**Keeping it simple: An experiment in online classification design**

Alexandra Swanson1, Brooke Simmons2, Helen Spiers1, Lucy Fortson3, Ted Cheeseman4, Chris Lintott1

1. Department of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH
2. Center for Astrophysics and Space Sciences, University of California, San Diego, La Jolla, CA 92093
3. Lucy affiliation
4. Ted affiliation

**Abstract**

As researchers increasingly engage volunteers in large-scale data interpretation through online citizen science platforms like [www.zooniverse.org](http://www.Zooniverse.org), projects need to become more efficient and engaging to keep pace with demand for data analysis. Here we introduce “cascade filtering” as an alternative approach to traditional image classification projects. By converting a complex task into a series of simple questions presented separately to volunteers, cascade filtering allows projects to better leverage mobile technology and machine learning methodologies. We evaluated the effects of a cascade filtering project design for the Zooniverse project *Snapshots at Sea,* which identifies humpback whale flukes. We split a set of 13,324 images between two versions, one implementing a ‘traditional’ (“Survey”) design, and the other using a cascade (“Yes/No”) design, and compared volunteer contributions and engagement across the two implementations. Volunteers completed the Yes/No workflow in half the time it took to complete the Survey workflow, with no meaningful reductions in accuracy or engagement on the discussion forum. Surveying project participants indicated that the speed and popularity of the cascade filtering approach likely derived from the simplicity of the task and the speed and ease with which volunteers were able to make a tangible contribution. Although cascade filtering increased inequality among volunteer contributions, this was driven primarily by extremely prolific classifiers. We recommend that researchers consider implementing cascade filtering, but acknowledge that this may not be appropriate for projects with complex or non-binary data, and more work needs to be done to assess potential long-term impacts to the volunteer community.

**Introduction**

The Zooniverse is a citizen science platform that engages volunteers 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 (Lintott et al. 2008, Fortson et al. 2012, Willet et al. 2013, Simpson et al. 2014, Candido dos Reis et al. 2015, Swanson et al. 2015, Kosmala et al. 2016). 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. In the two years since Zooniverse launched the project builder platform ([www.zooniverse.org/lab)](http://www.zooniverse.org/lab)), which allows researchers to launch projects without customized software development, the number of research projects within the Zooniverse has more than doubled (Figure 1). The accelerating production of big data and concomitant growth of citizen science projects threaten to outpace the growth of the volunteer community; citizen science projects need to become more efficient and more engaging to keep pace with demand for data analysis.

One solution is to combine human and machine classification; Wright *et al*. (2017) recently demonstrated that the combination of human and machine classifiers for a deployed Zooniverse project outperformed either one alone. However, the diversity of citizen science projects makes it hard to find a general solution, for example, machine learning solutions developed for one complex classification may not easily be transferred to other projects, as seen with Norouzzadeh *et al*.’s (2017) convolutional neural networks that successfully identify Serengeti wildlife but have yet to be applied to other species. An alternative solution to ensure project sustainability is to find novel ways to engage volunteers, for example through a mobile app. However, this too has challenges, for example, some projects cannot easily be reduced to interfaces simple enough for mobile development, such as ecology projects (e.g. [www.snapshotserengeti.org](http://www.snapshotserengeti.org), [www.cameracatalogue.org](http://www.cameracatalogue.org)) that ask volunteers to identify which of over forty different species is present in an image.

Here we introduce “cascade filtering” as an alternative approach to more traditional image classification projects. We introduce this term to describe the conversion of a complex task into a series of simple binary (i.e. yes or no) questions presented sequentially to volunteers. By reducing all projects to a similar set of problems, cascade filtering allows projects to deploy common approaches 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 (Figure 2).

Reducing a complex task into binary questions enables the integration of simpler machine learning for automated image identification at every step of the filtering process (Allwein et al. 2000). It also enables better portability to mobile devices because the classification task is simpler, thus requiring less screen “real-estate,” and the task can be completed via 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 via a mobile device and ability to integrate simple machine learning routines could provide benefits to efficiency that far outweigh the increase in classification counts.

However, altering project design can have wide-ranging and unanticipated impacts on the nature of volunteer participation. 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 spent more time classifying when they received high proportions of “blank” images, which took less time and effort to classify (Bowyer et al. 2016). Recent analysis by Spiers et al. (in prep) of volunteer patterns of behaviour in online citizen science identified that weekly releases of data in the *Supernova Hunters* project led to an age and sex biased community, and high levels of inequality in volunteer classification contribution, demonstrating the far-reaching consequences of subtle design choices. Thus, applying a cascade filtering approach rather than a more traditional survey could impact data quantity and quality, as well as the nature of volunteer communities contributing to a project.

In this paper, we compare classifications for two implementations of the same project: one using a ‘traditional’ design, which solicits a full set of answers, and the other utilizing a cascade design. While models based on gamification would suggest that the introduction of a repetitive, simple interface would increase participation (CITE), previous studies of complex citizen science projects have found that gamification of this kind may be problematic (Eveleigh et al. 2014). If volunteers are motivated directly by a desire to contribute to research, it seems possible that they will seek out experiences which seem to allow them to contribute most, providing all the information they can for each image.

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 images contributed by tourists. By providing volunteers with a choice of project workflows, we are able to evaluate differential effects of alternative workflow design on the project efficiency and accuracy, as well as effects on the volunteer community. The results have implications for researchers implementing online citizen science methodology, whether in-browser or in-app and, especially when running such a project in tandem with machine learning.

**Methods**

***Data Collection***

*Snapshots at Sea* was originally launched as a cascade filtering project on 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, which allow for high resolution population monitoring and conservation research (Katona and Whitehead 1981). 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. *Snapshots at Sea* has an associated discussion forum, referred to as *Talk*, where volunteers can engage with each other and the research team.

For this study, we integrated the binary questions into a more traditional “survey” design that allowed volunteers to select species from a list of multiple options, which is standard for ecology projects (e.g. Swanson et al. 2015). To ensure that images on each workflow were similar in content and appeal, we acquired a set of n = 13,264 images taken by cruise passengers with Fast Raft Ocean Safaris (http://fastraft.com), randomized them into two subsets of n = 6,632 images each, and assigned them to either the standard “Survey” workflow or the cascade filtering “Yes/No” workflow. Because each individual Yes/No question provides less information than the survey task, the cascade filtering approach ultimately requires more classifications to complete the same number of images. The difference in classification requirements depends on study design. This project required n = 95,195 classifications for the Survey workflow and n = 66,320 for the Yes/No workflow (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 (Figure 2). Each image was seen by five volunteers prior to retirement at each stage (although because volunteers could continue to classify images after the dataset was complete, some images received >5 classifications). 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 n = 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).

These thresholds were decided by the project lead during early data validation on *Snapshots at Sea* to rapidly filter unusable images or non-target species from the dataset while minimizing false negatives of images that could eventually be used to identify whales. Early validation revealed that images below 75% agreement on the presence of an animal contained poor-quality captures of animals that researchers preferred to exclude from the dataset, such as a barely-visible dorsal fin obscured primarily by a splash of water, or seabirds in the distance. In contrast, questions on the presence of a tail or the identification of a tail as belonging to a humpback had higher rates of false negative error among raw classifications. Thus, the lower percent-agreement requirements ensured that these potentially useable “borderline” images were included in the final dataset.

*‘Survey’ workflow*

The survey workflow provided 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 finished classifying on one image, they submitted identifications by selecting “Done,” and were then shown the next randomly-selected image.

Note that species-level information on other animals was not central to the *Snapshots at Sea* research goals of identifying, and thus was not collected on the Yes/No workflow. However, specifying the presence of other animals to category (bird, fish, or seal/sea otter) was generally straightforward, required minimal additional effort, and provided potentially useful information for future research. Because the Survey workflow task was more complex, effectively combining multiple Yes/No questions into a single task, images were shown to 10 volunteers before being retired from circulation; this also ensured sufficient classifications for species-specific sub-questions if not all volunteers agreed on the species present.

*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 familiar with 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 the now-retired, marine-focused project *Seafloor Explorer*, on the basis that these volunteers would likely be interested in this project.

When Q1 of the Yes/No workflow was completed, we activated Q2 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. No newsletters were sent for the launch of Q3 or Q4. 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 all *Snapshots at Sea* volunteers asking them to complete a survey about their demographic information and workflow preferences (See Appendix S1 for the full questionnaire and responses).

***Data analysis***

*Classification rates and volunteer engagement*

We retrieved all classifications made on *Snapshots at Sea* during the study and associated metadata, including volunteer *user name* (recorded as “not-logged-in” if an unregistered volunteer), the *date* and *time* of the classification (as recorded in the server logs), and the *device* (defined from the volunteer’s browser metadata) used to classify. We also tracked volunteer activity on the associated *Talk* forum. 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; data from unregistered volunteers were excluded from analyses of classification behavior (including duration and session length). 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 from the duration and session length analyses because 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 (excluding non-registered volunteers). Lorenz curves show 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 G = A/(A+B), where A is the area above the Lorenz curve and below the 45º line, and B is the area below the Lorenz curve). The Gini coefficient ranges from 0 to 1, where 0 represents completely equal contributions and 1 represents completely unequal contributions.

*Aggregation & Accuracy*

We evaluated agreement between volunteer answers and expert answers contributed by the project’s scientific lead, T.C. We calculated raw agreement as the number of classifications that agreed with the expert classification.

Because the high-levels of accuracy on Zooniverse projects rely on the aggregation of multiple volunteer classifiers, we also applied very simple aggregation algorithms to produce a single “consensus” answer for every image. For each binary question, 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, and 0.5 for Q3 and Q4, which are the same requirements for passing images into subsequent workflows.

While more sophisticated approaches exist for aggregating survey-style tasks (Swanson et al. 2014), we applied an analogous approach to the Survey workflow 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 as the Yes/No workflow.

**Results**

*Classification rates & volunteer engagement*

Volunteers completed the Yes/No workflow twice as quickly as the Survey workflow, contributing the 95,195 classifications needed in only 21 days, compared to the 41 days required to achieve the 66,320 classifications needed to complete the Survey workflow (Table 2, Figure 3). Classification rates following newsletters increased for both workflows, but were three times higher for the Yes/No workflow, which reached a maximum hourly classification rate of 3,000 classifications per hour, compared to 1,000 classifications/hour on the Survey workflow.

Daily per-volunteer contribution was higher on the Yes/No workflow (9.3 classifications per person per day) than on the Survey workflow (2.8 classifications per person per day). The Yes/No classifications were much faster to complete, ranging from 1.0 seconds (on mobile) to 3.7 seconds (on a computer), compared to 8.2 seconds (on a computer) to 10.7 seconds (on mobile) for the Survey workflow (Figure 4). These faster classification times resulted in overall increases in the mean number of classifications per session (n = 103 for Yes/No, n = 48 for Survey, Table 2), even without increasing session length. In fact, mean session lengths were similar for volunteers classifying on Yes/No vs. Survey tasks (12 minutes), median session length was much lower for Yes/No tasks (nYes/No= 3 minutes, nSurvey= 6 minutes), indicating that many volunteers actually spent less time on the Yes/No workflow than on the Survey workflow.

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 or *new* volunteers. Gini coefficients ranged from 0.762 (new users) to 0.745 (existing users) for the Survey workflow and 0.825 (new users) to 0.828 (existing users) on the Yes/No workflow. The difference in skewedness between the workflows was driven primarily by increases in the number of highly-active volunteers, with n = 24 volunteers who contributed over 1,000 classifications on the Yes/No workflow, compared to n= 11 volunteers on the Survey workflow (Figure 5b).

For the 31 days when both workflows were available on the Zooniverse, a similar number of volunteers contributed to both workflows as contributed to Yes/No only and Survey alone (Table 2). However, volunteers who contributed to both workflows classified many more images and participated more actively on *Talk* (14% of volunteers) compared to those who only classified on one workflow (4.7% Survery workflow volunteers; 3.7% Yes/No workflow volunteers). While the difference in *Talk* activity was significantfor those who participated on both workflows compared to only one workflow (p = X), there was no significant difference between activity from the Survey and Yes/No workflows (p = 0.75 in a two-tailed proportion test, package *stats*, function *prop.test*).

*Accuracy*

Accuracy was similar across workflows (Table 3). The accuracy of raw volunteer classifications varied dramatically across questions, ranging from 69 - 99% agreement with expert data. The aggregated answers were much more consistent, ranging from 93 - 100% agreement. It is notable that the error for the presence of an animal came almost entirely from false positives (e.g. the incorrect identification of an animal where none was present), and that most of these false positives represent images that technically contained an animal, but that the researchers considered “empty” because the animal was in the distance, out-of-focus, or completely obscured from view. On the Cascade workflow, false positives are quickly filtered out of the final dataset by subsequent questions. In contrast, error for the presence of a whale, tail or humpback derived primarily from false negatives (e.g. failing to identify a tail when present). The Survey workflow buffers against false negatives by allowing for the collection of data on follow-up questions on images that would otherwise be filtered out in the cascade filtering workflow, potentially allowing researchers the opportunity to re-evaluate their thresholds and recover data from false negatives, if desired.

We note that altering the number of required classifications per subject (currently set n = 5 for Yes/No and n = 10 for Survey) or the threshold agreement levels within the aggregation algorithms could improve overall accuracy, as could more sophisticated aggregation routines that incorporate concepts such as user-weighting. However, our goal here was not to improve the accuracy of this project, but rather to test whether there was a difference in accuracy between the different project designs.

*Follow-up survey*

We received n = 128 responses to the follow-up survey. According to the survey responses, *Snapshots at Sea* volunteers were twice as likely to be female as male, and 50% of volunteers had a university degree or higher (Figure 6, Supplementary Table 1). 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 for engaging volunteers in data interpretation. Here we demonstrate that this alternative approach to more traditional, complex citizen-science projects engages volunteers in both classification and discussion, quickly producing accurate data for scientific use while simultaneously providing opportunities to better leverage both mobile technology and machine learning. We recommend that researchers consider implementing cascade filtering for their own citizen science projects, but raise some points for consideration when deciding whether cascade filtering is appropriate for their project.

*Does cascade filtering work?*

Rapid production of data is critical to the success of most citizen science projects – the faster volunteers provide classifications, the sooner the researchers can generate new scientific knowledge. For *Snapshots at Sea,* cascade filtering provided a rapid route to data completion, with volunteers completing the Yes/No workflow in just half of the time it took to complete the Survey workflow (Figure 3). This increase was driven by higher rates of classification per volunteer, as opposed to a greater number of volunteers: they completed classifications in a fraction of the time (Figure 4) and contributed three times as many classifications per day (Table 1). We note that while reducing the number of classifications required on the survey task would have reduced time-to-completion, it would not have provided sufficient classifications for consensus on secondary tasks, as <100% of people typically agree on species IDs.

Importantly, this acceleration did not come at the cost of data quality: levels of accuracy were similar across both workflows (Table 3). Individual accuracy, however, varied dramatically across the questions within each workflow; accuracy was especially variable on the Survey workflow, ranging as low as 69% agreement on the presence of a humpback. This could suggest that the difficult tasks are easier to answer when isolated, perhaps because the task of learning a complex interface could interfere with the accuracy of the answers given (CITE). We should, however, note that while the cascade filtering workflow did not reduce data accuracy, it ultimately captured less information about each image than the Survey workflow, which collected follow-up information on images that did not have threshold agreement on the presence of a whale and also asked volunteers to classify non-target species present (see Figure 2). While the additional information gathered on the Survey workflow was not central to the research questions, it would provide greater flexibility for improving data accuracy through post-hoc analysis and aggregation.

In addition to classification counts and rates, volunteer engagement is important for the health of a citizen science project. Volunteer engagement is directly correlated with project success: projects that successfully engage the public in their research are more likely to produce science (Cox et al. 2015), and this deeper engagement can lead to scientific learning and changing attitudes towards science (Masters et al. 2016). Cascade filtering did not have a clear impact on volunteer engagement on *Talk,* the project-specific discussion forum where volunteers tend to discuss interesting images, share strategies for accurate classification, and inquire more broadly about the science behind the project. Volunteers who participated exclusively on the Yes/No workflow were just as likely to engage on the discussion forum as those who participated on the Survey workflow, but they made only one-third as many comments (Table 2).

These contrasting patterns could reflect that volunteers on the cascade filtering workflow had fewer questions about how to use the interface or that the project design itself discouraged volunteers from stopping in-between classifications to comment on *Talk.* Alternatively, this could reflect differences in the underlying volunteer communities: the follow-up survey revealed that volunteers who preferred the Yes/No workflow did so because it was quick and easy, whereas those who preferred the Survey workflow felt that they were contributing more information (Supplementary Table 1). Notably, volunteers who classified on *both* workflows were three times more likely to participate on *Talk* and commented an order of magnitude more frequently than those who participated on either workflow exclusively (Table 2). These volunteers were also more active classifiers in general (Table 2), mirroring patterns seen across Zooniverse projects in which higher classification activity increases the likelihood of volunteer engagement on *Talk* (CITE).

Ultimately, this experiment revealed that cascade filtering effectively engaged volunteers to classify images without imposing significant reductions in accuracy or volunteer engagement. The speed and popularity of the collaborative filtering approach likely derived from the simplicity of the task and the speed and ease with which volunteers were able to make a tangible contribution. Yes/No tasks were fast – three to ten times faster than Survey tasks (Figure 4) - and more predictable: because only one question was active at a time, volunteers knew whether they would be asked to identify the presence of an animal (generally easy) or identify a tail as belonging to a humpback (generally somewhat difficult). Thus, the cascade filtering approach may have allowed volunteers to “get into the flow,” which tends to increase the amount of time people spend on an activity, as seen with mobile phone games (CITE) and illustrated on other Zooniverse projects (Bowyer et al 2016). This was especially true for the Yes/No questions on the mobile swipe app, where volunteers could classify by merely swiping their thumb left or right. While a reluctance to “break the flow” could explain why volunteers comment less often from the Yes/No workflow, the increased classification activity seen on the cascade filtering workflow could eventually drive deeper, more sustained involvement on the project that would in turn drive higher rates of participation on *Talk* (as seen by CITE).

*Should you implement cascade filtering on your citizen science project?*

We recommend that research teams consider cascade filtering as a potential project design for their citizen science projects. Beyond engaging volunteers and producing rapid results, cascade filtering also produces standardized binary classifications that are significantly easier to parse and analyze. Thus, rather than designing project-specific algorithms that have to accommodate long, nested lists of complex data types, projects can leverage simple, generalizable code. The standardized format of binary classifications also enables easier integration of machine-learning routines at every stage within a cascade filtering workflow (Simpson et al. 2012, Marshall et al. 2016, Wright et al. 2017). This step would be critical for projects with very large data sets wishing to implement significant acceleration of classification results while retaining high levels of per-species classification completeness as well as purity (low levels of false negatives).

However, cascade filtering is not a panacea, and researchers should consider several limitations when deciding whether to implement this approach. First, we have only demonstrated the efficacy of cascade filtering for relatively simple projects with a small set of binary questions. With only four binary questions targeting a single focal species, *Snapshots at Sea* doubled the speed of data completion. However, it provided lower-resolution data than the survey task and expanding the cascade filtering workflow to capture the same resolution of secondary information would have dramatically reduced these time-savings. A crude extrapolation of the rates here suggests that projects would cease to see increases in the rate of completion by approximately eight cascade filtering questions, but the final time-savings would depend on the retirement rules (i.e. the number of volunteers who saw each image), the thresholds for passing images onto the next stage, and how many images are passed to subsequent questions.

Thus, cascade filtering might not be appropriate for projects with many different possible choices or for which the underlying data is not binary. For example, many traditional Survey-style projects (e.g. camera trapping studies such as [www.snapshotserengeti.org)](http://www.snapshotserengeti.org)) ask volunteers to differentiate from more than 50 different species and there are sometimes multiple species within an image; the sheer number of binary questions required to convert a standard survey project into an exclusively cascade filtering approach would outweigh any increases in classification activity. Alternatively, some projects in other disciplines use binary questions to calculate where an image falls along a spectrum of classification, and thus may require that all subjects be classified for all questions, or that very low thresholds between filtering workflows be used, reducing the efficiency of the cascade filtering approach.

Second, project design can impact the volunteer community (Spiers et al. in prep) and cascade filtering produced more unequal volunteer contributions than the Survey workflow (Figure 5). This was true regardless of whether or not volunteers had previously classified on *Snapshots at Sea*, suggesting that the skew resulted from the nature of the task and was not driven by volunteers seeking out what was familiar. Thus, changing the workflow design has a measurable impact on the volunteer community, though levels of skew on both *Snapshots at Sea* workflows were well within the bounds of other successful Zooniverse projects, which typically range from 0.7 - 0.9 (Cox et al. 2015, Spiers et al. in prep).

We should also note that while it seems obvious that a citizen science project aimed at engagement should seek to minimize inequality, such skew is not intrinsically bad. Examining the underlying volunteer contributions (Figure 5b) reveals that this inequality appears driven more by an increase in highly prolific classifiers than an increase in “visitors” who contribute very few classifications. In general, increasing the amount of time that volunteers participate on projects leads to a variety of benefits. For example, volunteer accuracy in well-designed projects tends to correlate with the number of classifications (e.g. Space Warps, Marshall et al. 2016, Snapshot Serengeti, Swanson et al. 2016). Increased participation also correlates with increased understanding of project-specific content, suggesting that people are actively learning about science through their participation (Masters et al. 2016), although changing the *way* in which volunteers spend their time on a project could change these outcomes in unintended ways.

Third, volunteers show clearly divergent preferences for different 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, easy, and immediately 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. These divergent preferences were not explained by differences across age, gender, or education level (Figure 6), but do highlight the diversity of preference in the volunteer community. While cascade filtering may, on one hand, lower the barrier to entry by providing a simple route to engagement (CITE), the binary questions might not be as engaging or as enriching in the long-term, potentially reducing long-term volunteer retention and learning. However, as this study only spanned a few months, we cannot directly compare the long-term impacts of cascade filtering on engagement or opportunities for learning.

Clearly, not all citizen science projects can or should attempt to implement an exclusively cascade-filtering approach. However, integrating both approaches into a project could simultaneously increase data throughput and provide multiple paths to engagement. The Zooniverse project builder ([www.zooniverse.org/lab)](http://www.zooniverse.org/lab)) makes it straightforward to build multiple routes to classification for a single project. Since completing this test on *Snapshots at Sea*, this strategy has been implemented via the introduction of two binary workflows on the *Panthera* conservation organization’s *Camera CATalogue* project ([www.cameracatalogue.org)](http://www.cameracatalogue.org)). Images are first passed through an “Empty or Not” workflow, and then to a “Vehicle or Not” workflow, before continuing on to the multi-species survey task. Images that are empty or contain only vehicles represent roughly half of the 30-million image dataset (Ross Pitman, *pers comm*). The binary filtering steps are retiring images ~10 times faster than the full survey task, potentially reducing the project’s time to completion by up to 40%.

While more work needs to be done to assess potential long-term impacts to the volunteer community, this study demonstrates that cascade filtering provides a rapid, efficient, and accurate alternative to traditional project design, and provides clear benefits for both research teams and the volunteer community.

**References**

Allwein, E.L., Schapire, R.E., and Singer, Y. (2000). Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. Journal of Machine Learning Research *1*, 113–141.

Bowyer, A., Maidel, V., Lintott, C., Swanson, A., and Miller, G. (2015). This Image Intentionally Left Blank: Mundane Images Increase Citizen Science Participation. In Human Computation and Crowdsourcing: Works in Progress and Demonstrations. An Adjunct to the Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing., p.

Cox, J., Oh, E.Y., Simmons, B., Lintott, C., Masters, K., Greenhill, A., Graham, G., and Holmes, K. (2015). Defining and Measuring Success in Online Citizen Science: A Case Study of Zooniverse Projects. Computing in Science & Engineering *17*, 28–41.

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., and Cox, A.L. (2014). Designing for Dabblers and Deterring Drop-outs in Citizen Science. In Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, (New York, NY, USA: ACM), pp. 2985–2994.

Fortson, L., Masters, K., Nichol, R., Borne, K.D., Edmondson, E.M., Lintott, C., Raddick, J., Schawinski, K., Wallin, J., Fortson, L., et al. (2012). Galaxy Zoo: Morphological Classification and Citizen Science, Galaxy Zoo: Morphological Classification and Citizen Science. In Advances in Machine Learning and Data Mining for Astronomy, Advances in Machine Learning and Data Mining for Astronomy, M.J. Way, J.D. Scargle, K.M. Ali, and A.N. Srivastava, eds. (CRC Press), p. 213.

Katona, S.K., and Whitehead, H.P. (1981). Identifying Humpback Whales using their natural markings. Polar Record *20*, 439.

Kosmala, M., Wiggins, A., Swanson, A., and Simmons, B. (2016). Assessing data quality in citizen science - preprint.

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., et al. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society *389*, 1179–1189.

Marshall, P.J., Verma, A., More, A., Davis, C.P., More, S., Kapadia, A., Parrish, M., Snyder, C., Wilcox, J., Baeten, E., et al. (2016). Space Warps – I. Crowdsourcing the discovery of gravitational lenses. Mon Not R Astron Soc *455*, 1171–1190.

Masters, K., Oh, E.Y., Cox, J., Simmons, B., Lintott, C., Graham, G., Greenhill, A., and Holmes, K. (2016). Science Learning via Participation in Online Citizen Science. ArXiv:1601.05973 [Astro-Ph, Physics:Physics].

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Packer, C., and Clune, J. (2017). Automatically identifying wild animals in camera trap images with deep learning. ArXiv:1703.05830 [Cs].

Simpson, E., Roberts, S., Psorakis, I., Smith, A., Simpson, E., Roberts, S., Psorakis, I., and Smith, A. (2012). Dynamic Bayesian Combination of Multiple Imperfect Classifiers, Dynamic Bayesian Combination of Multiple Imperfect Classifiers. ArXiv E-Prints, ArXiv e-Prints arXiv:1206.1831, arXiv:1206.1831.

Simpson, R., Page, K.R., and De Roure, D. (2014). Zooniverse: observing the world’s largest citizen science platform. (ACM Press), pp. 1049–1054.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. (2015). Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. Scientific Data *2*, 150026.

Swanson, A., Kosmala, M., Lintott, C., and Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology *30*, 520–531.

Tyfield, D., Lave, R., Randalls, S., and Thorpe, C. (2017). The Routledge Handbook of the Political Economy of Science (Taylor & Francis).

Willett, K.W., Lintott, C.J., Bamford, S.P., Masters, K.L., Simmons, B.D., Casteels, K.R.V., Edmondson, E.M., Fortson, L.F., Kaviraj, S., Keel, W.C., et al. (2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society *435*, 2835–2860.

Wright, D.E., Lintott, C.J., Smartt, S.J., Smith, K.W., Fortson, L., Trouille, L., Allen, C.R., Beck, M., Bouslog, M.C., Boyer, A., et al. A transient search using combined human and machine classifications. Mon Not R Astron Soc.

**Acknowledgments**

We are enormously grateful for the effort of the *Snapshots at Sea* volunteers, without whom this science would not be possible. We would like to thank the Zooniverse development team for their work in envisioning and creating the project builder platform.

This work was supported in part by [CHRIS, insert grants], and the U.S. National Science Foundation Award IIS-1619177.

**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.



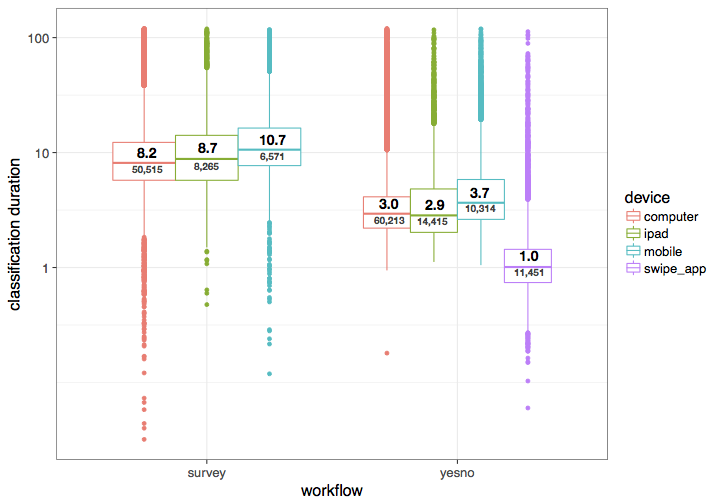
**Figure 2:** *Snapshots at Sea* workflows, including questions, retirement rules, images classified, and thresholds for passing images onto subsequent steps. Additional details of each workflow are described in the text under *Methods:Data Collection.*



**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 (n = 66,320 for the Survey workflow and n = 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 (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:** Results from the follow-up survey sent to volunteers who participated during the experiment. The full set of answers is given in Appendix 2.



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1:** Summary statistics for each workflow. Newsletters were calls to action sent to volunteers by email; the first three newsletters promoted both workflows, while the final email promoted the remaining survey workflow. Days to complete indicates the number of calendar days required to retire all 6,632 images. Gini coefficients reflect the amount of inequality in volunteer contributions, calculated as A/(A+B) where A is the area between the Lorenz curve and the line y = x, and B is the area between the Lorenz curve and the x and y axes. | | | | | | | | |
| **Workflow** | **Total images** | **Classifications received (vs. needed)** | **News-letters**  **sent** | **Registered volunteers** | **Days to completion** | **Gini coefficient** | **Classifications per volunteer per day** | **Classification duration (seconds)** | |
| **Survey** | 6,632 | 66,405 (66,320) | 4 | 588 | 41 | 0.756 | 2.75 | 8.47 | |
| **Yes/No** | 6,632 | 97,130 (95,195) | 3 | 473 | 22 | 0.830 | 9.33 | 2.83 | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2:** Volunteers were classified as those who participated on both workflows, the Survey workflow only, and the Yes/No workflow only. Data are limited to the period of time when both workflows were active, from June 6 2017 to July 5 2017. Proportion commenting is calculated as the proportion of volunteers classifying on each workflow type who also commented on the Talk forum. Total Talk comments reflects the number of comments made by users in each category. Session length and classifications per session are given as mean values with median values in parentheses, and were calculated for sessions in which volunteers classified on the specified workflows (regardless of their activity in other sessions). | | | | | | |
| **Workflows classified** | **Volunteers** | **Classifications made** | **Percent commenting** | **Total *Talk* comments** | **Mean and (median) session length (minutes)** | **Mean and (median) classifications per session** | |
| Both | 229 | 86,159 | 14.0% | 246 | 20 (9.6) | 110 (51) | |
| Survey only | 234 | 15,616 | 4.7% | 36 | 12.9 (6.1) | 48 (22) | |
| Yes/No only | 243 | 22,756 | 3.7% | 11 | 12.2 (3.4) | 103 (22) | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3:** Volunteer agreement with expert classifications. Raw agreement rates indicate the percentage of individual volunteer classifications that agreed with the expert classification for any given image. Aggregated agreement rates reflect the percentage of images that were correctly identified when compared with expert classifications. | | | | | | | | |
|  | **Survey** | | | | **Yes/No** | | | | |
| **Workflow** | **# images** | **# class-ifications** | **Raw % agreement** | **Aggregated % agreement** | **# images** | **# class-ifications** | **Raw % agreement** | **Aggregated % agreement** | |
| Presence of an animal | 55 | 407/419 | 97.14 | 96.36% | 30 | 477/551 | 86.56 | 93.33% | |
| Presence of whale/dolphin (given presence of animal) | 51 | 377/389 | 96.92 | 96.07% | 30 | 910/916 | 99.34 | 100.00% | |
| Presence of tail (given presence of whale/dolphin) | 41 | 270/309 | 87.38 | 97.56% | 30 | 384/425 | 90.35 | 96.67% | |
| Presence of humpback (given presence of tail) | 16 | 86/123 | 69.92 | 100.00% | 30 | 287/334 | 85.93 | 100.00% | |