Astronomy in Focus - XXX
Proceedings IAU Symposium No. XXX, 2018
M. T. Lago, ed.

doi:10.1017/S174392131900526X

Transforming research (and public engagement) through citizen science

Samantha Blickhan[1], **Laura Trouille**[1] **and Chris J. Lintott**[2]

[1]Department of Citizen Science, The Adler Planetarium
1300 South Lake Shore Dr. Chicago, IL 60605, USA
emails: samantha@zooniverse.org, trouille@zooniverse.org

[2]Department of Physics, The University of Oxford
Keble Road, Oxford, OX1 3RH, UK
email: chris.lintott@physics.ox.ac.uk

Abstract. Processing our increasingly large datasets poses a bottleneck for producing real scientific outcomes and citizen science - engaging the public in research - provides a solution, particularly when coupled with automated routines. In this talk we will provide a broad overview of citizen science approaches and best practices. We will also highlight in particular recent advances through Zooniverse, the world's largest platform for online citizen science, engaging more than 1.7 million volunteers in tasks including discovering exoplanets, identifying features on Mars' surface, transcribing artist's notebooks, and tracking resistance to antibiotics.

Keywords. citizen science, public engagement, machine learning

1. Introduction

As data rates and volumes grow, many research communities are turning to online citizen science as a method of coping with large datasets. Citizen science typically refers to collaborative research involving professional scientists and volunteer participants (Marshall *et al.* 2015). By involving the public in the research process through citizen science activities such as data collection, classification or evaluation, and analysis, we can process massive datasets in a reasonable amount of time while simultaneously engaging the public and promoting a greater understanding of the scientific process.

Zooniverse (www.zooniverse.org) is the largest platform for online citizen science in the world. It began with a single project called Galaxy Zoo, launched in July 2007. The goal of the project was to process just under one million images of galaxies from the Sloan Digital Sky Survey into single morphological categories. Several hundred thousand volunteers participated, and not only did they complete the task in a matter of months, but each image was classified an average of 38 times, instead of the single classification that would have been produced by an individual researcher. The results of the initial Galaxy Zoo project suggested that the general public can reliably classify large sets of galaxies with a similar accuracy to professional astronomers (Lintott *et al.* 2008). The success of Galaxy Zoo led to the foundation of the Zooniverse platform, which currently hosts more than 90 active projects and a volunteer community of 1.7 million registered participants in 234 countries. Among online citizen science platforms Zooniverse is unique, due to its 1) shared open-source software, communication and participation among inter- and multi-disciplinary users; 2) reliable, flexible, and scalable Application Programming Interface (API), which can be used for a variety of development tasks; 3) the Project Builder, a do-it-yourself (DIY) tool which allows anyone to build their own project for free (described further below); and 4) the size and scale of its audience. Zooniverse partners with hundreds of researchers across the disciplines, and has become a compelling option for research framework as citizen science gains popularity around the world. Since the launch of the Galaxy Zoo project in 2007, data from Zooniverse projects have been used in over 150 peer-reviewed publications across many disciplines, including astronomy (Fortson *et al.* 2012, Johnson *et al.* 2015, Lintott *et al.* 2008, Marshall *et al.* 2016, biomedical (dos Reis *et al.* 2015), climate science (Hennon *et al.* 2015, Rosenthal *et al.* 2018), ecology (Matsunaga *et al.* 2016, Swanson *et al.* 2016), and the humanities (Grayson 2016, Williams *et al.* 2014). These publications and the projects from which they stem have set a precedent for online citizen science producing significant, high-quality data.

2. Types of discoveries

With so many eyes on project data, combined with intuitive design and a supportive culture, Zooniverse is in a unique position to support multiple kinds of discovery: the Known Unknowns and the Unknown Unknowns. The former refers to making discoveries on the Zooniverse platform through the classification task; when the intended engagement with the data results in discovery. For example, in April 2017, the Exoplanet Explorers project (Schwamb *et al.* 2012) built on the work of the previous Planet Hunters project, in which citizen scientists examined time series photometry from *Kepler* to identify potential planetary transits. In Exoplanet Explorers, researchers at UC Santa Cruz and Caltech used processed *K2* time series data for candidates identified by machine learning. The project launch was set to coincide with a television program called Stargazing Live, on the Australian Broadcasting Corporation, during which the presenters of the program encouraged viewers to participate online, and gave live updates on the project results. Thousands of citizen scientists volunteered, and on the final night of the three-day program, researchers announced the discovery of a four-planet system by participants in the project. Since then, the near-resonant system has been named K2-138 and the research team have found that has a fifth planet, and possibly a sixth, although additional verification is needed (Christiansen *et al.* 2018). This example illustrates the enormous potential for discovery through crowdsourced data processing, and is one of the main reasons why the Zooniverse platform has been steadily growing for more than 10 years.

From its beginnings, the Zooniverse platform design has not only supported discovery through the classification task, but has also enabled *serendipitous* discovery; a result of having multiple independent examinations of the data, but also taking advantage of our innate human ability to notice things that are out of the ordinary. According to Luczak-Roesch *et al.* (2014), it is not uncommon for citizen scientists to go to great lengths to alert researchers to the existence of an unusual object of interest and/or work on independent research questions based on something they noticed while performing the prescribed main classification task. The Zooniverse 'Talk' discussion forums have been central to these volunteer-led investigations. The forums were intentionally designed to facilitate serendipitous discovery: when a volunteer finishes a classification on any Zooniverse project, they can choose to go directly to the Talk board to discuss the subject they have just seen. An example of the Talk board from the Gravity Spy project can be seen below, in Fig. 1.†

Talk provides a space for a wide range of information sharing and exchange of ideas between volunteers and researchers, enabling collaboration and community building. The interactions in Talk range from the purely social to goal-driven and benefit the project directly, as a means of responding to inquiry, and indirectly, as a way of ensuring long-term engagement (Mugar *et al.* 2014).

† https://www.zooniverse.org/projects/zooniverse/gravity-spy/talk

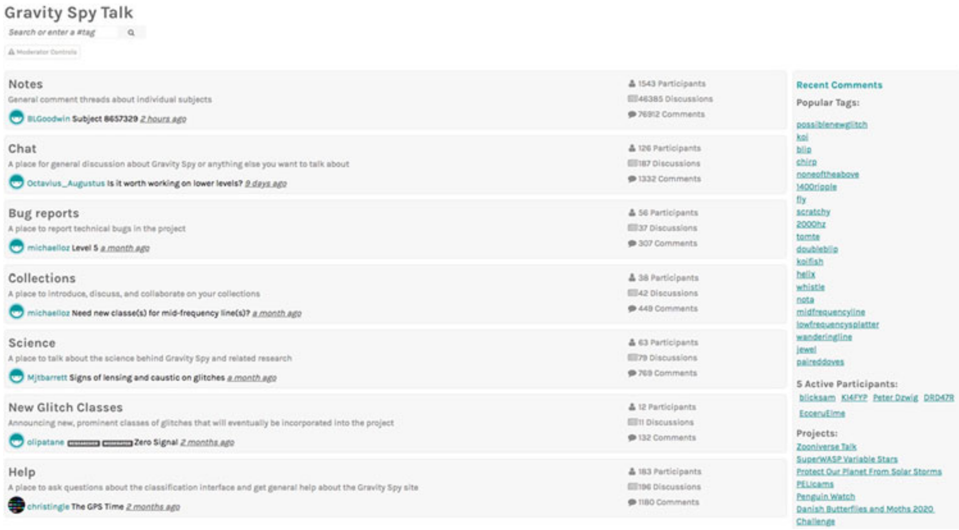

Figure 1. 'Talk' board for the Gravity Spy project.

Over the years, Zooniverse volunteer effort has led to a number of serendipitous discoveries that have been transformative in their respective fields. One thread in particular, posted on the Galaxy Zoo Talk board, led to a discovery which highlights the power of including message boards with each project. While classifying, several volunteers had noticed strange, compact green blobs which looked like peas, and subsequently started a Talk discussion thread titled 'Give peas a chance'. The researchers worked alongside the volunteers (who called themselves the Peas Corps) to refine the collection of objects, ultimately identifying 250 'green peas' in the million-galaxy data set. We now know that these compact green galaxies are low mass, low metallicity galaxies with high star formation rates (Cardamone *et al.* 2009). Yet when the Galaxy Zoo project began, this type of galaxy was not part of the taxonomy of galaxy types, let alone something that the research team was actively searching for with the classification task – a true 'Unknown Unknown'. According to Dr. Carolin Cardamone, one of the lead researchers for Galaxy Zoo, "No one person could have done this on their own. Even if we had managed to look through 10,000 of these images, we would have only come across a few Green Peas and wouldnt have recognized them as a unique class of galaxies."†

The 'Green Peas' are just one example of serendipitous discovery on the Zooniverse platform. Other examples include Hanny's Voorwerp (Lintott *et al.* 2009) and 'Boyajian's Star', KIC 8462852 (Boyajian *et al.* 2016), as well as the re-discovery of a group of 18th century female scientific illustrators on the Science Gossip project‡.

3. Project growth

As rates and volumes of data steadily climb, the number of research teams wanting to build crowdsourcing projects has similarly grown. As noted above, Zooniverse has grown from a single, custom-built project in 2007, to a platform with more than 90 active projects in which volunteers can choose to participate. The accelerated expansion of Zooniverse is a result of the launch in July 2015 of our free Project Builder tool,¶

† https://news.yale.edu/2009/07/27/galaxy-zoo-hunters-help-astronomers-discover-rare-green-pea- galaxies

‡ https://talk.sciencegossip.org/#/boards/BSC0000004/discussions/DSC00004s8

¶ https://www.zooniverse.org/lab

https://doi.org/10.1017/S174392131900526X Published online by Cambridge University Press

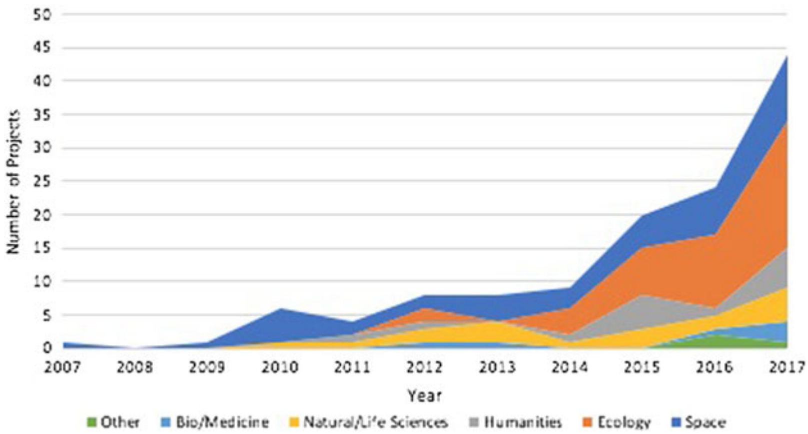

Figure 2. Zooniverse projects launched per year, 2007-2017.

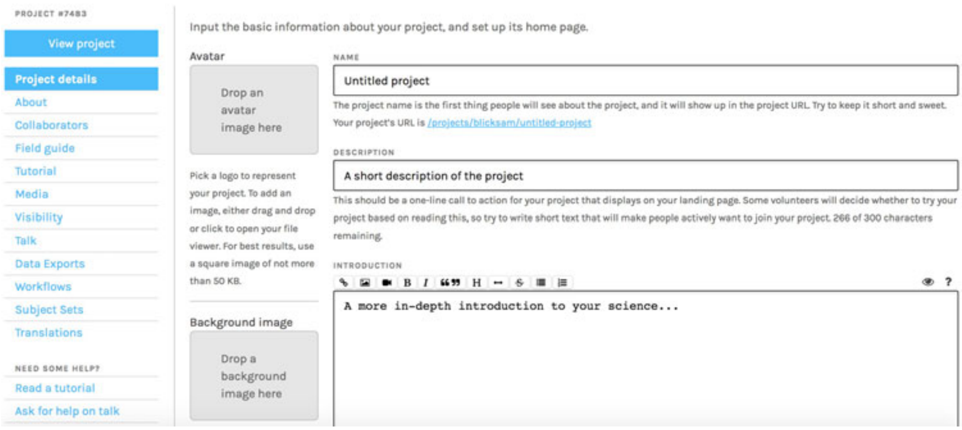

Figure 3. 'Project details' tab of the Zooniverse Project Builder.

which enables anyone to create an online citizen science project using a web browser-based toolkit. Prior to the development of the Project Builder, a typical online citizen science project was entirely custom-built, and required months to years of professional web development time. Fig. 2 shows project growth by year, from 2007 to 2017.

The Project Builder supports common types of interaction including classification, multiple-choice questions, comparison tasks, text entry, marking and drawing tools, or any combination thereof. The Project Builder front-end is a series of forms and text boxes a researcher fills out to create the project's classification interface and website (Fig. 3). All Project Builder projects come with a landing page, classification interface, discussion forum, and 'About' pages for content about the research, the research team, and results from the project.

Since the Project Builder launch, over 1,800 people have attempted building their own project. Some projects go through beta review and are launched publicly, while other projects remain private, for internal use among research teams, community-based projects, etc. One new public project is launched per week. Though we have over 1.7 million registered volunteers, we must acknowledge that this rate of project growth will either require renewed efforts toward recruiting even more volunteers, or finding other ways to affect project efficiency, as described below.

https://doi.org/10.1017/S174392131900526X Published online by Cambridge University Press

4. Future developments

As we plan to provide support for the growing number of Zooniverse projects, we also need to anticipate huge amounts of data from future surveys, such as LSST, which will produce 30 terabytes of data per night (Ivezic *et al.* 2018). In preparation for massive amounts of data, Zooniverse has explored the use of human-machine systems across several projects. For example, Beck *et al.* (2018) details an experiment with the Galaxy Zoo 2 dataset of 200,000 galaxy images, in which Zooniverse volunteers classified on the Galaxy Zoo project for over a year (from 2010-2011) in order to retire the full dataset. In 2017, the research team used the human classifications to train an off-the-shelf algorithm to simulate what would have happened had the project integrated machine learning in 2010. After four days of near real-time training (days 4-8 of the human classification data), the machine retired over 70,000 images on its first run. Through active learning, the machine retired all 200,000 images by day 30, over a factor of 8 decrease in time compared to the more than 400 days it took human volunteers to complete the dataset.

Zooniverse machine learning efforts are closely aligned with an auxiliary service known as Caesar; a secondary Rails module that monitors classifications in real time, supporting aggregation and subject retirement and promotion. Caesar can set rules and actions based on those rules, such as responsive retirement and linking subjects retired from one workflow to the next logical workflow, as well as supporting integration of machine learning algorithms into the Zooniverse system.

A number of projects have used volunteer classifications to generate training sets for automated methods to efficiently classify remaining data. For example, an early Supernova Hunters project on Zooniverse, using data from the Palomar Transient Factory, retired from the system after the volunteer classifications provided a large enough training set for the researchers' machine learning algorithms to automatically process the remaining data with confidence. Since the implementation of Caesar, projects have been able to utilize human effort concurrently with machine learning. The Camera CATalogue project trained models to provide predicted classifications for species identification in camera trap images, and used custom Caesar rules to retire an image if the first two human volunteers agreed with the model's prediction. Ultimately, this has allowed the project to maintain classification accuracy rates while reducing human effort by 43% (Willi *et al.* 2018).

As another example, the current Supernova Hunters project† uses machine learning to flag Pan-STARRS telescope images as containing potential supernovae candidates. Subjects which the machine deems unlikely to contain supernovae are retired automatically; the remaining subjects are uploaded to the project for volunteer classification. Wright *et al.* (2017) found that the human classifications and machine learning results were complementary; by combining the aggregated volunteer classifications with the machine learning results, they are able to create the most pure and complete sample of new supernovae candidates.

The integration of machine learning into Zooniverse may at first seem antithetical to the existence of a platform dedicated to crowdsourced participation in research. However, we have shown that machine and human *collaboration* can produce superior classification results than either one alone. Integration of machines can also greatly increase the efficiency of a citizen science system, but tensions exist when designing for engagement, inclusion and serendipity. Trouille *et al.* (2018) discuss these tensions in detail, and offer potential solutions, such as the Snapshot Wisconsin team experiment about including blank images in ecology camera-trap projects (Bowyer *et al.* 2015).

† zooniverse.org/projects/dwright04/supernova-hunters

5. Concluding remarks

Zooniverse is a transformative tool for research and public engagement in science. In the eleven years since the launch of Galaxy Zoo, the Zooniverse platform has grown exponentially, a result of steadily-growing data rates and volumes. Recent experiments with machine learning have shown that human-computer collaboration will be essential to the continued health of the platform - and, arguably, the field of citizen science - in years to come.

6. Acknowledgements

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google and by a grant from the Alfred P. Sloan Foundation. The development of Caesar was supported in part by support from STFC under grant ST/N003179/1. We gratefully acknowledge our amazing Zooniverse web development team, the research teams leading each Zooniverse project, and the worldwide community of Zooniverse volunteers who make this all possible.

References

Anderson, TM, *et al.* 2016, *Philosophical Transactions of the Royal Society B: Biological Sciences* 371(1703):20150314
Areta, C, *et al.* 2016, in *European Conference on Computer Vision* 483–98
Barr, AJ, *et al.* 2016, arXiv:1610.02214 [physics.soc-ph]
Beck, MR, *et al.* 2018, *MNRAS* 476(4): 5516–34
Bowyer, A, *et al.* 2015, 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*
Boyajian, TS, *et al.* 2016, *MNRAS* 457(4): 3988–4004
Cardamone, C, *et al.* 2009, *MNRAS* 399(3): 1191–1205
Christiansen, JL, *et al.* 2018, *AJ* 155(2): 57
Fortson, L, *et al.* 2012, in *Advances in Machine Learning and Data Mining for Astronomy* (CRC Press): 214-33
Grayson, RS 2016, *British Journal for Military History* 2(2): 160–85
Hennon, CC, *et al.* 2015, *Bul. Am. Met. Soc.* 96(4): 591–607
Ivezic, Z, *et al.* 2018, arXiv:0805.2366v5 [astro-ph]
Johnson, LC, *et al.* 2015, *ApJ* 802(2): 127
Lintott, CJ, *et al.* 2008, *MNRAS* 389(3): 1179–89
Lintott, CJ, *et al.* 2009, *MNRAS* 399(1): 129–40
Luczak-Roesch M, *et al.* 2014, in *ICWSM* aaai.org, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8092/8136
Marshall, PJ, *et al.* 2015, *ARAA* 53: 247–78
Marshall, PJ, *et al.* 2016, *MNRAS* 455: 1171–90
Matsunaga, A, *et al.* 2016, *Future Generation Computer Systems* 56: 526–36
Mugar, G, *et al.* 2014, in: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* (New York, NY, USA: ACM): 109–19
dos Reis, FJC, *et al.* 2015, *EBioMedicine* 2(7): 681–89
Rosenthal, IS, *et al.* 2018, *in press*
Schwamb, ME, *et al.* 2012, *apj* 754(2): 129–146
Swanson, A, *et al.* 2012, *Conservation Biology* 30(3): 520–31
Trouille, L, *et al.* 2018, accepted to *PNAS*
Willi, M, *et al.* 2018, submitted to *Methods in Ecology and Evolution*
Williams, AC, *et al.* 2014, in *IEEE International Conference on Big Data* 100–105
Wright, D, *et al.* 2017, in *Monthly Notices of the Royal Astronomical Society* 472(2): 1315–23
Zevin, M, *et al.* 2017, *Classical and Quantum Gravity* 34(6): 064003