ATLAS/Pan-STARRS Eyeballing for the Rubin Reliability Factor

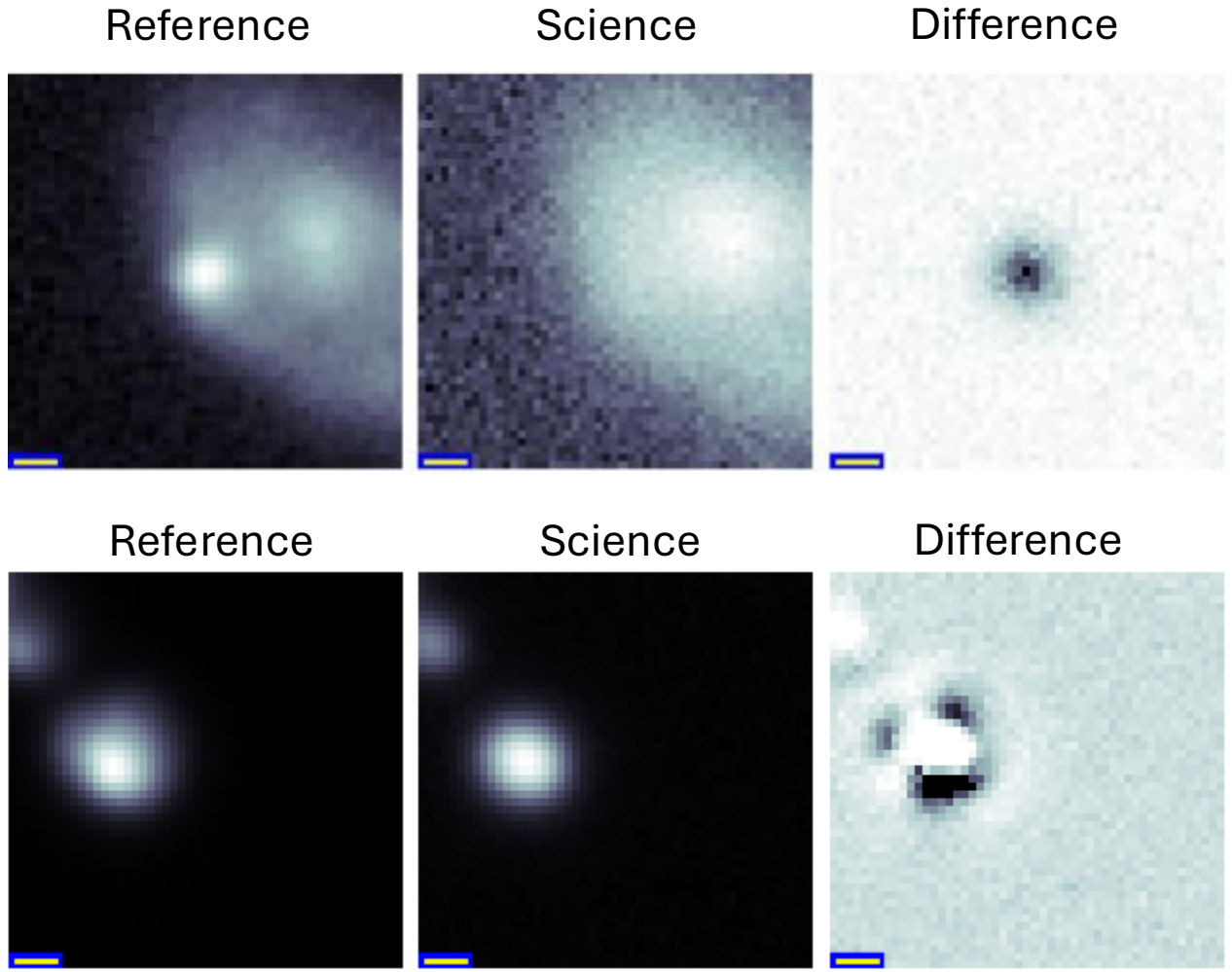
J. G. Weston, *T. Acero-Cuellar, S. J. Smartt, F. Bianco, E. Bellm, C. Lintott, F. Stoppa, H. F. Stevance, K. W. Smith, D. J. Magill, L. Eastman, X. Sheng, M. Nicholl*

Talk Slides

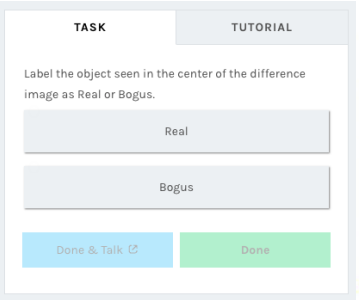
jweston04@qub.ac.uk

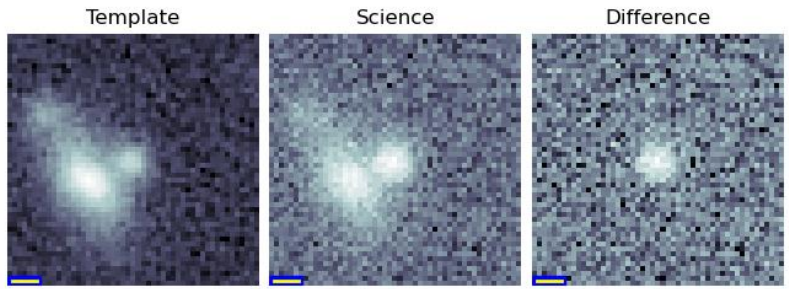
Rubin Reliability Factor

- The Rubin alert stream utilises a real/bogus ‘Reliability Factor’ to provide a prediction of an alert’s realness based on image stamp data.
- The Reliability Factor (RF) is generated by a Convolutional Neural Network, which is trained on and takes science/reference/difference images of a given alert to calculate a score.
- Training of a high-quality RF requires a high-quality training set – with a large number of accurate labels.



Rubin Difference Detectives





- **215,345 subjects** passed to citizen scientists via Zooniverse for real/bogus classification. 2,523 volunteers provided labels over 2 months to classify the entire dataset.
- In parallel, a team of experts labelled **55,161 subjects** over the course of 3 months on a separate project.

Tatiana Acero-Cuellar

PhD Student, UoD

Federica Bianco

Associate Professor, UoD

Eric Bellm

Associate Professor, UoW

Expert Labellers

Stephen Smartt

Professor, Ox/QUB

Xinyue Sheng

Research Fellow, QUB

Ken Smith

Software Dev, Ox/QUB

Josh Weston

PhD Student, QUB

Heloise Stevance

Research Fellow, Ox

Dylan Magill

PhD Student, QUB

Fiore Stoppa

Research Fellow, Ox

Lauren Eastman

PhD Student, Ox

- Expert Labellers have experience in ATLAS/Pan-STARRs image stamp classification and/or image stamp machine learning development.
- For citizen scientists, a subject was retired if the first two labels were in agreement. If they disagreed, five more labels were obtained.
- For experts, a subject was also retired if the first two labels were in agreement. If they disagreed, one more label was obtained.

Data Selection

The first tranche of data released to the citizen science project consisted of six subsets from the following groups:

- **Solar System Matches**: One set of Rubin Alerts cross-matched with an internal Solar System object catalogue, within one arcsecond.
- **Gaia Matches**: One set of Rubin Alerts cross-matched with Gaia star catalogues objects, within one arcsecond.
- **Transient Name Server (TNS) Matches**: Three sets of Rubin Alerts cross-matched with reported transients on the TNS, within one arcsecond.
- **Reliability Factor Subsets**: One set of Rubin Alerts assigned a Reliability Factor score $0.5 < RF < 0.9$.

The **expert project classified a 10% sample of the subsets above**, as well as additional subsets. For the construction of the expert data, a **random 10% sample of all subsets was taken**. For subsets with a total number of subjects below 1000, the full data was downloaded, providing 105,304 subjects in total for classification.

Measuring Agreement

Cohen's Kappa (κ) measures inter-rater agreement for labelled data, correcting for agreement occurring by chance. κ is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

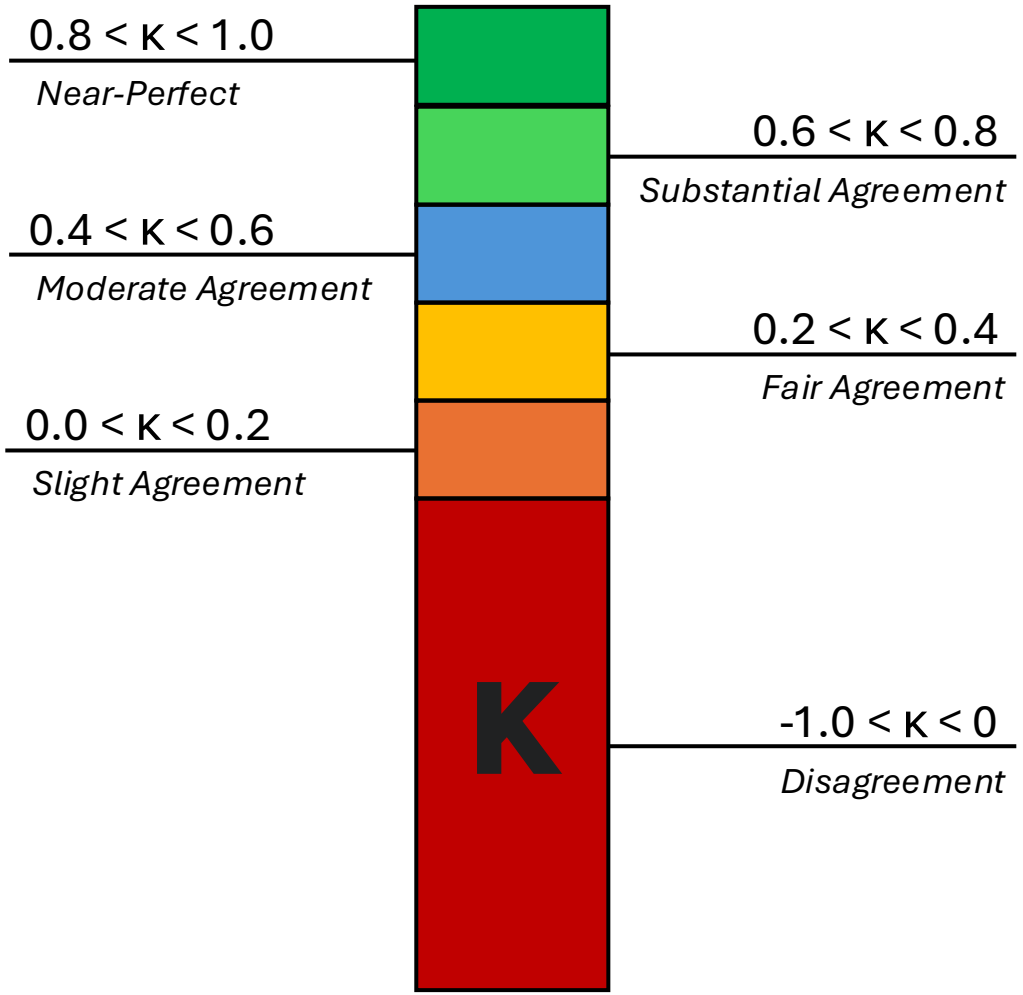
P_o = fraction of items that both labellers agreed upon.

P_e is the expected agreement proportion by chance:

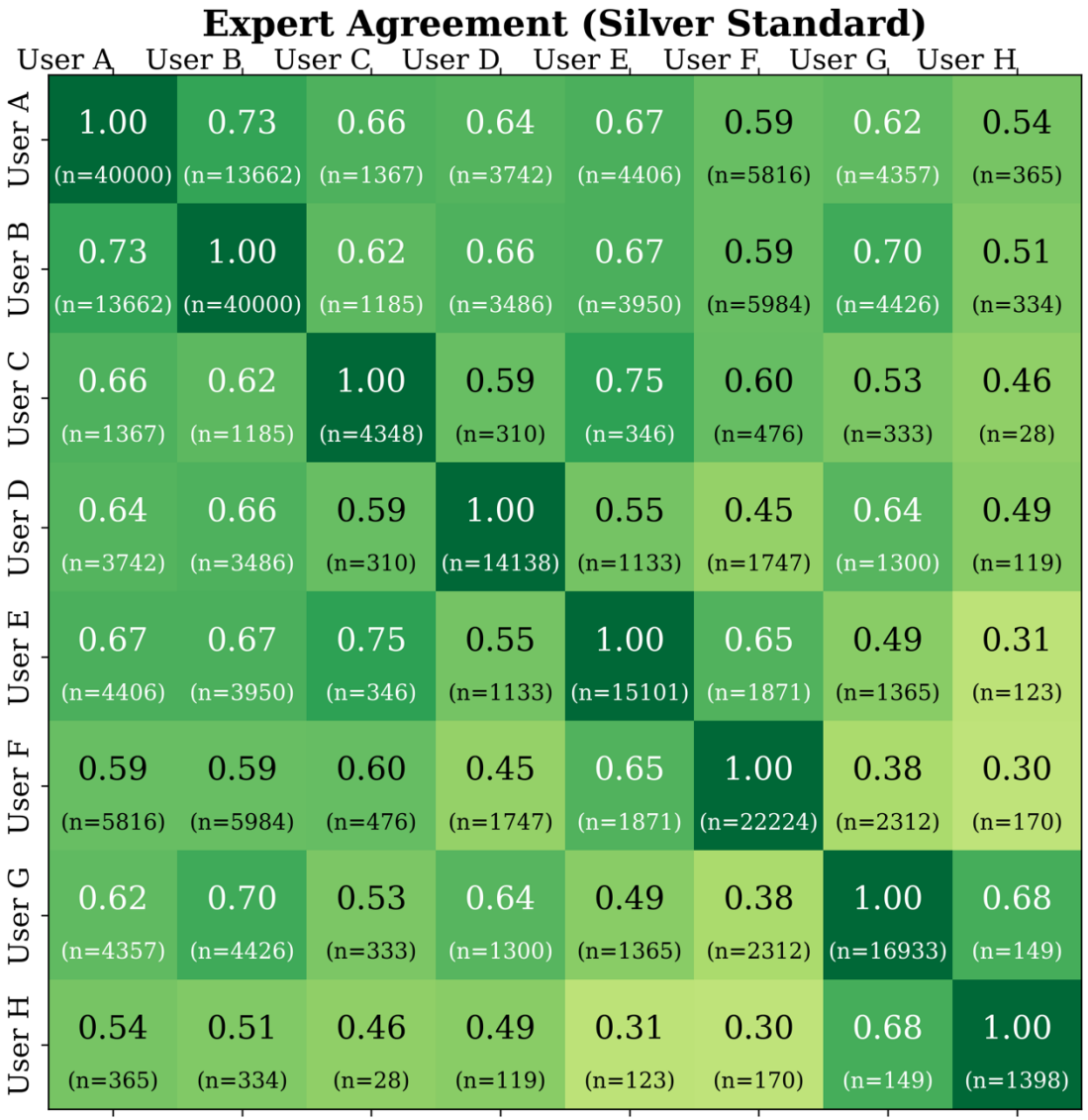
$$P_e = \sum_{i=1}^{k} P_{A,i} \cdot P_{B,i}$$

$P_{A,i}$ = proportion of items assigned to category C_i by labeller A

$P_{B,i}$ = proportion of items assigned to category C_i by labeller B



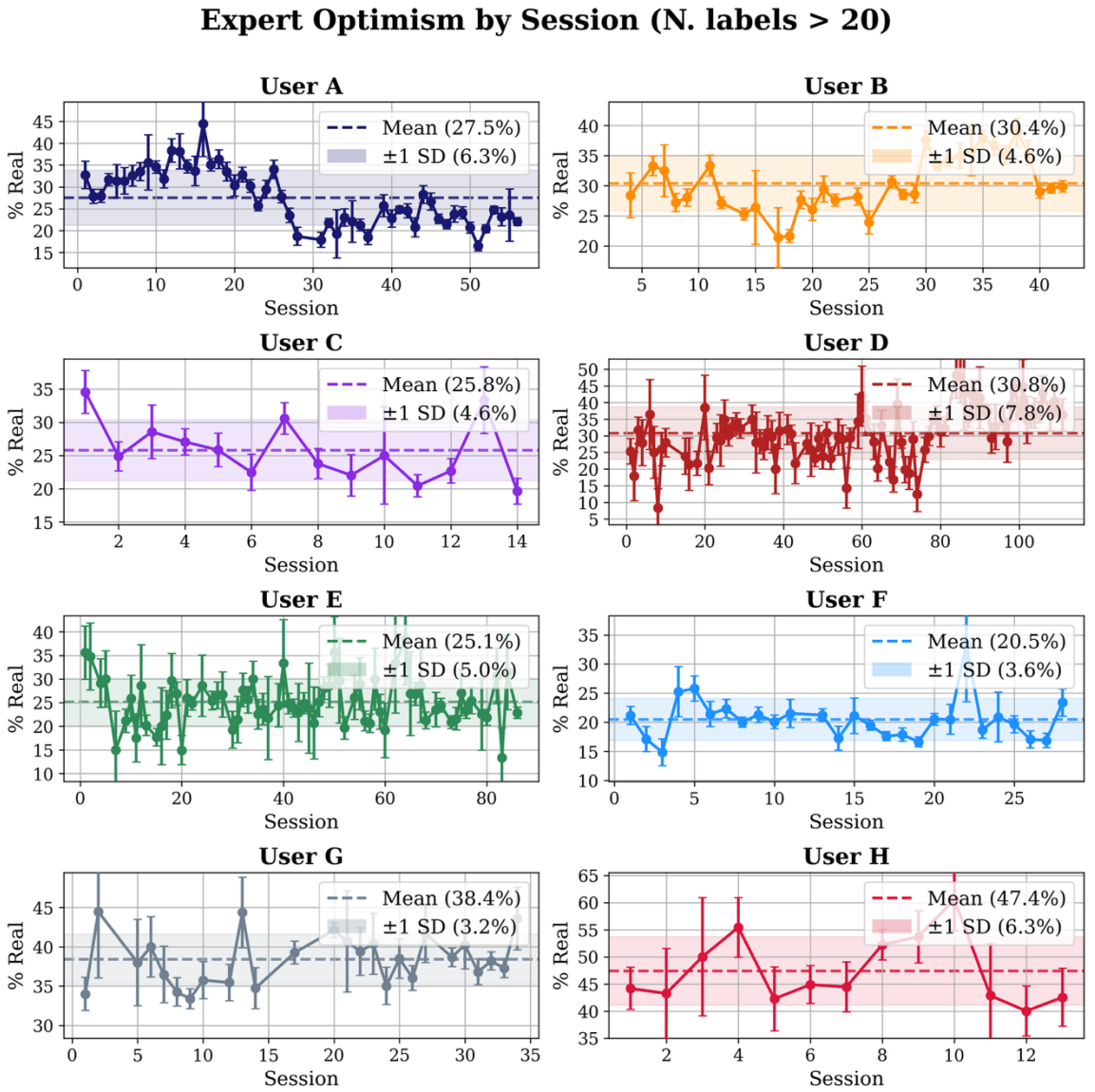
Expert Agreement



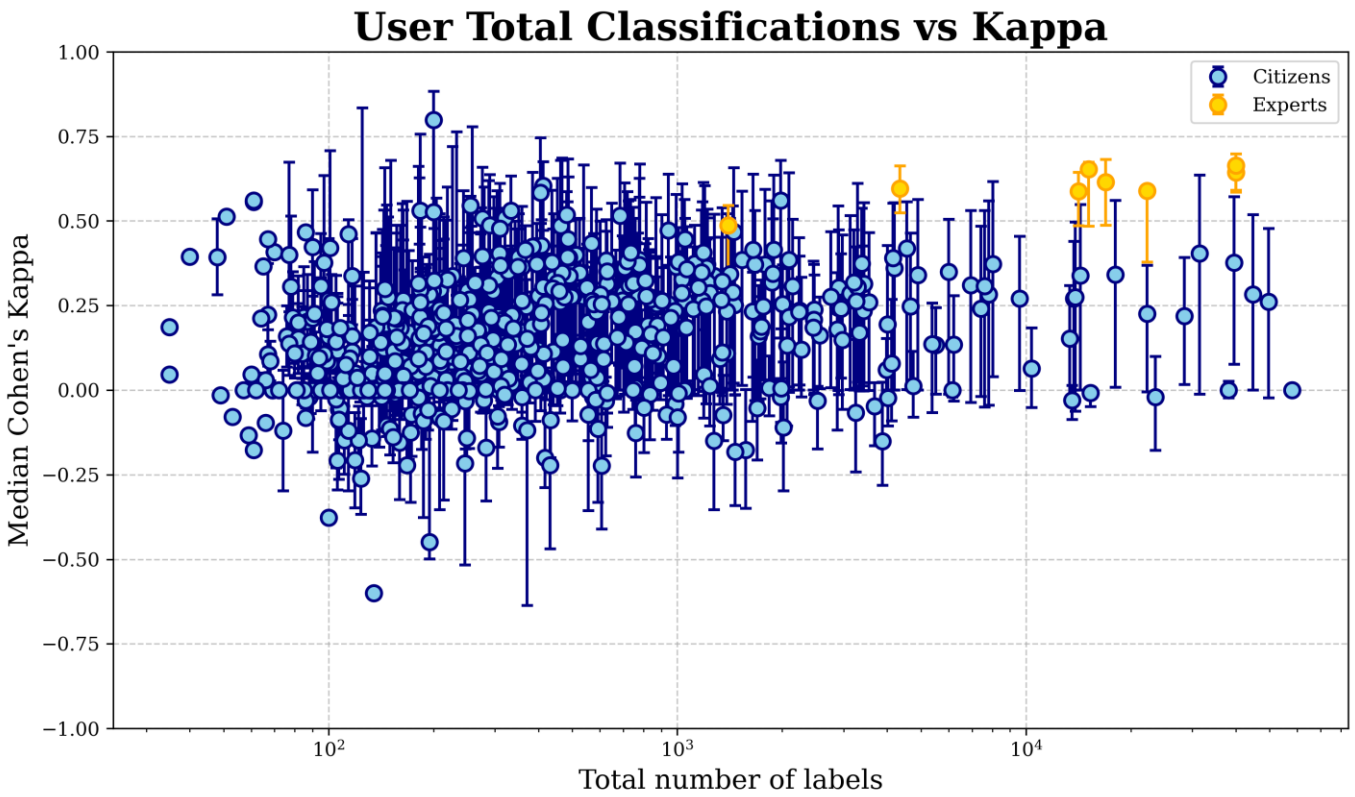
- There is a high level of agreement across experts.
- The least ‘agreeable’ expert agreed on a classification **73%** of the time. The most agreeable expert agreed on a classification **85%** of the time.
- The most common label pairings see a high level of agreement. Users A and B contribute the most joint classifications with a kappa of 0.73 (substantial agreement). Users C and H contribute the fewest joint classifications.
- Two subgroups appear when interpreting agreement. Users A, B, C, D and E have a high level of agreement with each other, with a minimum kappa of 0.55 (moderate agreement). Users F, G, and H agree less often with each other.

Expert Consistency

- There is a ~10% range of optimism between most Users. The optimism level of each User remains broadly consistent across sessions.
- Users who have a low number of classifications per session have the highest noise in their optimism, but most users show no change to their rate of Real labels over time.
- No significant correlation between session length/speed and the % labelled Real.

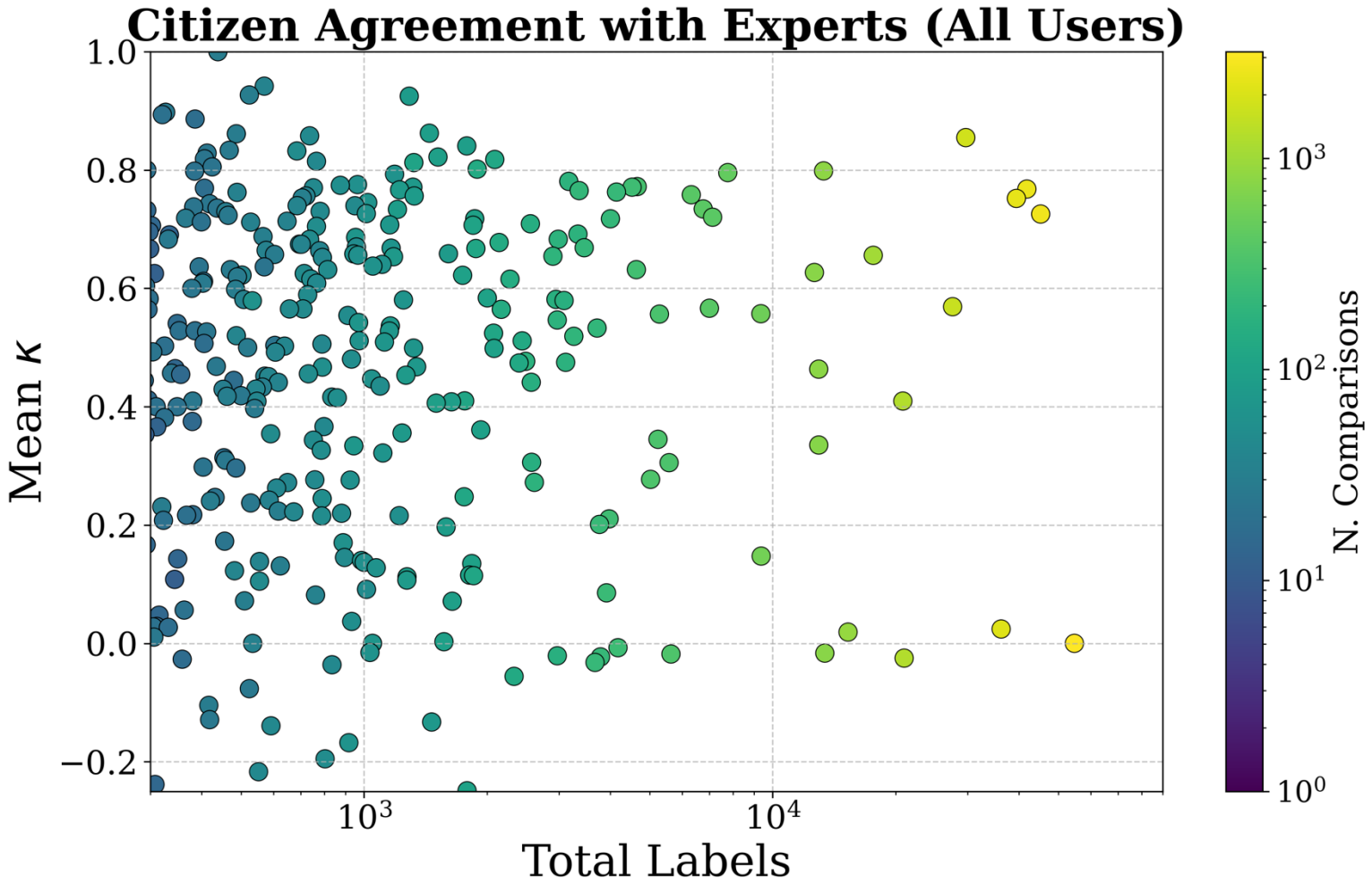


Citizen Science Agreement



- There are varying levels of agreement across citizen scientists. Without filtering, there is a large standard deviation in each user's agreement – indicating many different behaviours in the group.
- Several users have negative kappa values, indicating systematic disagreement.
- The contribution of citizen scientists does not correlate to their agreement level.

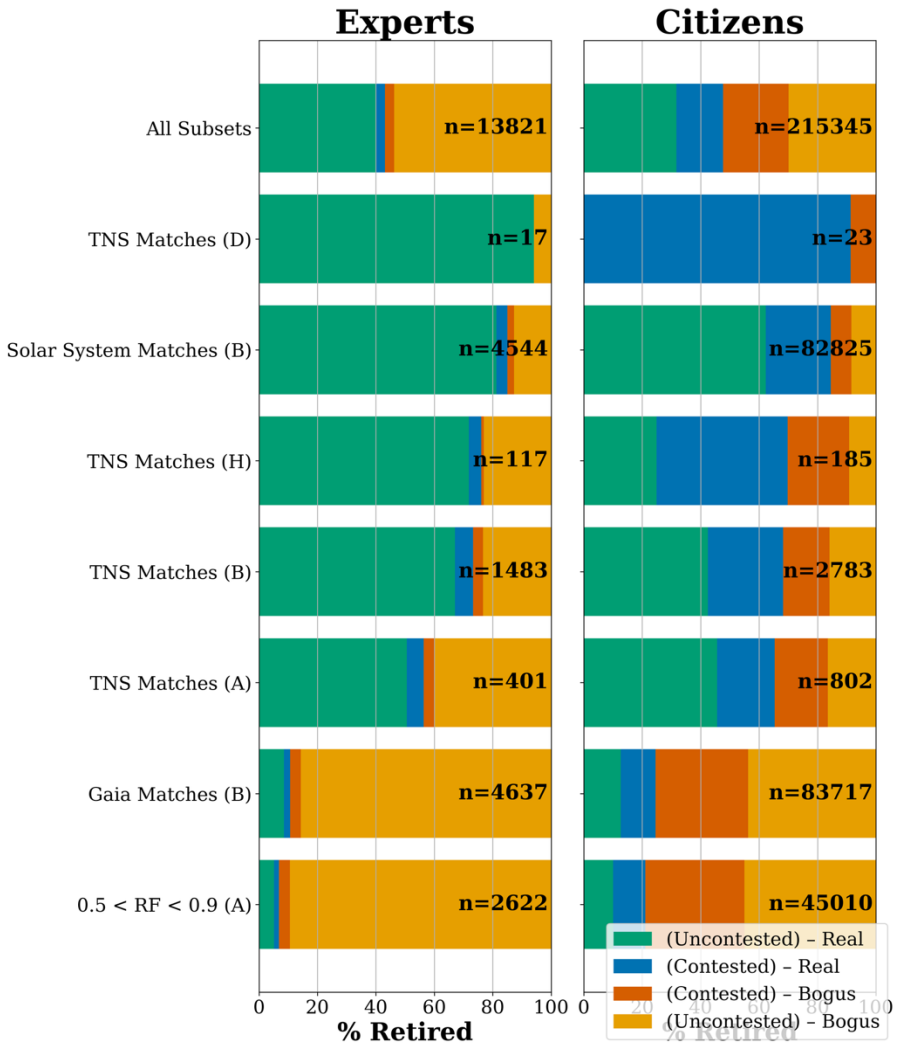
Citizen Science Quality



- Accuracy of citizen science labelling tends to increase with experience.
- Horizontal banding for low-volume users can be attributed to the limited number of subjects available for kappa calculation.
- Noise in the population is largely filtered out above 1,000 total labels. Users with a high overlap can be filtered on to obtain high-quality labels that can increase the size of the gold standard label set.

Citizen Science Subset Real/Bogus

- The percentage of real labels across the sample is 50%, 7% more than in the Expert sample.
- Citizen scientists contest subjects more often than expert labellers. A contested subject is more likely to be real if the subset is majority real, and more likely to be bogus if the subset is majority bogus.
- The largest differences in R/B ratio were in the Gaia and Reliability Factor Subsets.

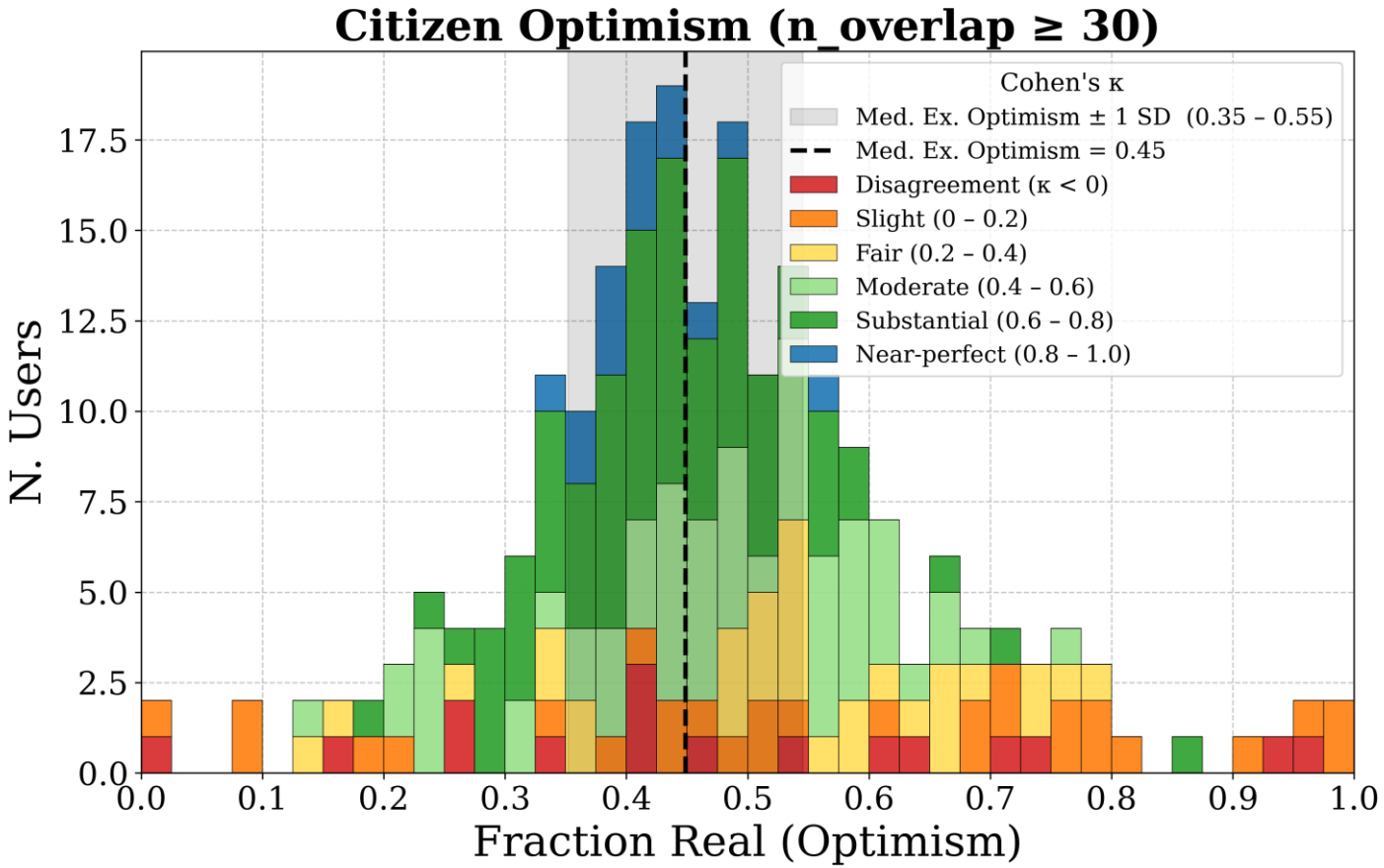


Next Steps

- Filter out the best citizen scientists and compare their performance/behaviours against experts.
- Rubin Alerts Production Team & Tatiana Acero-Cuellar (University of Delaware) exploring use of citizen science & expert labels in Rubin RF development.

Tatiana Acero-Cuellar

PhD Student, UoD



Conclusions

- **The Rubin Reliability Factor ground truth is stable.** Expert labels are consistent and reliable, examining optimism and agreement metrics.
- **High-agreement citizen scientists can augment the ground truth with additional data.** Citizen scientists effectively classify transients at scale, with label quality and quality-to-noise improving significantly with experience.

Talk Materials

Contact (Hire) me