may be more effective in the context of audio annotation with volunteer citizen scientists.

Volunteers' contributions and accomplishments. A reflective case study highlighting insights the Zooniverse UX team gained as the platform developed [52] further helps us to frame or findings. For instance, audio tasks on the Zooniverse platform are considered less likely to sustain volunteers' participation than image-oriented tasks, highlighting the importance of designing workflows that maximize the impact of individual contributions. A second insight is that it is important for volunteer contributors to gain a sense of accomplishment, and in more monotonous tasks (such as ours) this should be achieved quickly. While on the surface the simple binary task is completed more quickly, it is possible that annotators will be presented with a succession of examples in which the sound in question is not present. If accomplishment is more closely associated with positively identifying sounds than negatively identifying their absence, it is likely that the binary task becomes demotivating, and that the multi-label task, which enables contributors to make a positive identification in every case, leads to earlier and more consistent feelings of accomplishment. A third insight is that volunteer discussion and collaboration, beyond the initial requirements of the task, has resulted in a number of citizen-led discoveries. This indicates that citizen scientists are not necessarily looking for simple tasks that can be completed as quickly as possible. Rather, they may be motivated to extend a task, investigate further, and gain a deeper understanding. Having the full range of classifications options visible may prompt sharing and discussion, which are important to this process of social inquiry. However, one important caveat to this discussion is that our tasks required only a limited number of classification options. An increase in the number of categories from which labels are selected could quickly make multi-label tasks, in which all options were always visible, extremely challenging.

6 CONCLUSION

This paper contributes to our understanding of multi-label audio annotation crowdsourced with volunteer citizen scientists. We have described how a multi-labeling approach to annotation can result in annotations at a higher throughput and of comparable overall quality (as measured by F-score) to those obtained using binary-labeling, the technique most commonly used with paid crowdworkers. This supports previous work on image annotation with citizen scientists, and reminds us of the important differences between participants who volunteer their time and effort freely, and paid crowdworkers, which we unpack in light of insights gained from crowdsourcing, citizen science, and HCI literature.

7 ACKNOWLEDGMENTS

We would like to thank all the Zooniverse volunteers who continue to contribute to our project. This work is supported by National Science Foundation award 1544753 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1544753).

REFERENCES

[1] [n. d.]. Zooniverse. https://www.zooniverse.org/. Accessed: 2018-09-19.

[2] Ãsten Axelsson, Mats E Nilsson, and Birgitta Berglund. 2012. The Swedish soundscape-quality protocol. *The Journal of the Acoustical Society of America* 131, 4 (2012), 3476–3476.

[3] Östen Axelsson, Mats E Nilsson, and Birgitta Berglund. 2010. A principal components model of soundscape perception. *The Journal of the Acoustical Society of America* 128, 5 (2010), 2836–2846.

[4] Mathias Basner, Wolfgang Babisch, Adrian Davis, Mark Brink, Charlotte Clark, Sabine Janssen, and Stephen Stansfeld. 2014. Auditory and non-auditory effects of noise on health. *The Lancet* 383, 9925 (2014), 1325–1332.

[5] Juan Pablo Bello, Claudio Silva, Oded Nov, R Luke DuBois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. 2019. SONYC: A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution. *Commun. ACM* 62, 2 (2019).

[6] Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. 2013. Using gamification to inspire new citizen science volunteers. In *Proceedings of the International Conference on Gameful Design, Research, and Applications*. ACM, 18–25.

[7] Jonathan Bragg and Daniel S Weld. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.

[8] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. 2015. Multi-label vs. combined single-label sound event detection with deep neural networks. In *Proceedings of the European Signal Processing Conference*. IEEE, 2551–2555.

[9] John M Carroll. 1989. Evaluation, description and invention: Paradigms for human-computer interaction. In *Advances in Computers*. Vol. 29. Elsevier, 47–77.

[10] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. 2017. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction* 1, 1 (2017).

[11] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (2013), 123–133.

[12] Xiangyu Chen, Yadong Mu, Shuicheng Yan, and Tat-Seng Chua. 2010. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 35–44.

[13] 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.

[14] Caren B Cooper, Jennifer Shirk, and Benjamin Zuckerberg. 2014. The invisible prevalence of citizen science in global research: migratory birds and climate change. *PLOS ONE* 9, 9 (2014), e106508.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

IEEE, 248–255.

[16] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. 2014. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3099–3102.

[17] Martin Dittus, Giovanni Quattrone, and Licia Capra. 2016. Analysing volunteer engagement in humanitarian mapping: building contributor communities at large scale. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, 108–118.

[18] Margret C Domroese and Elizabeth A Johnson. 2017. Why watch bees? Motivations of citizen science volunteers in the Great Pollinator Project. *Biological Conservation* 208 (2017), 40–47.

[19] Alexandra Eveleigh, Charlene Jennett, Ann Blandford, Philip Brohan, and Anna L Cox. 2014. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2985–2994.

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dartaset for audio events. In *Proceedigns of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[21] Monica S Hammer, Tracy K Swinburn, and Richard L Neitzel. 2013. Environmental noise pollution in the United States: developing an effective public health response. *Environmental Health Perspectives* 122, 2 (2013), 115–119.

[22] Eric Humphrey, Simon Durand, and Brian McFee. 2018. OpenMIC-2018: an open dataset for multiple instrument recognition. In *Proceedings of the International Society for Music Information Retrieval Conference*.

[23] Edwin L Hutchins, James D Hollan, and Donald A Norman. 1985. Direct manipulation interfaces. *Human-computer Interaction* 1, 4 (1985), 311–338.

[24] Ioanna Iacovides, Charlene Jennett, Cassandra Cornish-Trestrail, and Anna L Cox. 2013. Do games attract or sustain engagement in citizen science?: a study of volunteer motivations. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1101–1106.

[25] Corey Brian Jackson, Kevin Crowston, Gabriel Mugar, and Carsten Østerlund. 2016. Guess what! You're the First to See this Event: Increasing Contribution to Online Production Communities. In *Proceedings of the International Conference on Supporting Group Work*. ACM, 171–179.

[26] Corey Brian Jackson, Carsten Østerlund, Gabriel Mugar, Katie DeVries Hassman, and Kevin Crowston. 2015. Motivations for sustained participation in crowdsourcing: case studies of citizen science on the role of talk. In *Proceedings of the Hawaii International Conference on System Sciences*. IEEE, 1624–1634.

[27] Aren Jansen, Jort F Gemmeke, Daniel PW Ellis, Xiaofeng Liu, Wade Lawrence, and Dylan Freedman. 2017. Large-scale audio event discovery in one million youtube videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 786–790.

[28] Charlene Jennett, Laure Kloetzer, Daniel Schneider, Ioanna Iacovides, Anna Cox, Margaret Gold, Brian Fuchs, Alexandra Eveleigh, Kathleen Methieu, Zoya Ajani, et al. 2016. Motivations, learning and creativity in online citizen science. *Journal of Science Communication* 15, 3 (2016).

[29] Dick Kasperowski and Thomas Hillman. 2018. The epistemic culture in an online citizen science project: Programs, antiprograms and epistemic subjects. *Social Studies of Science* (2018).

[30] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. In *Proceedings of Americas Conference on Information Systems*. 1–11.

[31] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the ACM Symposium on User Interface Software and Technology*. ACM, 43–52.

[32] Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. 2016. Embracing error to enable rapid crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3167–3179.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[34] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, and Dan Andreescu. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.

[35] Oisin Mac Aodha, Rory Gibb, Kate E Barlow, Ella Browning, Michael Firman, Robin Freeman, Briana Harder, Libby Kinsey, Gary R Mead, and Stuart E Newson. 2018. Bat detective—Deep learning tools for bat acoustic signal detection. *PLOS Computational Biology* 14, 3 (2018), e1005995.

[36] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.

[37] Winter Mason and Duncan J Watts. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 77–85.

[38] Greg Newman, Andrea Wiggins, Alycia Crall, Eric Graham, Sarah Newman, and Kevin Crowston. 2012. The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment* 10, 6 (2012), 298–304.

[39] Oded Nov, Ofer Arazy, and David Anderson. 2011. Dusting for science: motivation and participation of digital citizen science volunteers. In *Proceedings of the iConference*. ACM, 68–74.

[40] Oded Nov, Ofer Arazy, and David Anderson. 2014. Scientists@ Home: what drives the quantity and quality of online citizen science participation? *PLOS ONE* 9, 4 (2014), e90375.

[41] Nathan R Prestopnik and Kevin Crowston. 2011. Gaming for (citizen) science: Exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system. In *IEEE International Conference on e-Science Workshops*. IEEE, 28–33.

[42] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Carie Cardamone, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. 2013. Galaxy Zoo: Motivations of citizen scientists. (2013). arXiv:physics.ed-ph/1303.6886

[43] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. 2010. Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review* 9, 1 (2010), n1.

[44] Jason Reed, M Jordan Raddick, Andrea Lardner, and Karen Carney. 2013. An exploratory factor analysis of motivations for participating in Zooniverse, a collection of virtual citizen science projects. In *Proceedings of the Hawaii International Conference on System Sciences*. IEEE, 610–619.

[45] Hauke Riesch and Clive Potter. 2014. Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. *Public understanding of science* 23, 1 (2014), 107–120.

[46] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets.

Proceedigns of the International AAAI Conference on Web and Social Media, 17–21.

[47] Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. 2012. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, 217–226.

[48] Lior Shamir, Carol Yerby, Robert Simpson, Alexander M von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller, and John Wallin. 2014. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *The Journal of the Acoustical Society of America* 135, 2 (2014), 953–962.

[49] Ben Shneiderman. 1982. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology* 1, 3 (1982), 237–256.

[50] Gunnar A Sigurdsson, Olga Russakovsky, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Much ado about time: Exhaustive annotation of temporal data. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.

[51] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific data* 2 (2015), 150026.

[52] Ramine Tinati, Max Van Kleek, Elena Simperl, Markus Luczak-Rösch, Robert Simpson, and Nigel Shadbolt. 2015. Designing for citizen data analysis: a cross-sectional case study of a multi-domain citizen science platform. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 4069–4078.

[53] Anthony Truskinger, Mark Cottman-Fields, Daniel Johnson, and Paul Roe. 2013. Rapid scanning of spectrograms for efficient identification of bioacoustic events in big data. In *Proceedings of the IEEE International Conference on eScience*. IEEE, 270–277.

[54] Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision* 101, 1 (2013), 184–204.

[55] Chris Wood, Brian Sullivan, Marshall Iliff, Daniel Fink, and Steve Kelling. 2011. eBird: engaging birders in science and conservation. *PLOS Biology* 9, 12 (2011), e1001220.

[56] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464.