Anomaly Detection in Large Data Sets driven by Citizen Science

Chris Lintott

Focus areas: Human computer teaming, Computational citizen science, and Citizen Science Trust, Equity, Ethics, and Responsible AI

As datasets get larger, the task of anomaly detection – of finding the object, event or occurrence which brings new insight – is increasingly difficult. Though machine learning is effective at finding the unusual in large datasets, it is difficult to train an automated system to tell us what is interesting. This, then, offers an opening for citizen scientists, who are able to work within a ML-enriched environment to make discoveries. Crucially, this mode of operation gets us beyond thinking of volunteers as mere 'humans in the loop', providing data for training large ML models, but instead into a mode where the interests and abilities of citizen scientists are central, despite the presence of AI.

However, attempts to construct such projects have raised numerous questions which will be discussed in this workshop. Though applicable to all areas of online citizen science, we will draw mainly on experience from within the astrophysical community where we have rich data on patterns of participation and where confronting this challenge in the context of new surveys is essential.

- How do we best use modern AI (including now conventional tools such as CNNs and more modern innovations such as transformers and autoencoders) to filter datasets to be shown to citizen scientists?
- How do we design projects that give citizen scientists agency in controlling and using these AI tools, which may be unfamiliar to them and which are often hidden?
- How do we retain a space, through project design and perhaps the use of tools such as Large Language Models, for casual participation such as that seen in successful Zooniverse projects? How, in other words, do we make these projects open to all, instead of the preserve of a small number of already expert citizen scientists working with professionals?
- How do we ensure that short-term participation is still meaningful? One of the advantages of existing online citizen science experiences is that even those with a small amount of time can feel that they have contributed to a project, and that they have agency. As projects become more complex, this is not necessarily true.
- What tools are needed to enable discussion between participants in such projects? In an environment in which AI makes each encounter with data very individual, how do we still create a common space and sense of purpose?
- What interfaces are needed to the archives and publications produced by funded researchers so that citizen scientists can both publicize and receive credit for discoveries made in this space?

However, with projects such as the Vera Rubin Observatory – which are engaging with citizen science from the beginning - imminent, the time is ripe for a discussion of these ideas derived from principles of ethical, meaningful participation. Combining expertise from domain experts, citizen science practitioners and those with expertise in ML, the outcome of this discussion will be a set of focused research questions which can inform future experiments and projects.