**Abstract**

**Introduction**

The Zooniverse is an online citizen science platform that engages volunteers by asking for help in the interpretation of large-scale datasets. At [www.zooniverse.org](http://www.zooniverse.org), researchers from across disciplines ask volunteers to apply their pattern-recognition skills to classify what they see in images or videos, such as the species of wildlife captured in camera traps or the morphology of galaxies in telescope imagery. Since its inception with the launch of [www.galaxyzoo.org](http://www.galaxyzoo.org) in 2007, the Zooniverse has grown to include over 1.6 million registered volunteers who participate across more than 100 different research projects; the need for citizen science in data interpretation shows no signs of abating. Since launching the project builder platform (www.zooniverse.org/lab) in 2015, which allows researchers to launch projects without customized software development, the number of research projects within the Zooniverse has more than doubled (Figure 1). However, such rapid growth in citizen science projects threatens to outpace the growth of volunteer community; citizen science projects need to become more efficient and more engaging to keep pace with the data-analysis demand.

Mobile technology andmachine learning advancements provide novel opportunities for enhancing efficiency in online citizen data processing. Mobile phones allow increased volunteer engagement throughout the day, whereas machine learning enables automatic image identification, thus reducing the number of images requiring human input. However, not all online citizen science projects are immediately suited to these approaches: for example, ecology projects (e.g. [www.snapshotserengeti.org](http://www.snapshotserengeti.org), [www.cameracatalogue.org](http://www.cameracatalogue.org)) often ask volunteers to identify which of over forty different species is present in an image, which can be challenging to communicate quickly via a mobile interface. Similarly, identifying which of dozens of species is present in a photograph is a largely unsolved machine learning problem.

“Collaborative filtering” is an alternative approach to more traditional survey-based image classification. By converting a complex task into a series of simple binary (i.e. yes or no) questions presented separately to volunteers, the collaborative filtering approach can allow projects to better leverage mobile technology and machine learning methodologies. A traditional species identification task could, for example, be broken down into a series of questions that asks first about the presence of any animal, then about the presence of a specific animal or group of animals, and so on, with each binary question successively filtering the dataset down to the images the research team needs (see Figure 2).

Simplifying a complex task into binary questions not only provides more portability to mobile devices, but also enables the integration of simpler machine learning for automated image identification at every step of the filtering process. Thus, although images would ultimately require more classifications to produce the same amount of final information, the ease of classification and ability to integrate existing machine learning routines could provide benefits to efficiency that far outweigh the increase in classification counts.

However, altering interface design can have unanticipated impacts on the volunteer community and classification rates (Bowyer et al. 2016, Spiers et al. in press). Thus, changing the classification methodology from a traditional survey design to a collaborative filtering approach could impact data quantity and quality as well as the nature of volunteer communities contributing to a project.

Here we evaluate the effects of these different designs within a single Zooniverse project, Snapshots at Sea ([www.snapshotsatsea.org](http://www.snapshotsatsea.org)), which asks volunteers to identify humpback whale flukes (tails) from tourist-contributed images. By providing volunteers with a choice of workflows to classify with, we were able to evaluate differential effects of the alternative workflows on the project efficiency and accuracy, as well as effects on the volunteer community, with implications for researchers engaging in citizen science for data interpretation.

**Methods**

**Data Collection**

Snapshots at Sea was launched June 9, 2016 to identify images containing humpback whale flukes from photographs taken by tourists and researchers on board cruise ships. The undersides of humpback whale flukes are individually identifiable by color pattern and shape; images containing humpback whale flukes with visible undersides are passed to another citizen science project, Whales as Individuals, where volunteers help prepare the photos for individual recognition algorithms.

We acquired 13,264 images taken by cruise passengers on board X [Need info from Ted]. We randomized the images into two subsets of 6,632 images each, which were then assigned to a “Survey” workflow (standard for many ecology projects) and a “Yes/No” workflow (“Collaborative filtering”). Because each individual Yes/No question provides less information than the survey task, the collaborative filtering approach requires more classifications to complete the same number of images (see Figure 2 for a detailed description of each workflow, retirement rules, and efficiency).

*‘Yes/No’ workflow*

The collaborative filtering workflow was comprised of four separate binary (yes/no) questions, of which only one question was active on the website at a time (detailed in Figure 2).

Each image was seen by five volunteers prior to retirement. Following retirement of all images associated with one question, all images of interest were identified based on the percentage of volunteers who answered “yes” to that question. Images of interest were then taken forward, forming the subject set of the next question. All 6,632 images were passed to Q1; images with 75% agreement on the presence of an animal were passed to Q2 (n = 5,910); images with 75% agreement on the presence of a whale or dolphin were passed to Q3 (n = 5,020); images with 50% agreement on the presence of a fluke were passed to Q4 (n = 1,470).

*‘Survey’ workflow*

The survey workflow provided a more direct route to classification by presenting volunteers with a list of options for every image: “Whale or Dolphin,” “Seal, sea otter, etc.,” “Bird,” “Fish,” and “No animal.” When volunteers select one of these options, a pop-up screen appeared with reference photos and additional details about their selection, and they could choose to select either “cancel” or “identify.” If volunteers selected “Whale or Dolphin” from the menu, the pop-up screen questions: “Is the underside of the tail visible?” (answers: “yes” or “no”), and “If YES, does the tail belong to a humpback whale?” (answers: “yes” or “no”). Volunteers could select and identify multiple species in an image. When they were done classifying, they submitted their identifications by selecting “Done” below the species list. Because the Survey workflow task was more complex than the binary questions in the Yes/No workflow, images were shown to 10 volunteers before being retired from circulation.

*Volunteer communication*

We launched the experiment on June 7, and announced it via newsletter on June 8. Because Snapshots at Sea already had an active volunteer community used to the collaborative filtering approach, we sent one email to the existing community announcing new data and a new, alternative workflow, and one email to volunteers who had previously participated on another, now-retired, marine-focused project.

When Question 1 of the Yes/No workflow was completed, we activated Question 2 and sent out another newsletter announcing both new data and the alternative survey workflow. This newsletter went to all volunteers who had classified on Snapshots at Sea before or during the experiment, thus including all newly recruited volunteers. After all four questions on the Yes/No workflow were completed, we de-activated the Yes/No workflow and sent out a final newsletter to the Snapshots at Sea community alerting them to the remaining data on the survey workflow.

**Data analysis**

We retrieved all classifications made on Snapshots at Sea during the study and associated metadata, including volunteer user name, the date and time of classification, and the device used to classify. We included the following fields in our analysis:

* *user\_name:* The volunteer’s distinct Zooniverse login id. Non-logged volunteers are recorded in the metadata as “not-logged-in.”
* created\_at: The date and time the classification was recorded in the database.
* *duration:* Calculated as the time from when the image was loaded on the volunteer’s screen to when the volunteer submitted the classification. Durations > 2 minutes were excluded as these typically represent incidents when the volunteer had loaded a classification but was not actively classifying.
* *device:* Defined from user-agent strings as computer, tablet, mobile, and swipe app for classifications made using the Zooniverse mobile app where users swipe left for “no” and right for “yes.
* *session:* Defined as a continuous period of classifications with less than 30 minutes between consecutive classifications.

*Classification rates and volunteer contributions*

We classified volunteers as existing, if they had classified on Snapshots at Sea prior to launching the survey workflow, and as new, if their first classification was during the experiment. We evaluated skew among volunteer contributions by calculating Lorenz curves (package *ineq*, function *Lc*) and Gini coefficients (package *reldist*, function *gini*) for each workflow for new and existing volunteers. Lorenz curves give the proportion of classifications vs. proportion of volunteers; y = x represents hypothetical perfect equality, where every volunteer contributes the same number of classifications. The Gini coefficient is a measure of inequality (calculated as the ratio of the area above the Lorenz curve and below the line of y = x to the total area below the line of y = x) and ranges from 0 to 1, where 0 represents completely equal contributions and 1 represents completely unequal contributions.

*Engagement*

We identified volunteers as classifying on both workflows, yes/no only, and survey only, and summarized the number of classifications made, the proportion of each group who commented on the Talk forums, and the number of talk comments made during the 31 days in which both workflows were available for classifying.

*Aggregation & Accuracy*

We aggregated multiple volunteer answers into a single final “consensus” answer for every image. For Yes/No questions, we calculated the proportion of volunteers who answered “yes” for a given image and calculated the consensus answers as “yes” if the proportion exceeded 0.75 for Q1 and Q2, 0.45 for Q3, and 0.5 for Q4. While more sophisticated approaches exist for aggregating survey-style tasks (Swanson et al. 2014), we applied an analogous approach to the survey task to facilitate comparison between the workflows. For every image, we recorded the proportion of volunteers who identified the presence of any animal, a whale or dolphin, a fluke, and whether or not the fluke belonged to a humpback, and calculated consensus answers according to the same thresholds.

We compared volunteer answers to a dataset of X classifications created by the project’s scientific lead, T.C. Because false positive errors are easily filtered out in subsequent image processing, we compared rates of false negatives.

**Results**

*Classification rates & volunteer contributions*

Volunteers completed the Yes/No workflow twice as quickly as the Survey workflow, contributing the 95,195 classifications needed in only 22 days, compared to the 41 days to contribute the 66,320 classifications needed to complete the Survey workflow (Table 2, Figure 3). Classification rates following newsletters spiked for both workflows, but the increases were much higher for yes/no questions. This was especially apparent when volunteers had a choice of which workflow to work on: survey workflow classification rates peaked following the July 14 newsletter, when it was the only available workflow.

Volunteers contributed more per-capita classifications on the Yes/No questions, which took much less time to complete, than on the Survey questions: (Figure 4), volunteers actually spent more time classifying on the collaborative filtering questions (*might need to revise this analysis to just sum duration of classifications instead of sessions, as sessions can be mixed)*.

There was greater inequality in volunteer contributions on the Yes/No workflow than on the survey workflow (Figure 5), regardless of whether the volunteers were *existing* volunteers (those who had classified prior to the experiment) or *new* volunteers (who joined Snapshots at Sea during the experiment), with Gini coefficients ranged from 0.762 (new users) to 0.745 (existing users) for the survey task and 0.825 (new users) to 0.828 (existing users) on the yes/no questions.

*Volunteer engagement*

For the 31 days in which both workflows were available on the website, a similar number of volunteers contributed to both workflows as contributed to Yes/No only and Survey only (Table 2). However, volunteers who contributed to both workflows classified many more images and participated more actively on the discussion forums than those who only classified on one workflow or the other: 14% of volunteers who classified on both workflows commented on *Talk*, while only 4.7% of survey-only and 3.7% of yes/no-only volunteers commented on *Talk*. Volunteers who contributed to both workflows still contributed many more yes/no classifications than survey classifications, within and across sessions [a plot seems unnecessary here – do I need to even mention this?].

*Accuracy*

Accuracy was similar across workflows, with 0% false negative error on most questions, and low rates (2.2% survey, 5% yes/no) for identifying the presence of a tail (Table 3). However, in the survey approach, false negative error for the presence of tails can be corrected if volunteers identify a humpback in an image; these false negatives cannot be corrected in the yes/no questions because they are filtered out of the dataset.

After applying Ted’s aggregation rules for % agreement necessary, it looks like false negative rates are pretty equivalent at capturing the presence of a whale or dolphin (0% Type 2 error through both approaches), capturing the presence of a tail (5% yes/no and 2.2% survey), and capturing the presence of a humpback given a tail (0% false negatives for both).

**Discussion**

Collaborative filtering provides an effective and efficient route to classifying an otherwise complex task.

* While the simple yes/no questions engaged a similar number of volunteers as the survey workflow, each volunteer spent more time and contributed more classifications than on the survey workflow. Thus, despite the collaborative filtering approach being inherently less efficient, data were processed more quickly via this route.
  + This is probably because classifications were so much faster on the yes/no surveys, especially on the swipe app available on mobile phones.
  + Do we have references evaluating how users contribute more when it’s easier to do? Blanks experiment on SS suggests users stay longer & contribute more classifications when they are easier
* Collaborative filtering does appear to engage volunteers differently than the survey task – the volunteer contribution was more unequal, regardless of whether volunteers had previously classified on SAS or not.
  + However, the implications of this inequality, and how much skew is “too much”, are not clear.
  + The collaborative filtering contributions are well within the range of other successful Zooniverse projects (~ 0.7 - 0.9, as per Cox et al. & presumably Spiers et al)

Mobile is a viable route for classifications: Classifications on the yes/no workflows are much faster on mobile. Also, despite no announcement of Snapshots at Sea’s recent availability on mobile phones, mobile classifications make up ~20% of all classifications contributed and > 30% of yes/no classifications, suggesting that mobile is a viable and largely unexplored route for contribution.

* + Need something about machine learning – ummm, are there actual existing techniques that could be applied here?
* However, engagement on *Talk* is similar across workflows, suggesting that this might not necessarily be a negative impact. While the percentage of users from each workflow who comment is similar, the users on the survey task only comment more frequently on the discussion forums than those who only classify on the yes/no tasks. However, these numbers are dwarfed by the engagement from users who contribute to both workflows. Since volunteers who contribute to both workflows still contribute overwhelmingly to yes/no questions, perhaps the difference in engagement reflects something about the interest of the volunteer or a limitation of the device on which they are classifying, instead of whether the task itself encourages engagement.

Take-home message:

* Collaborative filtering approaches work really well in this case – people like them and they get through data quickly. However, this type of workflow does appear to change the relative engagement of volunteers in the community, increasing the inequality of classifications and perhaps encouraging users who engage little on Talk. [I don’t think this is bad, I think it just might be an alternative route for users who don’t want to engage all that much].
* Limitations: Snapshots at Sea is a relatively simple survey question, and the collaborative filtering approach only requires 4 distinct binary questions to achieve the same level of information as a survey question. Thus, a collaborative filtering approach may not be all that inefficient. In contrast, trying to use collaborative filtering to identify all 60 species from Snapshot Serengeti would require significantly more questions.
* Applications: Perhaps as a filtering step in an otherwise fully complex survey task, say, filtering out empty images or those with the most common animals (e.g. filtering out wildebeest and blanks in SS would reduce the dataset by 80%). Allowing users to engage in multiple routes to classification would provide opportunities to engage with the images at different levels, according to the volunteer’s desires.

[Possibly raise in discussion?] For example, Supernova Hunters posts new data every day at 2pm and announces when the data are available. [Volunteers complete a new dataset within X minutes, meaning that only a small percentage of would-be contributors ever have the opportunity to participate, and driving the demographics to overwhelmingly older white male volunteers.]??

* Any broader citations that might cause us to expect a change in community?

- Here we explore the potential consequences of implementing a collaborative filtering approach in terms of user contributions, engagement, and accuracy by simultaneously running both types of workflow simultaneously on an existing Zooniverse citizen science project, Snapshots at Sea.

**Tables and Figures**

Figure 1: Growth of Zooniverse projects through time. Each point and vertical gray line reflect the launch of a new project. Blue line denotes the launch of the Zooniverse project builder platform, which enables research teams to build their own Zooniverse project without custom web development.

Figure 2: Snapshots at Sea workflows and retirement rules.

Figure 3: Instantaneous (a) and Cumulative (b) classifications through time for each approach. Cumulative classifications are plotted as the proportion of total required to complete the dataset (66,320 for the survey workflow and 95,195 for the collaborative filtering workflow). Data additions are indicated in gray dashed vertical lines, newsletters in blue.

Figure 4: Distribution of classification durations (time to complete a single classification) for each workflow across different devices.

Figure 5: Lorenz curves for each workflow. users split into *existing* if they had classified prior to the experiment and *new* if they only began classifying during the experiment.

Table 1: Summary stats per workflow

Table 2: Users, Classifications, and Talk comments

Table 3: Accuracy