Publications of the Astronomical Society of Australia (PASA)

doi: 10.1017/pas.2015.xxx.

The Role of Citizen Science in Future Surveys in Astrophysics

Chris Lintott[1], Brooke Simmons[1], Campbell Allen[1], Andy J. Connolly[1], Peter T. Darch[1], Rob Fender[1], Lucy Fortson[1], Bryan Gaensler[1], Henry Hseih[1], Matt Jarvis[1], Sarah Kendrew[1], Sandor Kruk[1], Stuart Lynn[1], Karen L. Masters[1], Adam McMaster[1], Grant Miller[1], Robert C. Nichol[1], Edward Paget,[1] Stephen J. Roberts[1], Claudia Scarlata[1], Stephen Serjeant[1], Edwin Simpson[1], Rebecca Smethurst[1], Aprajita Verma[1], Shiang-Yu Wang[1], and Kyle W. Willett [1]

[1]St. Catherine's College, Oxford University

Abstract

...

Keywords: keyword1 – keyword2 – keyword3 – keyword4 – keyword5

1 INTRODUCTION

Citizen science in the form of distributed data analysis has made significant contributions to astronomical science through projects such as Galaxy Zoo which allow many hundreds of thousands of people to make contributions. As astronomers continue to develop surveys which will produce ever growing volumes of (often) publicly accessible data, it seems likely that further opportunities for such collaboration will occur, though the increasing sophistication of machine learning and the need in some cases for rapid classification may well requiring increasingly sophisticated frameworks to be put in place. This paper is the result of discussions at a workshop held at St Catherine's College in Oxford from April 15th-17th 2015, hosted by the Zooniverse and sponsored by the Kavli Foundation, at which participants sought to explore the possibilities for citizen science with the crop of surveys currently under development.

2 THE PROBLEM: SURVEYS AND BIG DATA

3 ZOONIVERSE ARCHITECTURE

4 REAL TIME CLASSIFICATION

What 'real time' means depends on the science case. Radio transients observed within minutes of discovery have been found to be past their peak brightness, supernovae may need follow-up in days whereas surveys such as DES think of follow-up on timescales of months in deciding which objects should be targeted by fibres. However, for most of these use cases then *classification* is still urgent so that decisions can be taken, even if action can wait.

It is useful to think about problems involving streams of data for classification as being divided by their cadence. Projects can usefully be divided into the following categories :

- **Batch projects:** Provide data on, for example, monthly timescales. Examples include projects using the main (annual) LSST data release.
- **Sparse projects:** Alerts at irregular intervals when intervention is required. Examples include supernova identification with pre-filtering.
- **Continuous projects:** These projects would require substantial datasets to be reviewed, with enough data to keep volunteers busy all the time (rather than requiring a small number of actions after an alert).
- **Campaign projects:** Similar to continuous projects, but running for a short while; an example was the Snapshot Supernova project which reviewed Skymapper data while PESSTO follow-up was available.

In all cases, the common problem is the identification of several different categories of interesting object alongside a variety of artefacts; we essentially face a filtering problem. This filtering is a classification problem, but the performance of classification is also effected by

the bias of astronomers affected by time constraints. A striking example was the recent discovery of superluminous supernovae, of which ~ 10 are known. These objects are identified by their long decay times, and have been found in archival data in SDSS, but were not recognised as a category of interest until recently, despite many searches for transients in these data sets.

A similar story emerges from the identification, by citizen scientists, of a set of extreme star forming dwarf galaxies known as the 'Green Peas'; identified by visual inspection in the first instance, they can be identified through judicious cuts in colour/colour space. The citizen science task may be less repeated classification but rather the identification of categories of interest. Machine learning can then be used to 'amplify' categorization provided by inspection in a series of filters or, perhaps, a random forest populated by classifiers.

An example project could proceed by starting with a simple set of questions about expected common object types. As classification proceeds we should expect the performance of algorithms to improve as a large training set is built up. This will allow the filtering of the dataset to remove common object types, increasing the potential for rarer objects to be found. Classification instructions can then be adjusted to concentrate on these new targets, until they too obtain training sets large enough for significant algorithmic filtering. The advantages of such a scheme include the fact that it is responsive; it is capable of adjusting to new categories of interest and to improvements in algorithmic performance.

In essence, this proposal recognises that whether or not real time observing is happen what is important is on-the-fly learning. Training sets developed dynamically could respond to changes in conditions or instrument performance, but more importantly also to the changing characteristics of a filtered dataset. This is especially important as training sets for rarer objects can only be produced once more common ones have been removed. An ambitious system - one which is truly efficient - would spend about the same amount of time assembling a training set for an algorithm as it takes to run it.

At present, responsive classification is seen in Zooniverse projects such as Space Warps, Planet Hunters and Disk Detective as taking place through the discussion forums provided. In these venues, subsets of volunteers have proved capable of developing detailed and complex classifications of problematic objects, but using this information in a programatic way is hard. Future projects should probably look to design modes of interaction that lie between hard-wired classification and free discussion if this sort of project is to be truly useful.

5 MACHINE LEARNING

6 EDUCATION

Existing projects have demonstrated the potential for outreach and education using citizen science, which has been recognised by its inclusion in the planned effort for surveys such as LSST. Contributors to Galaxy Zoo demonstrated an increase in knowledge typical of Astronomy 101 courses, and the fact that this must have involved looking beyond the Galaxy Zoo site suggests that the projects are motivating self-study or seeking out of scientific content elsewhere. The language used by experienced volunteers in discussion is also significantly more technical than that used by recent arrivals. A survey of Zooniverse volunteers showed an increase in (self-reported) visits to science museums and consumption of scientific content over time. However, use of authentic citizen science

7 RECOMMENDATIONS AND CONCLUSIONS

1. We expect that while machine learning advances will allow algorithmic approaches to solve many common or simple data driven problems, there will still be problems that require human intervention. Examples include anomaly detection, the identification of rare objects, or cases where there are a wide variety of morphologies. For many well-developed science cases, even algorithms approaching 99% accuracy will be insufficient to prevent significant missed discoveries in future survey volumes.
2. The improvement we expect and want to see in machine learning will come about in many cases from the development of large training sets of the kind provided by citizen science. This sort of validation and assessment of machine learning will be needed while surveys are operational.
3. We expect citizen science to be useful throughout the process of data collection and analysis; we need to explore how to tie different tasks together, allow volunteers to work at different levels and to be proactive in leading their own investigations and in shaping the scientific process. Tool development beyond simple classification interfaces will also be needed.
4. The most challenging class of projects given our current technology is that in which continuous inspection of large data sets in near-real time is required for exploratory analysis. In this case, dynamic combinations of human and machine classifications will be needed, and projects which can explore how to do this well should be prioritised in the next few years.

5. Automated extraction of the knowledge contained within broader discussions of unusual objects or new classifications by citizen scientists (e.g., within Zooniverse Talk, an open discussion forum parallel to directed classification tasks) provides the ability to enhance and expand on classifications and annotations predefined by a project. This has the potential to enable open and exploratory data analyses with citizen science projects.
6. Best practices for citizen science:
    a. Simplicity is good; tasks should be understandable and avoid using jargon. The typical interface used by astronomers is not necessarily best for volunteers.
    b. We shouldn't use complex weightings or task assignment unless we have to. If the data rate allows, then the system should be designed to simplify systematics and the selection function.
    c. Citizen science data, like any other data, must be calibrated; gold standard data or simulations must be used to make that calibration.
    d. Data produced by a citizen science project should (after a suitable time) be made public.
    e. It is necessary to engage with the broad citizen science community. We expect science teams who invest time and resources in encouraging education and outreach around their projects to benefit from greater scientific return in addition to impact.
7. Citizen science is a tool for science; attention needs to be paid to data curation such that data is presented in a form suitable both for the volunteers and to enable the scientific returns that are expected.
8. Lack of data sharing is fatal to many interesting investigations, and maximal commensal observing should be part of the standard plan for such surveys. A failure to do this will inhibit projects including cross identification of sources and the development of new citizen science and machine learning approaches. Real-time classification will only be of use if the policies and resources are available for follow-up.
9. Citizen science itself should be open, and we should facilitate sharing code and cloning projects to allow for rapid iteration. For example, projects set up to find one set of objects can quickly be redeveloped to allow the project to adjust to new (and possibly rare) targets.
10. Citizen science projects have enormous potential for education, and can act as a shop window for the science being carried out by surveys, universities and for citizen science itself. The fact that projects can now easily be created - supported by reuse of datasets - makes bringing project design for all a reality. However work and documentation will be needed to allow educators in both formal and informal sectors to quickly understand how to make use of a complex set of tools.