Approaching the Grand Challenges in Citizen Science + ML/AI

The following is grounded in my role as co-PI and now PI for Zooniverse, as well as the influence of Adler Planetarium's practices and lessons around public engagement. Since 2009, Zooniverse has grown to 2.7M registered participants worldwide who have contributed over 800M classifications. The 400+ projects have led to numerous discoveries, policy impacts, and over 450 peer-reviewed publications.

Advancements in ML/AI have transformed our ability to process and analyze vast datasets. The convergence of citizen science and ML/AI technologies offers unprecedented opportunities for integrating public participation with cutting-edge computational methods. But also see risks/rewards below.

ML applications on our platform include:

- Using crowd-generated labels as training data.
- Workflows for volunteer review or validation of machine-generated outputs.
- Human-in-the-loop systems.
- Combining multiple approaches to enhance efficiency/scalability while optimizing human effort.

Recent innovations include:

- **Correct-a-Machine Infrastructure**: Volunteers generate data to train ML models, which then produce predictions on new data. Volunteers correct these predictions, creating a feedback loop that refines the model. This infrastructure has been applied across disciplines, from historical text transcription to annotating electron microscopy images.
- **3D Image Annotation**: Utilizing the correct-a-machine framework for complex 3D image data, such as tracing neuron pathways in brain images. Volunteers choose between using predicted annotations or creating their own.

Risks and Rewards - The intersection of ML and citizen science presents both risks and rewards:

- **Over-optimization**: Risks decreasing volunteer enjoyment but can allow participants to focus on more compelling tasks if designed wisely.
- **Lack of content expertise**: Can lead to decreased data quality, but better models can enhance overall data quality.
- **Scalability**: Might cause data analysis bottlenecks but also increases the potential for discovery.
- **Transparency**: Lack of transparency can erode trust, while better transparency can build trust and improve public understanding of ML/AI.
- **Shared infrastructure**: Can lead to inappropriate use by unqualified teams but also democratizes access to ML, fostering inclusivity in scientific research.

Considerations for Policymakers to harness the full potential of ML/AI in citizen science:

- **Transparency**: Critical for ethical volunteer engagement and knowledge-building. Transparency about methods, data handling, and publication processes is essential.
- **Investment in Shared Infrastructure**: Enhances scientific replicability and allows resources to be reinvested in innovative uses of existing technologies.
- **Pluralism**: Ensuring diversity in participant demographics to avoid bias in crowd-generated data. Projects must actively work to include diverse audiences beyond just offering multilingual options.