**Title?? [Suggestions welcome]**

Alexandra Swanson, Helen Spiers, Lucy Fortson, Ted Cheeseman, Chris Lintott

**Abstract**

[Coming…]

**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 the 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, while progress has been made by applying deep learning techniques on images that contain a single species (Norouzzadeh et al 2017), identifying which of dozens of species is present in a photograph continues to be a challenging machine learning problem.

“Cascade 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 cascade 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).

This process enables the integration of simpler machine learning for automated image identification at every step of the filtering process (see e.g. Allwein et al 2000). For example, one need not focus on optimizing a learning model for an arbitrary number of species that may be project dependent. Simple binary classifiers, while potentially not optimum alone, will be sufficient with human classifiers implemented alongside machines as is enabled by citizen science platforms such as Zooniverse (Wright et al 2017).

Reducing a complex task into binary questions also enables better portability to mobile devices: the classification task is simpler, thus requiring less screen “real-estate,” and can also be reduced to gross motor movements such as “swiping” left or right. Thus, although images would ultimately require more classifications to produce the same amount of final information, the ease of classification and ability to integrate simple 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 cascade 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. The results have implications for researchers engaging in citizen science for data interpretation and design decisions related to task structure whether in-browser or in-app and eventually in tandem with machine learning.

**Methods**

**Data Collection**

*Snapshots at Sea* was launched June 9, 2016 in cascade filtering mode 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, which allow for high resolution population monitoring. 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. [TED – please add a sentence about how the researchers are using the data, e.g. tracking behavior/populations/etc?]

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, e.g. Swanson et al. 2015) and a “Yes/No” workflow (“Cascade filtering”). Because each individual Yes/No question provides less information than the survey task, the cascade 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 cascade 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 simultaneous 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 on one image, they submitted their identifications by selecting “Done” below the species list and were then shown the next randomly selected image. 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, 2017, and announced it via newsletter on June 8, 2017. Because *Snapshots at Sea* already had an active volunteer community used to the cascade 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. When both workflows were complete, we emailed volunteers asking them to complete a survey about their demographic information and workflow preferences.

**Data analysis**

*Classification rates and volunteer contributions*

We retrieved all classifications made on *Snapshots at Sea* during the study and associated metadata, including volunteer user name (or “not-logged-in” if an unregistered volunteer), the date and time of the classification, and the *device* (defined from the user-agent strings) used to classify. We classified registered 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 calculated the time to complete a classification (*duration*) as the time from when the image was loaded on the volunteer’s screen to when the volunteer submitted the classification; classifications with *duration* > 2 minutes were excluded as these typically represent incidents when the volunteer had loaded a classification but was not actively classifying. We identified a *session* asperiod of continued classifications with less than 30 minutes of inactivity between consecutive classifications. All analyses were conducted in Program R 3.4.0 and code is available online at [www.github.com/aliburchard/snapsatsea](http://www.github.com/aliburchard/snapsatsea).

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 registered 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 evaluated rates of false negatives in volunteer consensus answers to a dataset of 135 classifications created by the project’s scientific lead, T.C. (20 images for each of the four Yes/No questions and 55 Survey images).

**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.

Daily per capita volunteer contribution was higher on the Yes/No workflow (5.00 classifications per person per day) than on the Survey workflow (1.95 classifications per person per day). The Yes/No classifications were much faster to complete (Figure 4): the median duration a classification ranged from 8.2 – 10.7 seconds for survey tasks and 1.0 – 3.7 seconds for a Yes/No task, depending on the device used to classify. However, volunteers spent longer classifying when working on Yes/No tasks: median session length for volunteers classifying on exclusively Yes/No questions was 29 minutes compared to 21 minutes when classifying on both tasks and 16.7 minutes when classifying exclusively on the Survey task.

There was greater inequality in volunteer contributions on the Yes/No workflow than on the survey workflow (Figure 5a), 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). 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. The increased skew on the Yes/No workflow classifications appears to be driven primarily by increases in the number of users contributing >10,000 classifications (Figure 5b), though also reflect increases in the number of users contributing <10 classifications.

*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*.

*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. However, in the Survey approach, false negative error for the presence of tails could be corrected if volunteers identify a humpback in an image; these false negatives cannot be corrected in the Yes/No workflow because they are filtered out of the dataset.

*Follow-up survey*

Of the X volunteers who received the survey, 128 responded, but not all respondents answered all questions. According to the survey responses, *Snapshots at Sea* volunteers were twice as likely to be female as male, and 50% of volunteers had at least a university degree (Figure 6; supplementary table with responses??). Overall, 40% of volunteers preferred the Yes/No workflow and only 20% preferred the Survey workflow, but the proportion of volunteers who preferred each workflow did not vary significantly across age, gender, or education levels.

**Discussion**

Cascade filtering provides an effective and efficient route to for engaging volunteers in data interpretation, offering benefits for both researchers eager to process raw data and volunteers eager to contribute. While the simple Yes/No questions engaged a similar number of volunteers as the survey workflow, volunteers contributed more than twice as many Yes/No classifications per day than on the Survey workflow (Table 1, Figure 3), completing the dataset in half the time (21 vs. 41 days) with no decline in accuracy or volunteer engagement.

The success of cascade filtering is likely driven by simplicity of the task and the speed and ease with which volunteers are able to make a tangible contribution. An average Yes/No task took anywhere from one-tenth to one-half the time of a Survey task, depending on the complexity of the image and the device used. Peoples’ preference for simple tasks is well documented in [gaming and volunteer??] literature (e.g. ??), and demonstrated on other Zooniverse projects. For example, a study of volunteer behavior on *Snapshot Serengeti* (www.snapshotserengeti.org), which asks volunteers to identify African animals as one of >50 species (Swanson et al. 2015), revealed that volunteers classified more images when they received high proportions of “blank” images, which typically take less time and effort to classify (Bowyer et al. 2016).

In addition to being faster, Yes/No questions were also more predictable: because only one Yes/No question was active at a time, volunteers knew before classifying whether they would be asked to identify the presence of an animal – a generally easy task – or to identify a visible whale fluke as a humpback – a generally more difficult task. In contrast, volunteers classifying on the survey workflow could receive images ranging from very easy to very difficult in any given session. Thus, if a volunteer were seeking the type of low mental-overhead activity characteristic of popular mobile phone games such as [candy crush???], they could only reliably find it only on the Yes/No workflow.

In addition to providing faster data completion, the cascade filtering approach facilitates easier data analysis: binary classifications are easier to parse, analyze, and “retire” from circulation. The raw annotation data from Zooniverse projects comes as JSON strings embedded within a comma separated text file. JSON is a flexible, non-tabular data format that is common data in web development but uncommon in many scientific disciplines. Data from binary question tasks are easy to “flatten” into more recognizable data-frames and tables, and aggregating individual volunteer answers into a consensus dataset can be as simple as calculating a single proportion for each question. In contrast, data from survey tasks contain nested lists and arrays that are complex to flatten; aggregating data requires complex logic and data restructuring to accommodate the data format, the potential presence of multiple species in each image, and disagreement among main questions that propagates to species-specific sub-questions.

The standardized format of binary classifications further enables more straightforward retirement rules and, eventually, easier integration of machine learning routines. The Zooniverse platform allows researchers to apply logical rules to image circulation, such as retiring an image as soon as three consecutive volunteers have identified that it has no animals present, or retiring an image after at least five volunteers have classified *and* their answers have at least 80% agreement. The rules for binary classifications are significantly more tractable to implement than for survey tasks, which require many layers of conditional logic. Machine learning routines designed for binary questions could also be integrated at every stage within a cascade filtering workflow [WORDS?].

The cascade filtering approach provides clear benefits to research goals for citizen science projects. However, research teams interested in designing a project should consider potential implications for their volunteer community as well, as changing the design can affect who participates and how they engage. For example, [?? community from Supernova Hunters, where 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.??]

However, while the cascade filtering approach does appear to have implications for volunteer engagement, they do not appear to be negative or costly. Volunteer contribution was more unequal on the Yes/No workflow than on the Survey workflow; this was true regardless of whether or not the volunteers had previously classified on *Snapshots at Sea*, indicating that the skew resulted from the nature of the task and not driven by volunteers seeking out what was familiar. However, this skew is not necessarily negative, and extremely low skew could actually reflect low retention. With Gini coefficients of 0.756 for the Survey task and 0.830 for the Yes/No questions, this skew was well within the range of other successful Zooniverse projects, which range from 0.7 - 0.9 (Cox et al. 2015, Spiers et al. in prep).

Examining the underlying volunteer contributions reveals that this inequality appears driven more by an increase in super-users than an increase in visitors. Figure 5b demonstrates slight increases in both types of volunteer, but more dramatic increases in especially active ones: on the Survey workflow, 24 volunteers contributed only 1 classification compared to 34 volunteers on the Yes/No workflow, but only 11 volunteers contributed >1000 Survey classifications compared to 24 volunteers on the Yes/No workflow.

Thus, the simplicity of the Yes/No task does not appear to limit volunteer activity or engagement. Volunteers classifying Yes/No questions spent nearly twice as long in a single sitting as those classifying Survey questions and participated just as actively on the *Talk* forums. Interestingly, volunteers who classified on both workflows were three times more likely to participate on *Talk* than those who only classified on one workflow. These differences may reflect something about the interest and engagement of the individual volunteer instead of whether the task itself encourages a deeper engagement with the science.

In general, increasing the amount of time that volunteers participate on projects leads to good things, like [CITATIONS?? Increased activity and retention leads to…stuff? Accuracy tends to improve, but do we have data? People learn more and, um, stuff?]

There are, however, limitations to this design.

*Snapshots at Sea* is a relatively simple survey question and the cascade filtering approach only requires four binary questions to produce the necessary resolution of information. In contrast, many camera trap projects ask volunteers to identify animals from among 20-60 different wildlife species and answer follow-up questions about individual counts or behaviors. Thus, the sheer number of binary questions required to convert a standard survey project to cascade filtering might outweigh any increases in efficiency.

Furthermore, volunteers show clearly divergent preferences for styles of contribution. The follow-up survey revealed that the 40% of volunteers who preferred the Yes/No tasks did so mostly because the questions were quick and rewarding, while the 20% of volunteers who preferred the Survey workflow did so because they were able to contribute a more robust and meaningful classification.

Integrating both approaches into a project could simultaneously increase data throughput and allow volunteers to engage with a given project according to their own preferences. For example, a more standard camera trap project like *Snapshot Serengeti* could add an “Empty or Not” workflow that allows volunteers to filter out the 60-80% of empty images without animals, before passing animal-images on to the standard Survey workflow. The Zooniverse project builder ([www.zooniverse.org/lab)](http://www.zooniverse.org/lab)) makes it trivial to build multiple routes to classification for a single project. In fact, since completing this test on *Snapshots at Sea*, we have implemented both “Empty or Not” and a “Vehicle or Not” workflows on Panthera’s *Camera CATalogue* project ([www.cameracatalogue.org)](http://www.cameracatalogue.org)).

Cascade filtering provides a rapid, efficient, and accurate alternative to traditional project design, providing clear benefits for both research teams and the volunteer community. [Some fancy uplifting wrap up words about allowing researchers to engage volunteers and keep pace with their data production, doing bigger and better science…Suggestions welcome!!!]

**Tables and Figures**

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



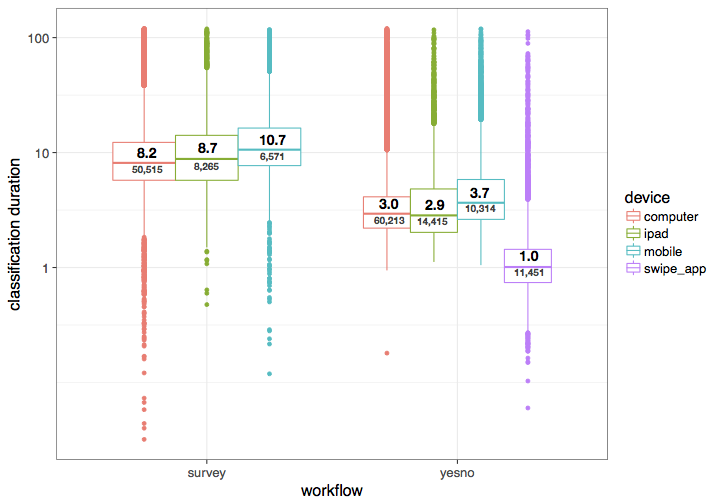
**Figure 2:** *Snapshots at Sea* workflows and retirement rules.



**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 Yes/No workflow). Data additions are indicated in gray dashed vertical lines, newsletters in blue. The completed Yes/No workflow was completed on June 27 and de-activated (i.e. removed from the website) on July 5.



**Figure 4:** Classification duration in seconds (calculated as time that an image loaded into a volunteer’s browser to the time that “Done” was selected) for each workflow across different devices. Median durations given in black text above the median line; sample sizes given in gray text below the median line. Note that the Y-axis is on a Log10 scale.



**Figure 5:** (a) Lorenz curves for each workflow. Registered volunteers (i.e. those with Zooniverse accounts) were categorized as *existing* if they had classified prior to the experiment and *new* if they only began classifying during the experiment. (b) Probability density distributions for classifications per registered volunteer for each workflow (note that the x-axis is plotted on a log-scale).



**Figure 6:** (Not sure this is necessary, mostly here for your info/input – if we keep, will clean up.)





****

****