Teaching Citizen Scientists to Categorize Glitches using Machine Learning Guided Training

Corey Jackson[a,b,*], Carsten Østerlund[a], Kevin Crowston[a], Mahboobeh Harandi[a], Sarah Allen[c], Sara Bahaadini[d], Scott Coughlin[e], Vicky Kalogera[e], Aggelos Katsaggelos[d], Shane Larson[e], Neda Rohani[d], Joshua Smith[f], Laura Trouille[c], Michael Zevin[e]

[a]*Syracuse University School of Information Studies, Syracuse, NY 13244 USA*
[b]*School of Information, University of California, Berkeley, Berkeley CA 94720 USA*
[c]*Adler Planetarium, Chicago, IL 60605 USA*
[d]*Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 606201 USA*
[e]*Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) and Dept. of Physics and Astronomy, Northwestern University, 2145 Sheridan Rd, Evanston, IL 60208 USA*
[f]*Department of Physics, California State University, Fullerton, CA 92831 USA*

Abstract

Existing literature points to scaffolded training as an effective yet resource-intensive approach to help newcomers learn and stay motivated. Experts need to select relevant learning materials and continuously assess learners' progress. Peer production communities such as Wikipedia and Open Source Software Development projects face the additional problem of turning volunteers into productive participants as soon as possible. To address these challenges, we designed and tested a training regime combining scaffolded instruction and machine learning to select learning materials and gradually introduces new materials to individuals as their competences improve. We evaluated the training regime on 386 participants that contribute to Gravity Spy, an online citizen science project where people are asked to categorize glitches to assist scientists in the search for gravitational waves. Volunteers were assigned to one of two conditions; (1) a machine learning guided training (MLGT) system that continuously assesses volunteers skill level and adjusts the learning materials or (2) an unscaffolded training program where all learning materials were administered at once. Our analysis revealed

*Corey Jackson
Email address: coreybjackson@berkeley.edu (Corey Jackson)
URL: www.coreybjackson.com (Corey Jackson)

that volunteers in the MLGT condition were more accurate on the categorization task (an average accuracy of 90% vs. 54%), executed more tasks (an average of 228 vs. 121 tasks), and were retained for a longer period (an average of 2.5 vs. 2 sessions) than volunteers in the unscaffolded training. The results suggest that MLGT is an effective pedagogical approach for training volunteers in categorization tasks and increases volunteers' motivation.
Keywords: citizen science, experiment, training, online communities, Zooniverse, user studies, scaffolding, learning

1. Introduction

Peer production platforms like Wikipedia, open-source software (OSS), and citizen science projects rely on a steady stream of newcomers. In Wikipedia, new editors compose and edit articles and new software developers write and debug software code in OSS. Citizen scientists contribute to research by collecting, analyzing, or interpreting data Bonney et al. (2009). Peer production platforms need to facilitate learning among newcomers while maintaining them as motivated and productive participants to be successful (Kraut et al., 2012). This is not an easy task as learning, motivation, and productivity easily goes counter to one another. Some peer production platforms provide explicit training to help volunteers learn about the community, its technical infrastructure, the tasks to be completed, and how best to contribute (Malinen, 2015; Ducheneaut, 2005). But, if the training takes a lot of effort, it might demotivate some volunteers and take away from the time they could serve as productive members. In contrast, a short training regime could lead to under-qualified participants that find the task too hard and lose interest. The training and the tasks should exist in the zone of proximal development (Vygotsky, 1980) where volunteers can execute tasks with minimal assistance while remaining productive members of the community (Brown & Duguid, 1991; Lave & Wenger, 1991; Lave, 1991; Downes, 2006).

The online learning and e-learning literature point to scaffolded instruction as an effective approach to instruct newcomers (Rienties et al., 2012; Molenaar et al., 2012; Tuckman, 2007; Haythornthwaite, 2014; Jones & de Laat, 2016; Østerlund & Carlile,

2005; Luckin, 2008; Haythornthwaite & Andrews, 2011; Downes, 2006). Dickson et al. (1993) describes scaffolded instruction as "the systematic sequencing of prompted content, materials, tasks, and teacher and peer support to optimize learning." Scaffolding allows instructors to accommodate individual student needs by providing among others tailored assistance, feedback, and actively diagnosing student needs and comprehension (Hogan & Pressley, 1997).

The added benefit of scaffolding training comes at a cost often difficult for peer production communities to afford. First, developing and making the training regime can be resource-intensive. It often requires experts to select and organize training materials and continuously evaluate learners as they move through the scaffolded process and require gradually more challenging materials (Ford & Geiger, 2012). On many projects, one finds a short supply of experts. For instance, citizen science projects often struggle to maintain the continuous attention of science teams. Second, training takes away from the time left for productive activities. To avoid these costs, organizations sometimes apply on the job training, by having newcomers work and receive feedback simultaneously (Van Maanen & Schein, 1979). This requires a division of labor affording tasks at a range of difficulties. Thirds, too difficult or easy tasks demotivate volunteers. Newcomers to peer production communities come with a range of abilities, and they learn as they go. Trying to fit the task to the individual member's skill level requires a flexible division of labor and the attention of experts.

To address these challenges, we propose an innovative computer-supported approach labeled machine-learning-guided training (henceforth referred to as MLGT) intended to scaffold the learning of categories needed to execute a task and assess learning of categories computationally. In this approach, a machine-learning classifier selects genuine tasks (i.e., tasks currently being worked on in the system) to provide newcomers so that newcomers are introduced to new categories of data gradually rather than all at once. Whether or not a machine-learning-guided training regime using genuine tasks can be effective is an open question. It could be that the selected categories are not appropriate for training, meaning that the ML-chosen categories are no better than random tasks for helping volunteers learn. It may also be the case that the tasks are not of the appropriate level of difficulty, leading to poor retention of volunteers. In

this paper, we ask: *What effect does a machine-learning-guided training regime have on retention and contribution?* To answer this question, we evaluated the MLGT in an online field experiment against unscaffolded training. We do so in the context of a citizen science project where volunteers help analyze data in the search for gravitational waves. Our results reveal that compared to the unscaffolded training, volunteers in the MLGT condition were more accurate, executed more classification tasks, and were retained for a longer period.

1.1. Citizen Science

Citizen science includes scientific research projects that rely on contributions from members of the general public (i.e., citizens in the broadest sense of the term). There are several kinds of citizen-science projects: some have volunteers collect data, while others, including the one we examine in this paper, have volunteers analyze preexisting data-sets Bonney et al. (2009). The interactions between volunteers and the project organizers are mediated over the Internet, i.e., on a digital platform that accepts contributed data or that presents data to be analyzed and collects volunteers' classifications (e.g., Zooniverse), making citizen science a form of peer production.

Volunteers typically execute small micro-tasks, that is, granular units of work easily handled by amateurs in a short time period. For example, in Galaxy Zoo, a citizen science project dedicated to helping astrophysicists analyze galaxies, questions about the shape and emergent properties of galaxies are posed. The results contribute to scientific research projects. Other classification tasks include transcribing ship logs (i.e., OldWeather), counting and labeling and describing the behaviors of chimpanzees (i.e., Chimp & See), and recording the presence of exotic nanoparticles (i.e., Higgs Hunters).

Developing effective citizen science training materials face the same core challenges as many other online production communities, i.e., help volunteers (1) learn, (2) stay motivated, and (3) productive (Mugar et al., 2014). Yet, one needs to see these issues in the context of the particular demographic and motivational forces characterizing citizen science.

First, a small number of studies mographics represented across citizen science does

not resemble the overall population or online user demographics (Pandya, 2012; Masters et al., 2016). The volunteer population tend to bee predominately male (Raddick et al., 2010; Estrada et al., 2013), well educated (46 percent of Foldit players indicated having an undergraduate degree (Curtis, 2015) and Raddick et al. (2013) found that Galaxy Zoo participants were more educated than the general online population), and come from English speaking Western nations (Raddick et al., 2013).

Second, one finds a variety of motivations present among citizen science volunteers including a desire to contribute to science Raddick et al. (2013); Lee et al. (2018), interacting with the site (Reed et al., 2012), social engagement (Reed et al., 2012), curiosity (Curtis, 2015), and intellectual challenge (Curtis, 2015). Examining motivation at the project level, one finds a variety of motivational drivers. (Raddick et al., 2013), who surveyed 10,000 volunteers in Galaxy Zoo, identified twelve categories, including the desire to learn, discover, social interaction, help, contribute and use the project as a resource for teaching, the beauty of the images, fun, largeness of the universe, interest in the project theme, astronomy, and scientific focus. One expects the find a similar breadth of motivational drivers in other projects.

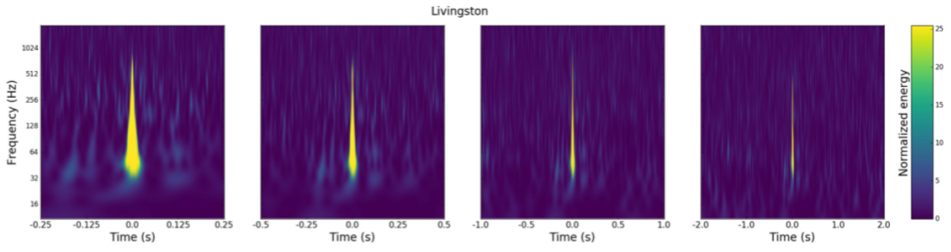
1.2. Gravity Spy Project

We carried out our experiment in Gravity Spy (Zevin et al., 2017; Bahaadini et al., 2018, 2017), an online citizen science project hosted on the Zooniverse platform (Simpson et al., 2014). Gravity Spy leverages human classification and machine learning to aid the Laser Interferometer Gravitational-wave Observatory (LIGO) collaboration in its search for gravitational waves. Gravitational waves are extremely faint distortions in the fabric of space created by astronomical events such as merging black holes or neutron stars. Astrophysicists use Interferometry to detect gravitational waves by bouncing lasers off mirrors to look for small changes in the distance the light traveled. However, the technique is highly sensitive to non-gravitational wave disturbances, e.g., from terrestrial interference or internal faults or interactions in and around the detector. Glitches are produced in a wide variety of time-frequency-amplitude morphologies and occur hundreds or thousands of times a day (depending on the threshold used). They are problematic because they can obscure or even masquerade true gravitational-wave

signals, reducing the efficacy of the search.

The Gravity Spy project couples human volunteers with machine learning to develop a catalog of glitches having similar morphological characteristics that allow researchers to focus their search for glitch sources in their effort to improve the detector and future data analysis. Data from the interferometer are presented to volunteers in a visual representation of the strength of the signal at different frequencies over time, called a spectrogram. Present in some spectrograms are transient, non-Gaussian noise features known as glitches. Figure 1 shows spectrograms of two glitches: on the top a blip and on the bottom a whistle. The four images for each glitch cover different durations, to show the glitch in detail (on the left) or in context (on the right). While computer vision algorithms perform well at the categorization task for known classes of glitch, there are many that do not fit these classes, so human classifiers are still needed to examine the spectrograms.

(a) Blip glitch



(b) Whistle glitch

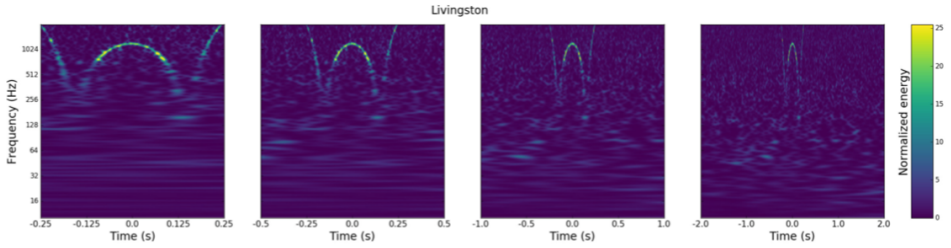


Figure 1: An example of various morphologies of a blip glitch (top) and a whistle (bottom).

Human volunteers contribute to Gravity Spy by classifying the spectrograms into one of the families of known glitches or “None of the Above”. There is also a possibility of new families of glitches being discovered as the detector is altered and improved.

Shown in Figure 2 is the classification interface. On the left of the interface, a spectrogram is presented, with time on the x-axis and frequency on the y-axis. The intensity of the glitch is represented by the color appearing in the spectrogram from blue to yellow. On the right are the glitch classes from which a volunteer can select to classify the glitch. Each spectrogram is analyzed by multiple volunteers, and a consensus glitch class is supplied to them during the data transfer to improve the quality of the data.

Classifying spectrograms is a perceptual categorization task and, more specifically, what Ashby & Maddox (2005) calls an information-integration category learning task, where two or more dimensions of a stimulus are considered prior to making a decision. In Gravity Spy, for example, volunteers might examine the duration of a glitch, its frequency, and its morphological features in order to arrive at a decision. However, instructors often struggle to verbally describe perceptual dimensions making it difficult to develop training materials needed to teach information integration tasks.

For glitch classifying in Gravity Spy, the information integration task structure poses additional constraints on the effectiveness of traditional training regimes. First, feedback on the task is crucial to improve learning. Without feedback, learners will be left without reasoning as to why a particular spectrogram was assigned to a category. Second, category bounds are of concern. That is, the delineation of some categories may be opaque for some types of glitches such that distinguishing them may be impossible for learners.

2. Theory

Gravity Spy was designed drawing on theory to increase both volunteer learning and motivation. By learning, we mean specifically learning to do the classification task accurately. The literature on optimal learning strategies for information categorization tasks suggests that procedural learning, feedback, and timing are important. Below, we describe how these structures homogenize in Gravity Spy.

2.1. Learning to Classify Glitches

The Gravity Spy system integrates four approaches to learning to promote accurate glitch classification. Of these four, the first three (explicit training, feedback, and

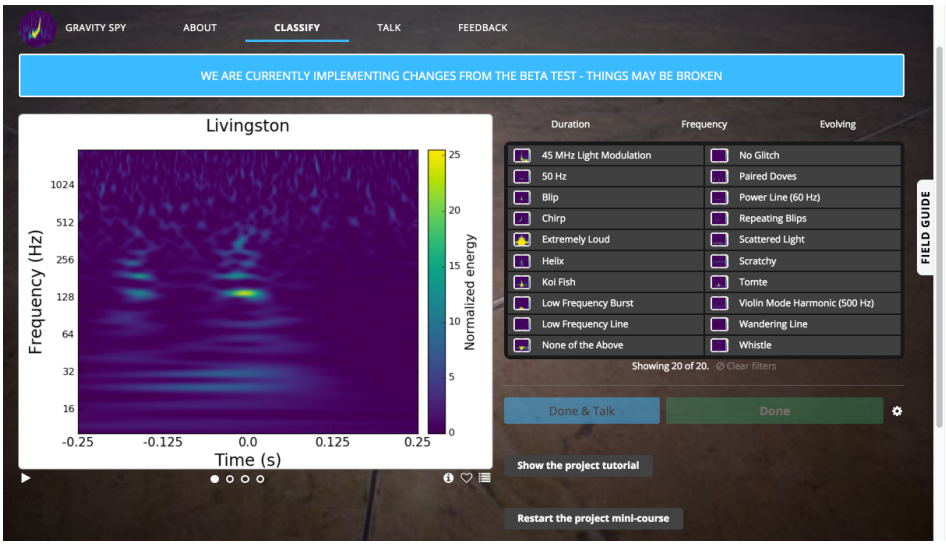


Figure 2: An example of the Gravity Spy classification interface. Volunteers categorize the glitch in the spectrograph on the left by selecting the glitch classes it most resembles from the list on the right.

presentation of prototypes and exemplars) can be found in many crowd-sourcing and citizen-science projects. The fourth approach (scaffolding supported by ML) adds a new dimension to training and is the focus of this study.

2.1.1. Explicit Training

On-line production community sites typically provide a brief introduction to the project that explains its goals and tasks. Citizen-science projects, in particular, provide training on the scientific tasks, how to interpret images, and how to use the classification interface. In the Gravity Spy, training is provided as a pop-up when a volunteer first starts the classification task and be revisited at any time via a link on the classification interface titled “Show the project tutorial”.

2.1.2. Feedback on Classification

Feedback on performance is critical in the learning process (Corbalan et al., 2009; Leutner, 1993; Moreno & Valdez, 2005; Easterday et al., 2017; Goldin et al., 2017). The Gravity Spy system, therefore, has beginning volunteers classify some glitches from a gold-standard dataset (i.e., glitches previously classified by members of the sci-

ence team). Knowing the correct classification makes it possible to give the volunteers feedback on the correctness of their classifications and to assess their accuracy in classifying those classes of glitch. As an added benefit, Corbalan et al. (2009) found that when feedback was provided for participants on their performance, they were more motivated to contribute than when feedback was not provided.

2.1.3. Prototypical and Exemplary Glitches

Cognitive theories suggest that people learn perceptual categories through exposure to prototypes and exemplars of known categories. Prototypes serve as a heuristic: an average representation of an entire category (Kim & Murphy, 2011). Prototypes may help learners categorize new stimuli by severing as a reference to the prototype. Stimuli, which match the prototype are said to be a member of the category. Another form of category learning emerges from the presentation of many references for the category or exemplars (Kulatunga-Moruzi et al., 2011). Exemplars function as multiple examples of a category and should exist across the spectrum of possible stimuli to be a part of the category. Learners may use these previously exposed examples as references for categorizing new stimuli.

The information integration tasks volunteers are asked to execute map well to the use of exemplars and prototypes to learn perceptual categories. When individuals are asked to generalize a category, they evaluate several characteristics and weight each of these characteristics (Jones et al., 2005; Nosofsky, 1986; Shepard, 1987; Sinha & Russell, 2011). That is, individuals, decide whether an image belongs to a category depending on how much the image is similar to or different from the prototypes and exemplars in certain characteristics and the importance of the characteristics (i.e., weights). As individuals are exposed to images, they update the weights for the stimuli characteristics. Therefore, to support the learning of image classification, volunteers should be continuously provided with good prototypes and exemplars of the classes.

Gravity Spy presents prototypical and exemplary images of glitches to volunteers in two ways. First, the classification interface shows volunteers prototypical instances of the various classes to guide their selection. When a class is selected, a larger image of a prototypical example and a brief description are displayed to reinforce the

exemplar. Second, Gravity Spy, like many Zooniverse projects, provides a field guide, with prototypical glitches, several exemplars, and discussions of the kinds of data to be classified. The field guide can be accessed in the classification interface.

2.2. Scaffolded Learning

The zone of proximal development emphasizes the need to adjust learning opportunities to the learner's current abilities (Engeström, 2014). Optimal learning opportunities exist at the intersection of independent learning and achievement with guidance and encouragement from a skilled partner (Vygotsky, 1980). Thus, training should exist in the zone of proximal development. Studies of learning through legitimate peripheral participation similarly suggest that learners gradually expand their access to central activities (Lave & Wenger, 1991). The emerging literature on learning in online settings specifically and eLearning more broadly (Haythornthwaite, 2014; Jones & de Laat, 2016; Østerlund & Carlile, 2005; Luckin, 2008; Haythornthwaite & Andrews, 2011; Downes, 2006) suggests that learning emerges as participants gradually expand their engagement with a task. Bringing these concerns together, the concept of scaffolding suggests the importance of carefully sequencing participants' learning opportunities (Johri & Yang, 2017).

Volunteers' progress in online production communities has been analyzed and supported by these perspectives. In a study of Wikipedia, for instance, Bryant et al. (2005) showed how novices often start out by reading other's articles before making their initial contributions and gradually access more tasks crucial to the community. Preece & Shneiderman (2009) similarly suggested that participants in peer-production sites move from "readers to leaders". The Fold It citizen science system (Cooper et al., 2010) provides a series of increasingly challenging tasks to help newcomers learn how to fold protein cells so they may take on more serious tasks in the future. In other learning environments, we see increased calls for scaffolded mechanisms to deliver formative feedback (Rose & Ferschke, 2016) to guide users in the learning process.

2.3. Motivation to Contribute

As citizen-science projects depend entirely on the contributions of volunteers, volunteer motivation has been a consistent topic of research, and researchers have iden-

tified a range of motivations for citizen science. For example, surveys and interviews show that citizen science volunteers are motivated to participate in projects by the opportunity to make a contribution to science (Brossard et al., 2005; Land-Zandstra et al., 2016a,b; Raddick et al., 2010). Accordingly, presenting volunteers with authentic tasks should be more motivating than providing a standalone training program. Following Crowston & Fagnot (2018), we consider the motivation for an initial contribution and for sustained contribution separately.

3. The Machine Learning Guided Training (MLGT)

In the Gravity Spy system, we implemented an innovative training program using real tasks selected by a machine-learning (ML) system to scaffold the training materials, in such a way that a volunteer's current competencies in the classification task match the difficulty of the categorization task. We describe the MLGT components below.

3.1. Machine Learning

Selection of learning tasks is a common issue for scaffolded instruction (Kicken et al., 2008; Winn, 2007). Tasks must be within a learner's zone of proximal development and promote individual learning to be effective. We chose to utilize machine learning to select tasks. The process is described in detail in (Zevin et al., 2017; Bahaadini et al., 2018, 2017). An *image classifier* is maintained on the system which processes each spectrograms using a Convolutional Neural Network (CNN) classifier that outputs a vector indicating the confidence that the glitch is of a particular class. Each image has a probability of belonging to each of the 20 known classes. Based on these image classifier scores, spectrograms are then assigned to a workflow based on confidence thresholds with high confidence spectrograms being assigned to early workflows, and more difficult to machine classify spectrograms assigned to higher workflows.

3.2. Guided Training

In the MLGT, newcomers begin at Level 1 (shown in Figure 3) where they are exposed to gold standard spectrograms and spectrograms requiring human annota-

tors. Prior to classifying, volunteers are shown a short tutorial introducing them to the project and teaching them how to execute a classification. As volunteers classify gold standard spectrograms, the system maintains a *citizen score*, which is a real-time assessment of each volunteer's performance classifying data; the score increases as volunteers agree with gold-standard data decreases when responses diverge. Volunteers periodically receive feedback about their current abilities to identify glitches by providing answers to gold data. Gold data are spectrograms that have been assessed by the science team. If the volunteer submits the same answer as the expert, they see a message that reads *"Good work! When our experts classified this image, they also thought it was a Blip!*" (or whichever class was chosen). If the answer is incorrect, the message reads *"You responded Whistle. When our experts classified this image, they labeled it as a Blip.*" The citizen score is used to determine whether a volunteer should be promoted to the next level. The MLGT training shows spectrograms rated by the machine learning component having the highest confidence of classification are assigned to the first workflow (or Level 1). For instance, only blips and whistles with high machine confidence scores are shown in Level 1; once a volunteer is promoted to Level 2, the confidence threshold is relaxed for blips and whistles and volunteers are introduced to 4 new classes: koi fish, power line, violin mode having high confidence. The levels and new glitch classes are shown in Table 1.

Once volunteers have completed all rounds of training, introducing the classes of glitches, they are considered fully qualified and are given glitches to classify at varying levels of ML confidence in all known classes or even glitches for which the ML has no good score from the image classifier, thus contributing to novel areas of research in the project. And as the ML can be wrong, it is still useful to the project to have a human

Figure 3: Levels for the Gravity Spy project. When the buttons are shaded (e.g., Level 4: Black Hole Merger and Level 5: Universe Cosmic Background) a volunteer cannot access the level.

judgment even for glitches for which the ML reports high confidence. For example, when a new class of glitch appears in the data, the ML will attempt to classify them as one of the known classes. It has happened that the new glitches are confused with an existing class, resulting in incorrect ML classifications with high confidence. These classification errors can be corrected even by new volunteers. For instance, in the same Gravity Spy classification task, (Crowston et al., 2019) found volunteers and the machine learning agreed 91% of the time on the classification of glitches.

	Glitch Name
Level 1 (3)	blip, whistle, none of the above
Level 2 (6)	blip, whistle, koi fish, power line, violin mode, none of the above
Level 3 (10)	blip, whistle, koi fish, power line, violin mode, chirp, low frequency burst, no glitch, scattered light, none of the above
Level 4 (20)	blip, whistle, koi fish, power line, violin mode, chirp, low frequency burst, no glitch, scattered light, helix, 45Mhz light modulation, low frequency noise fluctuations, paired doves, 50hz, repeating blips, scratchy, tomte, wandering line, extremely loud, none of the above

Table 1: Gravity Spy glitch classes by training level.

We expect the MLGT program to support better learning (i.e., to help volunteers become more accurate at classifying) for two reasons. First, because the ML has high confidence in the classification of the glitches, it is most likely that they are of the identified class and so will be exemplary glitches that will help the volunteer to learn how to identify that class. Second, focusing attention initially on just a few classes enables volunteers to master those classes before adding complexity (i.e., staying in the volunteer's zone of proximal development).

4. Hypotheses

The first three types of learning support described in Section 2, i.e., tutorials, feedback, and the inclusion of prototypical and exemplary images in the interface are standard in online citizen-science projects, but the fourth approach (scaffolded introduction to categories) is novel for citizen-science projects. Furthermore, while scaffolding is a well-accepted approach, implementing scaffolding by using ML to select among real tasks is, as far as we know, novel. Assessing the success of the scaffolded machine learning guided training (MLGT) approach is the focus of the experiment reported in this paper. We hypothesize:

H1: Volunteers who go through the MLGT will be more accurate in their classifications than volunteers who do not go through the MLGT.

Citizen-science projects, like most online production communities, exhibit a highly-skewed distribution of contribution: a few volunteers contribute a lot while many contribute only a little Sauermann & Franzoni (2015). Indeed, many new visitors to a project do not contribute at all but rather leave before making a classification. We hypothesized that exposure to complex task causes volunteers to feel overloaded and discouraged from contributing. For example, the full Gravity Spy interface offers 22 options (increased from the original 20), some with fairly subtle distinctions. A new user could feel unable to perform the task accurately or at all. This problem is not unique to Gravity Spy: the Zooniverse Snapshot Serengeti project, for example, asks volunteers to identify animals shown in images into one of 54 species, many of which would be unfamiliar to a novice. We expected that the MLGT, with its scaffolded design introducing volunteers to the classes a few at a time, will be less challenging and thereby more inviting. We therefore hypothesize:

H2: Initial exposure to fewer glitch categories in the MLGT will motivate more volunteers to provide an initial classification than exposure to all glitch categories in the non-MLGT.

We expect a scaffolded approach to presenting new tasks will motivate volunteers.

The system is designed to appeal to volunteers' sense of accomplishment. The initial Gravity Spy page (Figure 3) shows all of the training levels, but volunteers can only access the ones they have "unlocked" by successfully completing the lower levels (Note that volunteers are free to choose to work on any of the unlocked levels, not just the highest one.). The system also provides an encouraging messaging when mastery at the current level is achieved, and the next level is unlocked. In a sense, the MLGT contains elements of gamification through the existence of levels of possible accomplishment (Morschheuser et al., 2017), which we expected to motivate volunteers to continue to contribute. Iacovides et al. (2013); Bowser et al. (2014) propose that gamification can be an effective motivator for citizen scientists. We therefore hypothesize:

H3: Volunteers who go through the MLGT will contribute more classifications than volunteers who do not go through the MLGT.

H4: Volunteers who go through the MLGT will contribute for a longer period of time than volunteers who do not go through the MLGT.

5. Experiment Design

To test the hypotheses developed above, we conducted a randomized controlled online experiment in the Gravity Spy project.

5.1. Procedure

The experiment tested the impact of the MLGT on volunteer accuracy and contribution. During the experiment, volunteers who visited the project site were randomly assigned to the treatment condition and the other half to the control condition. Subjects for the experiment were volunteers who joined the Gravity Spy project during the experimental period from 30 October 2016 to 19 December 2016. When Zooniverse volunteers created an account or logged into their account for the first time after the experiment launched, they were randomly assigned to either the control or to the treatment group. Volunteers retain this assignment when they visit the project on future sessions. To assess the impact of the treatment on a volunteer, it is necessary that

they receive the treatment from their initial interaction with the Gravity Spy system. Therefore, we did not include data from volunteers who had already contributed to the Gravity Spy system before the start of the experiment.

Treatment: Volunteers who were assigned to the treatment received the scaffolded MLGT. When first creating an account, volunteers are shown the tutorial, which consists of five pages (406 words). An example of the pop-up tutorial is shown in Figure 4. The tutorial is overlayed on the classification interface, is self-directed, and volunteers can exit the tutorial on any page. The tutorial contains descriptions of the project, the task, what functions different buttons on the interface perform. The estimated read time, calculated based on the average reading speed of 200 words per minute (wpm)

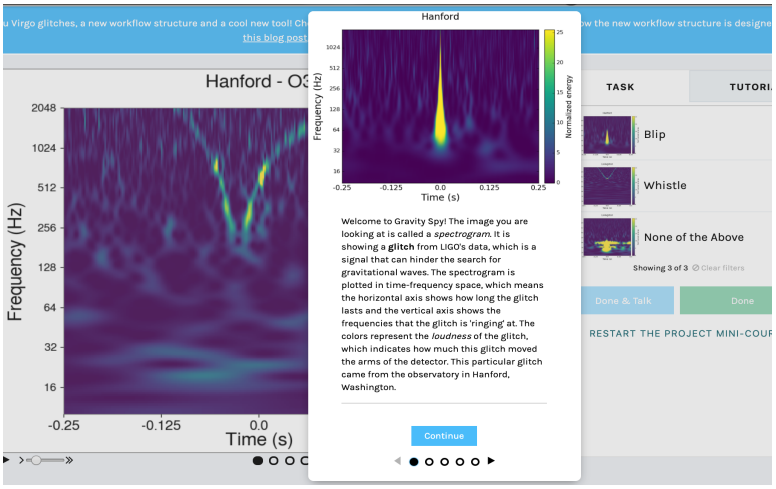


Figure 4: An example of the tutorial that volunteers in the control group are shown.

Newcomers in the treatment begin in the Level 1 workflow, where they are presented with glitches to classify that are expected to be of one of only two distinctive classes– blips and whistles in the current system—and given those two choices or "none of the above" in a simplified version of the classification interface. All spectrograms have been assessed by the (CNN) classifier, and only those having high likelihood (approximately $> 90\%$) of being categorized as blip or whistle by the classifier are assigned to the Level 1 workflow. Periodically, volunteers are administered gold-standard data to assess their performance and after a period of time when a volunteer has classified

a sufficient number of gold-standard data to evaluate their performance the promotion algorithm decides whether to promote the volunteer to the next level or keep them in the current level and administer more training. At the next level, the volunteer receives additional training with information about new glitches at that level in addition to prototypical categories. Table 1 shows the Gravity Spy workflow levels with the number of glitch class options and the names of the glitch categories presented in each level.

Control: Upon creating an account, volunteers assigned to the control condition were directed to a tutorial, again introducing them to the project and the classification task. The tutorial volunteers receive a slightly modified version of the one presented in the MLGT. The modification needed to the level was to expand the tutorial to include content included in the tutorials from lower levels, to ensure that the tutorial content was the same in both conditions: presented in four parts in the treatment and all at once in the control. The tutorial is somewhat longer than the MLGT at seven pages (525 words) with an estimated read time of 2 minutes, 38 seconds.

As the control condition, we assigned volunteers to a slightly modified version of the Gravity Spy Level 4 (labeled M.A.), in which volunteers can categorize a spectrogram using any of the glitch categories. The modified Level 4 was used as the control condition as it matches the approach taken in other online citizen-science projects in the Zooniverse that make all options available to all volunteers from the beginning of their participation. However, the design of the Zooniverse system is such volunteers could choose to classify at a lower level even if they were initially assigned to classify at Level 4 (see Figure 3).

6. Data Collection and Analysis

We obtained two datasets from the Zooniverse database dumps. The first dataset included records of classifications executed by volunteers in Gravity Spy. Each record contained a unique user identification, the experimental group to which the user was assigned, whether the classification was of gold data, a subject identification (a unique identifier for the spectrograms), a timestamp indicating when the classification was executed, and the volunteer's response. The second dataset included the gold-data

classification expert responses, which included a subject identification.

The data were combined using the subject identification value in both datasets. The data were aggregated and analyzed at the session-level. We define a session as a group of consecutive activities separated by no more than 30 minutes. The intuition behind the definition of a session is that volunteers often log in to the system, contribute for some time, and then take a break, e.g., until the next day. As volunteers do not always log out of the system when they are done classifying. A gap of more than thirty minutes indicates the start of a new classification session.

We computed three additional variables, which we consider as dependent variables in our assessment of the MLGT: accuracy, contributions, and retention dependent variables. **Accuracy** was computed by comparing a volunteer's response on gold-data to that of the expert. Accuracy is measured by a volunteers' ability to classify gold standard data correctly. Accuracy is a continuous variable representing the fraction of gold-standard classifications a volunteer answered correctly. **Contributions** is the total number of classifications a volunteer executed during a session. **Retention** was measured as the number of sessions in which a volunteer has contributed.

To test our hypothesis, we conducted significance tests to compare the control and treatment groups on the independent variables described above. For H1 and H3, we first tested the data for normality using the Shapiro-Wilk test. For variables that were normally distributed, we used the independent samples t-test, which is a standard test for difference in population means. For data that were not normally distributed, we used the corresponding non-parametric Mann-Whitney-Wilcox test, which is used to determine whether the data come from the same distribution. For H2, we used a χ^2 test of proportions. All statistical analyses were conducted using R Studio.

6.1. Ethics Review

The experiment protocol was reviewed by our university's human subjects institutional review board (IRB). The experimental procedure posed minimal or no risk to the participants, as the control process was the process used in nearly all other citizen-science projects on the Zooniverse, and the treatment was the same as used by default in Gravity Spy. We did not collect any data about the subjects; only the count and

timing of the classifications they did and the agreement of those classifications with gold-standard data. Indeed, the site does not collect demographic information of any kind, and volunteers are identified only by a self-selected volunteer ID. A section of the initial volunteer agreement provided when volunteers sign up for a Zooniverse account is a disclosure that site administrators run experiments to improve the system and volunteer experience. As collecting informed consent would require volunteers to provide identifying information that was not otherwise collected, we were permitted to run the experiment without requesting specific, informed consent for this experiment.

7. Results

The chart in Figure 5 shows the flow of new volunteers through the experiment. After the experiment had been run, we discovered an omission in the data collection. It appears that the system assigned volunteers to control or treatment when they first visited the site but did not record the assignment until they actually came to the classification page. As a result, volunteers who dropped out while viewing the tutorial were not recorded in the system. As the assignment to control or treatment was random, we believe that roughly equal numbers of volunteers were assigned to each. However, we ended up with unequal numbers of volunteers in the control and treatment groups, apparently because fewer volunteers finished the longer tutorial in the control group. The final population of new volunteers and the population we analyzed was 386: 246 volunteers in the treatment and 140 volunteers in the control.

Two hundred twenty-two of the treatment and 99 of the control group volunteers contributed classifications. Figure 5 shows the number of volunteers that contributed at each workflow level. In the treatment condition, volunteers have to perform well at each level before they are promoted to the next. As a result, contributing at one level is a prerequisite for being able to contribute at a higher level. That is, for volunteers in the treatment, contributing to level 1 is a prerequisite for being allowed to contribute at level 2, and contributing to level 2 is a prerequisite for contributing to level 3. In contrast, volunteers in the control condition began in a modified version of Level 4 (M.A.). However, the system allows volunteers to contribute at lower levels, meaning

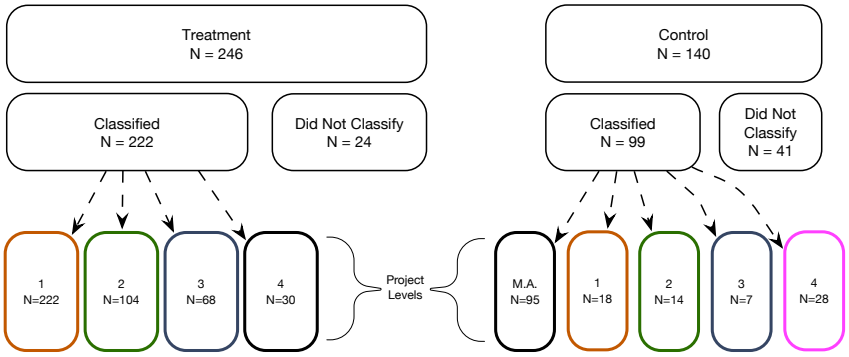


Figure 5: Flow chart showing how many volunteers visited Gravity Spy, created an account and were assigned to either the treatment or control.

that those in the control group could contribute to levels 1-3 if they choose to do so. Because all levels are available for the members of the control group, contributing to level 1 is not a prerequisite for contributing to level 2 and so on. As a result, the counts of contributions at each level are not cumulative for the control group.

7.1. Hypothesis 1: Learning

We first report on the effect of the two training regimes on volunteer classification accuracy in all work levels. We assessed each volunteer's accuracy by examining whether their answers agreed with the science team answers for the gold-standard data. However, of the 321 volunteers who did classifications, 160 did not see any gold-standard data, decreasing the sample size in both the control (N = 46) and the treatment (N = 115). As hypothesized, the average accuracy was significantly higher in the treatment group. The average level of agreement with gold-standard data was 60% (SD = 35.9) for volunteers in the control and 95% (SD = 9.1) for volunteers in the treatment. A Mann-Whitney-Wilcoxon test indicates the difference is significant, W = 2395.5, $p < 0.001$. As a result, H1 is supported.

7.1.1. Gold Data for Level 4 and Modified Apprentice Workflow.

The analysis above includes data for all workflow levels. However, note that in the initial training levels in the MLGT, volunteers select from only a subset of the classes, which could explain the higher accuracy. To address this bias, we compared accuracy

	Treatment	Control	Mann-Whitney
Accuracy on Gold Data			
All Levels	90% (SD = 1%) N = 115	54% (SD = 23%) N = 46	103.5***
Levels 4 & M.A.	56% (SD = 9%) N = 15	46% (SD = 25%) N = 45	t(56.82) = -2.09*

Table 2: Volunteer accuracy on gold data in control and treatment groups, overall and for just level 4. *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$.

for classifications done in the two groups at Level 4, in which there is the same number of options. The results are shown in Table 2. In the control group, which starts at Level 4, 45 volunteers saw gold-standard data. Of the 30 volunteers in the treatment who reached Level 4, 15 saw gold data. We found 46% (SD = 25%) agreement in the control and 56% (SD = 9%) in the treatment group. While the distribution of accuracy scores for all project classifications did not follow a normal distribution, the distribution of accuracy scores for Level 4 and MA classifications did, so we analyzed these accuracy scores using the parametric independent samples t-test. We found the difference in accuracy to be statistically significant at $t(56.82) = -2.09$, $p = 0.04$.

7.2. Hypotheses 2: Initial contribution

Our second hypothesis was that volunteers who went through the MLGT are more likely to contribute classifications. In the experiment, 41 (30%) of volunteers in the control did not classify versus 24 (10%) volunteers in the treatment. We conducted a test of proportions to determine whether the number of volunteers classifying in each group was significantly different. The results of the chi-squared ($\tilde{\chi}^2$) revealed volunteers in the treatment were more likely to make an initial classification $\tilde{\chi}^2(1) = 37.84$, $p < 0.001$. Accordingly, H2 is supported. In addition, we believe that the dropout rate during the initial tutorial was much higher for the control group compared to the training group, as reflected in the different final sample sizes. Unfortunately, we do not have the data on the size of the dropout to test the hypothesis at this point in the volunteers' interaction.

7.3. Hypotheses 3: Contributions

Our third hypothesis was that volunteers who went through the MLGT would contribute more than those who did not. We found that volunteers in the treatment group contributed many more classifications than volunteers in the control: the average number of classifications for volunteers in the control group was 121.1 (SD = 722.7) compared with 228.2 (SD = 677.8) classifications in the treatment group (Table 3). The results of the Wilcoxon rank-sum test revealed a significant effect of the MLGT on the total number of classifications volunteers contributed (W = 7609.5.5, $p < 0.001$). Accordingly, H3 is supported.

	Treatment	**Control**	**Mann-Whitney**
		Classifications	
All Levels	228.2 (SD = 677.8) N = 222	121.1 (SD = 722.7) N = 99	6770.5***
		Sessions	
All Levels	2.5 (SD = 5.8) N = 222	2 (SD = 5.9) N = 99	8783**

Table 3: Volunteer total classifications and number of sessions in control and treatment groups. *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$

7.4. Hypothesis 4: Time as a Contributor

Our final hypothesis concerned the duration of engagement with the Gravity Spy project. The data on the number of sessions suggests that the MLGT increased interest in the project. Volunteers in the control contributed on average 2 (SD = 5.9) sessions while volunteers in the treatment contributed in an average of 2.5 (SD = 5.8) sessions (Table 3). A Mann-Whitney-Wilcoxon test indicates that the distribution of sessions in the training and control groups are significantly different, W = 8783, $p = 0.005$. Accordingly, H4 is supported.

8. Discussion

The research described in this article contributes to existing knowledge about how training impacts learning and motivation in online production communities. Through the use of machine learning algorithms, we addressed two persistent issues for scaffolded instruction - selecting learning resources that are congruent with an individual's competencies and having individuals train and execute tasks concurrently.

Building on prior research on learning and motivation, we proposed that a scaffolded introduction to the work of the Gravity Spy project would be more beneficial and motivating for volunteers (and the project) than simply having volunteers begin without scaffolded training. Further, we implemented an approach to scaffolding in which the materials provided to newcomers were selected by an ML classifier from the actual tasks of the project, rather than being curated by an instructional designer. Overall, the results show that this approach had the hypothesized impacts of improving volunteer accuracy (H1), increasing conversion to a contributor (H2), increasing the number of classifications (H3), and improving retention (H4).

Below, we discuss the role of scaffolding in support of learning and motivation of newcomers and the implication of our findings for the design of informal online learning environments in particular.

8.1. The Benefits of Scaffolding Work for Learning and Motivation

The literature on learning systems has increased in calls for scaffolding access to materials suggesting learners in online settings might find materials confusing and thus increase attrition (Rose & Ferschke, 2016). In computer-supported collaborative learning environments, scaffolding has been shown to enhance participant learning (Rienties et al., 2012) and support community-level benefits such as increasing the volume of contribution and retention among learners (Tuckman, 2007). The study we presented above showed similar findings; however, our main contribution is that scaffolding has similar effects when materials are served to learners using ML-supported training.

The scaffolded MLGT works in several ways. First, the system gradually expands the number of categories presented to the volunteers. As they were promoted, classification options were introduced a few at a time, expanding the number of options

and exemplary glitches shown in the classification interface and the classes of gold-standard data and glitches presented. Additionally, each level includes its own tutorial to gradually introduce features of the system that aid in the classification task and information about the new glitch classes. Finally, as volunteers graduate to new workflows, the category boundaries are expanded.

From a theoretical learning perspective, one can conceptualize this process as 1) legitimate peripheral participation where newcomers gradually expand their access to central activities (Lave & Wenger, 1991); and 2) an ML-supported approach that introduces work fitting each participant's zone of proximal development (Engeström, 2014).

While these two perspectives do not exclude one another, we can argue that the first highlights the phased introduction of the tutorials, prototypical and exemplary glitches, and feedback on gold-standard data. The second emphasizes the ML-supported introduction of work fitting the participants' level of skill. The research findings do not indicate if leveling or the gradual introduction of more difficult glitches matter the most. Future research may help untangle the benefits of these different design principles.

When it comes to motivation, the work design literature suggests that motivating work exists at the intersection of familiarity and challenge (Herzberg, 1968). We suspect that the ML-supported selection of glitches helps to approximate the right level of familiarity and challenge. Again, we are not certain whether it is the phased introduction of work, organized into levels, or the ML-supported selection of image difficulty that plays the most important motivational role. Future research can parse out the effects of these two design choices.

Interestingly, we found that a number of volunteers in the non-scaffolded condition who started in the modified Level 4 workflow, classified in lower level workflows after the experiment completed (Figure 4). For instance, 18 volunteers classified data in workflow one after having been assigned to the control. Volunteers who have worked at Level 4 have already been presented with a comprehensive tutorial, exemplars, and prototypical glitches; we suspect these demoting volunteers were overwhelmed in the modified apprentice workflow and preferred the scaffolded learning. Likewise, there should be little appeal to those motivated by levels to move backward. As a result, the most convincing explanation is that these volunteers seek work matching their zone of

development, neither too hard nor too easy.

To be successful in information integration tasks similar to those in Gravity Spy, Ashby & Maddox (2005) suggests procedural learning opportunities are necessary. Our MLGT includes training, feedback, and scaffolding to introduce categories in which glitches may appear quite different than the prototypes presented to learners in the tutorial or field guide. Categories composed of a large number of exemplars presented contiguously in the classification task facilitates learning. Additionally, learning information integration tasks improve with materials vetted for difficulty. The category boundaries increasingly expand to include more challenging information integration tasks. Exposure to numerous glitch examples that increasing departures from a prototype also appear to improve this scaffolded learning.

Beyond the positive findings associated with learning and motivation among volunteers, the Gravity Spy design offers benefits to the science team behind the project. Scaffolding allowed volunteers to work on real data during their training and not solely on pre-classified glitches. As a result, participants start contributing to science work from the very beginning. They do not have to complete the training before they become productive members of the project.

These findings are also relevant for other human categorization tasks such as teaching radiologist to categorize different diseases based on the appearance of various anomalies in medical images.

8.2. Training Volunteers in online Production Communities

Transforming non-experts into high performing contributors remains a challenge for many online systems. While our results are promising, they also point to future research opportunities.

First, volunteers come to online communities with varied competencies. Some newcomers might arrive with more background knowledge on the task or be more proficient learners. For example, among citizen scientists, many people contribute because of a prior interest in science (Jennett & Cox, 2017; Rotman et al., 2012). Therefore, participants would likely benefit from a personalized tutoring system that starts at their current level rather than from scratch (Karataev & Zadorozhny, 2017).

To properly target training requires an estimate of a volunteer's current level of knowledge. Few citizen-science projects evaluate volunteers' knowledge level at all. Those that do generally rely on proxies, such as the number of classifications contributed. These are quite crude measurements of volunteers' skill level. In the present project, we rely on responses to gold-standard data to assess a volunteer's knowledge, but as noted above, this approach is limited by the amount of gold data volunteers see.

Bayesian methods offer a promising approach to modeling user knowledge as they can incorporate prior knowledge about a volunteer and update it from experience. Such models are widely used to improve the performance of ML systems and human learning (Tenenbaum, 1999; Khajah et al., 2016; Vie et al., 2018). For instance, Corbett & Anderson (1994) Bayesian knowledge tracing (BKT) model has been applied to model learning in the tutoring system as students practice different skills. In our setting, such models could use responses to both gold and non-gold data. However, the citizen science context has to account for the possibility that the ML classification might be incorrect, rather than the volunteer's classification.

Knowing the volunteers' level also opens up new possibilities for interpreting their contributions, specifically for making decisions about the class of a glitch. It should be possible to achieve confidence in the collective assessment with judgments from fewer experienced participants than novices. Similarly, images that are confidently rated by the ML might be retired with fewer classifications or with classifications from less experienced volunteers. We also suspect that asking users to classify a set of predefined glitches to assess their performance (see Vie et al. (2018)) might also be useful for personalized learning.

Second, volunteers learn from resources beyond engaging in the task and receiving feedback from their tagging of gold-standard data. It is well-known in learning systems that formative feedback is important for learning. However, creating meaningful opportunities for feedback remains a challenge (Goldin et al., 2017). As indicated by the learning literature, a number of resources can help participants master a task. In the context of Gravity Spy, volunteers may engage with FAQs, tutorials, and comment forums to learn the practices and norms for contribution. However, these resources can be voluminous and disorganized, making it difficult for volunteers to know which are

relevant to them at the given time. Also, they are relatively static and less personalized. As a result, volunteers could benefit from a scaffolded access to project resources in addition to the introduction to the tasks. To implement such an approach requires a better understanding of the types of resources, participants find helpful at various stages of their participation. For instance, it seems intuitive that tutorials would be most effective with newcomers, while FAQ and comment forums would predominantly benefit more advanced participants, but these intuitions should be tested with data. Additionally, given the rich conversations that occur via discussion fora, one might explore the extent to which these materials can act as feedback opportunities. Several studies have experimented with this form of feedback as an intelligent tutoring system, e.g., Rose & Ferschke (2016); Easterday et al. (2017). For instance, Easterday et al. (2017) suggests several features for a crowd-based design critique system where designers learn through formative feedback from peers that might be applicable in this context.

Third, learners may face a paradox of choice caused by overexposure to numerous categories. Introducing categories gradually across four levels helps reduce this paradox by allowing learners to focus on the morphological characteristics and perceptual distinctions of a smaller set of categories at a time.

Finally, this paper has focused on the image classification task, but as we noted, volunteers may learn and be motivated by the broader scientific questions behind the project, in this case, gravitational waves. Additional learning resources exist to support this form of learning, but we know little about how best to scaffold this material.

8.3. Limitations

Field experiments in online production systems pose challenges that may limit researchers' inferences about volunteers and the community of the study. The research presented here is no different. While the true experimental design does control many threats to internal validity, there are two caveats.

The most apparent limitation here is the nature of the assignment of volunteers to the conditions of the study. First, the system did not correctly record the assignment of volunteers to condition, meaning that our analysis starts partway into the study. We believe, but do not have data to show that more volunteers dropped out in control con-

dition than in the treatment condition. However, this threat, known as experimental mortality, would be expected to leave the control condition with more motivated volunteers than the treatment, and so does not explain our findings that the treatment group seemed motivated. In other words, while it is unfortunate that the experimental data were not completely captured, this lacuna does not compromise the main contribution of the study.

Further, the system did not prevent volunteers in the control group from contributing at other levels, and in fact, 18 control group members at some point did visit other levels. This threat is known as design contamination, meaning that some of the control group received the treatment. This contamination would make the control and treatment groups more similar than otherwise, so again, this threat does not explain our findings. Indeed, as both threats tend to reduce the difference between the control and treatment groups, the actual effect of the treatment may be greater than we observed.

The second limitation concern the measurement of volunteer accuracy. We could not control the frequency with which volunteers see gold data or which class of glitch is shown. As a result, some volunteers did not see any gold data and so could not be included in our analysis of accuracy. However, this omission should affect the control and treatment groups equally.

Finally, the trade-off for design with strong internal validity is weaker external validity. We have shown that our training approach works in the Gravity Spy setting, but can not say for sure how the approach will work elsewhere. However, the underlying theoretical rationale for the approach suggests that it could be useful in citizen science projects more generally and perhaps for other kinds of online communities. Furthermore, the training has a number of parameters, e.g., how many classes to introduce and after what level of performance. The experiment has tested the only point in this design space, and so does not provide insight into the optimal settings.

9. Conclusions

In summary, we have presented an approach to newcomer training that offers learners ML-selected tasks in an attempt to fit their zone of proximal development, work that

is not too easy or too difficult. Our experiment shows that this approach is successful in increasing the accuracy of the volunteers while also increasing motivation and contribution. Even these initial classifications are useful to the project, as ML assessment is not perfect and so needs to be checked.

This approach to training addresses the dilemma faced by online communities in particular, as making good use of newcomers' contributions is important in setting where many volunteers only contribute a few times. Equally important, this approach to training scales to large numbers of participants who can engage in on-the-job training without requiring more experienced workers to evaluate the work quality.

Although our focus has been on learning in an online production community, it should be possible to apply the approach to other settings in which many newcomers need to learn to perform a variety of tasks. Other citizen science projects are also using scaffolded training with machine learning (see: Supernova Hunters Wright et al. (2017)). The main limitation is the need to train an ML-model to do the task. However, ML technology is rapidly improving and being applied to more kinds of work, suggesting that there will be many future applications. Often, the hope is to use the ML to complete automate the task, but in many cases, this hope may be too optimistic. The approach presented here offers a path to creating a collaboration between human and machine learning that takes advantage of the strengths of each while enabling both to learn, improve, and contribute.

Acknowledgements

We thank the citizen science volunteers and Zooniverse team for access to data.

References

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annu. Rev. Psychol.*, *56*, 149–178.

Bahaadini, S., Noroozi, V., Rohani, N., Coughlin, S., Zevin, M., Smith, J. R., Kalogera, V., & Katsaggelos, A. (2018). Machine Learning for Gravity Spy: Glitch Classification and Dataset. *Information Sciences*, (pp. 1–34).

Bahaadini, S., Rohani, N., Coughlin, S., Zevin, M., Kalogera, V., & Katsaggelos, A. K. (2017). Deep multi-view models for glitch classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2931–2935). IEEE.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C. (2009). *Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report.*. Technical Report.

Bowser, A., Hansen, D., He, Y., Boston, C., Reid, M., Gunnell, L., & Preece, J. (2014). Using gamification to inspire new citizen science volunteers. In *Gamification '13: Proceedings of the First International Conference on Gameful Design, Research, and Applications* (pp. 1–8). New York, New York, USA: ACM.

Brossard, D., Lewenstein, B., & Bonney, R. (2005). Scientific knowledge and attitude change: The impact of a citizen science project. *International Journal of Science Education*, *27*, 1099–1121.

Brown, J. S., & Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science*, *2*, 40–57.

Bryant, S. L., Forte, A., & Bruckman, A. S. (2005). Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (pp. 1–10). ACM.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., & players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*, 756–760.

Corbalan, G., Kester, L., & van Merriënboer, J. J. (2009). Dynamic task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and Instruction*, *19*, 455–465.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278. doi:10.1007/BF01099821.

Crowston, K., & Fagnot, I. (2018). Stages of motivation for contributing user-generated content: A theory and empirical test. *International Journal of Human-Computer Studies*, *109*, 89–101. doi:10.1016/j.ijhcs.2017.08.005.

Crowston, K., Østerlund, C., Lee, T. K., Jackson, C. B., Harandi, M., Allen, S., Bahaadini, S., Coughlin, S., Katsaggelos, A., Larson, S., Rohani, N., Smith, J., Trouille, L., & Zevin, M. (2019). Knowledge Tracing to Model Learning in Online Citizen Science Projects. *IEEE Trans. Learning Technol.*, .

Curtis, V. (2015). Motivation to Participate in an Online Citizen Science Game: A Study of Foldit. *Science Communication*, *37*, 723–746.

Dickson, S. V., Chard, D. J., & Simmons, D. C. (1993). An integrated reading/writing curriculum: A focus on scaffolding. In *LD Forum* (pp. 12–16).

Downes, S. (2006). Learning networks and connective knowledge. *Collective intelligence and elearning*, *20*, 1–26.

Ducheneaut, N. (2005). Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work*, (p. 323–368).

Easterday, M. W., Lewis, D. R., & Gerber, E. M. (2017). Designing Crowdcritique Systems for Formative Feedback. *International Journal of Artificial Intelligence in Education*, (pp. 1–41).

Engeström, Y. (2014). *Learning by expanding*. Cambridge University Press.

Estrada, T., Pusecker, K. L., Torres, M. R., Cohoon, J., & Taufer, M. (2013). Benchmarking gender differences in volunteer computing projects. In *eScience (eScience), 2013 IEEE 9th International Conference on* (pp. 342–349). IEEE.

Ford, H., & Geiger, R. S. (2012). "writing up rather than writing down": Becoming wikipedia literate. In *Proceedings of the Eighth Annual International Symposium*

on Wikis and Open Collaboration WikiSym '12 (pp. 16:1–16:4). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2462932.2462954. doi:10.1145/2462932.2462954.

Goldin, I., Narciss, S., Foltz, P., & Bauer, M. (2017). New Directions in Formative Feedback in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education*, (pp. 1–8).

Haythornthwaite, C. (2014). New Media, New Literacies, and New Forms of Learning. *International Journal of Learning and Media*, *4*, 1–8.

Haythornthwaite, C., & Andrews, R. (2011). *E-learning theory and practice*. Sage Publications.

Herzberg, F. (1968). One More Time: How Do You Motivate Employees? *Harvard Business Review*, *46*, 53–62.

Hogan, K., & Pressley, M. (1997). Advances in learning & teaching. Scaffolding student learning: Instructional approaches and issues. US: Brookline Books Cambridge, MA.

Iacovides, I., Jennett, C., Cornish-Trestrail, C., & Cox, A. L. (2013). Do games attract or sustain engagement in citizen science? A study of volunteer motivations. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 1101–1106). ACM.

Jennett, C., & Cox, A. L. (2017). *Digital Citizen Science and the Motivations of Volunteers* volume 343. Chichester, UK: John Wiley & Sons, Ltd.

Johri, A., & Yang, S. (2017). Scaffolded help for learning: How experts collaboratively support newcomer participation in online communities. In *Proceedings of the 8th International Conference on Communities and Technologies* C&T '17 (pp. 149–158). New York, NY, USA: ACM.

Jones, C., & de Laat, M. (2016). Network Learning. In C. Haythornthwaite, R. Andrews, J. Fransman, & E. M. Meyers (Eds.), *The SAGE Handbook of E-learning Research* (pp. 44–61).

Jones, M., Love, B. C., & Maddox, W. T. (2005). Stimulus generalization in category learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 27.

Karataev, E., & Zadorozhny, V. (2017). Adaptive Social Learning Based on Crowdsourcing. *IEEE Trans. Learning Technol.*, *10*, 128–139.

Khajah, M. M., Roads, B. D., Lindsey, R. V., Liu, Y.-E., & Mozer, M. C. (2016). Designing engaging games using bayesian optimization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* CHI '16 (pp. 5571–5582). New York, NY, USA: ACM. doi:10.1145/2858036.2858253.

Kicken, W., Brand Gruwel, S., & van Merriënboer, J. J. G. (2008). Scaffolding advice on task selection: a safe path toward self-directed learning in on-demand education. *Journal of Vocational Education & Training*, *60*, 223–239.

Kim, S., & Murphy, G. L. (2011). Ideals and category typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1092.

Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., Konstan, J., Ren, Y., & Riedl, J. (2012). *Building Successful Online Communities*. Evidence-Based Social Design. MIT Press.

Kulatunga-Moruzi, C., Brooks, L. R., & Norman, G. R. (2011). Teaching posttraining: Influencing diagnostic strategy with instructions at test. *Journal of Experimental Psychology: Applied*, *17*, 195.

Land-Zandstra, A. M., van Beusekom, M., Koppeschaar, C., & van den Broek, J. (2016a). Motivation and learning impact of Dutch flu-trackers. *Journal of Science Communication*, *15*, A04. URL: https://jcom.sissa.it/archive/15/01/JCOM_1501_2016_A04.

Land-Zandstra, A. M., Devilee, J. L. A., Snik, F., Buurmeijer, F., & Broek, J. M. v. d. (2016b). Citizen science on a smartphone: Participants' motivations and learning. *Public Understanding of Science*, *25*, 45–60. doi:10.1177/0963662515602406.

Lave, J. (1991). Situating learning in communities of practice. *Perspectives on socially shared cognition*, *2*, 63–82.

Lave, J., & Wenger, E. (1991). *Situated Learning. Legitimate Peripheral Participation*. NY: Cambridge University Press.

Lee, T. K., Crowston, K., Harandi, M., Østerlund, C., & Miller, G. (2018). Appealing to different motivations in a message to recruit citizen scientists: results of a field experiment. *Journal of Science Communication*, *17*, 1–22.

Leutner, D. (1993). Guided discovery learning with computer-based simulation games: Effects of adaptive and non-adaptive instructional support. *Learning and Instruction*, *3*, 113–132.

Luckin, R. (2008). The learner centric ecology of resources: A framework for using technology to scaffold learning. *Computers & Education*, *50*, 449–462.

Malinen, S. (2015). Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in Human Behavior*, *46*, 228–238.

Masters, K., Oh, E. Y., Cox, J., Simmons, B., Lintott, C., Graham, G., Greenhill, A., & Holmes, K. (2016). *Science Learning via Participation in Online Citizen Science*. Technical Report.

Molenaar, I., Roda, C., van Boxtel, C., & Sleegers, P. (2012). Dynamic scaffolding of socially regulated learning in a computer-based learning environment. *Computers & Education*, *59*, 515–523.

Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development*, *53*, 35–45.

Morschheuser, B., Hamari, J., Koivisto, J., & Maedche, A. (2017). Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International*

Journal of Human-Computer Studies, *106*, 26–43. doi:10.1016/j.ijhcs.2017.04.005.

Mugar, G., Østerlund, C., Hassman, K. D., Crowston, K., & Jackson, C. B. (2014). Planet hunters and seafloor explorers: legitimate peripheral participation through practice proxies in online citizen science. In *Proceedings of the ACM 2014 conference on Computer Supported Cooperative Work* (pp. 109–119). New York, New York, USA: ACM Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification—categorization relationship. *Journal of experimental psychology: General*, *115*, 39.

Østerlund, C., & Carlile, P. (2005). Relations in Practice: Sorting Through Practice Theories on Knowledge Sharing in Complex Organizations. *The Information Society*, *21*, 91–107.

Pandya, R. E. (2012). A framework for engaging diverse communities in citizen science in the US. *Frontiers in Ecology and the Environment*, *10*, 314–317.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, *1*, 13–32.

Raddick, J., Bracey, G., Gay, P. L., Lintott, C. J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. (2013). Galaxy Zoo: Motivations of Citizen Scientists. *arXiv*, . arXiv:1303.6886v1.

Raddick, J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. (2010). Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review*, *9*, 1–18.

Reed, J., Raddick, J., Lardner, A., & Carney, K. (2012). An Exploratory Factor Analysis of Motivations for Participating in Zooniverse, a Collection of Virtual Citizen Science Projects. In *2013 46th Hawaii International Conference on System Sciences (HICSS)* (pp. 610–619). IEEE.

Rienties, B., Giesbers, B., Tempelaar, D., Lygo-Baker, S., Segers, M., & Gijselaers, W. (2012). The role of scaffolding and motivation in cscl. *Computers & Education*, *59*, 893 – 906.

Rose, C. P., & Ferschke, O. (2016). Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses. *International Journal of Artificial Intelligence in Education*, (pp. 1–19).

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., & Jacobs, D. (2012). Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 217–226). ACM.

Sauermann, H., & Franzoni, C. (2015). Crowd science user contribution patterns and their implications. *Proc. Natl. Acad. Sci. U.S.A.*, *112*, 679–684.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web* WWW '14 Companion (pp. 1049–1054). New York, NY, USA: ACM. doi:10.1145/2567948.2579215.

Sinha, P., & Russell, R. (2011). A perceptually based comparison of image similarity metrics. *Perception*, *40*, 1269–1281.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II* (pp. 59–65). Cambridge, MA, USA: MIT Press.

Tuckman, B. W. (2007). The effect of motivational scaffolding on procrastinators' distance learning outcomes. *Computers & Education*, *49*, 414–422.

Van Maanen, J., & Schein, E. H. (1979). Toward of Theory of Organizational Socialization. *Research in Organizational Behavior*, *1*, 209–264.

Vie, J.-J., Popineau, F., Bruillard, É., & Bourda, Y. (2018). Automated Test Assembly for Handling Learner Cold-Start in Large-Scale Assessments. *International Journal of Artificial Intelligence in Education*, (pp. 1–16).

Vygotsky, L. S. (1980). *Mind in Society*. The Development of Higher Psychological Processes. Harvard University Press.

Winn, J. A. (2007). Promises and Challenges of Scaffolded Instruction. *Learning Disability Quarterly*, *17*, 89–104.

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., Chambers, K. C., Flewelling, H., Granger, W., Magnier, E. A., McMaster, A., Miller, G. R. M., O'Donnell, J. E., Spiers, H., Tonry, J. L., Veldthuis, M., Wainscoat, R. J., Waters, C., Willman, M., Wolfenbarger, Z., & Young, D. R. (2017). A transient search using combined human and machine classifications. *arXiv*, (pp. 1315–1323). arXiv:1707.05223v1.

Zevin, M., Coughlin, S., Bahaadini, S., Besler, E., Rohani, N., Allen, S., Cabero, M., Crowston, K., Katsaggelos, A., Larson, S., Lee, T. K., Lintott, C., Littenberg, T., Lundgren, A., Oesterlund, C., Smith, J., Trouille, L., & Kalogera, V. (2017). Gravity spy: Integrating advanced LIGO detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, *34*. doi:10.1088/1361-6382/aa5cea.