



QUEEN'S
UNIVERSITY
BELFAST



UNIVERSITY OF
OXFORD

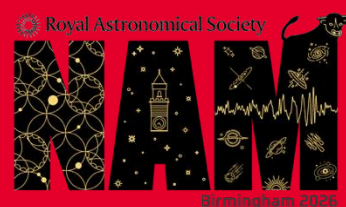
UNIVERSITY OF
DELAWARE

UNIVERSITY of
WASHINGTON

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*

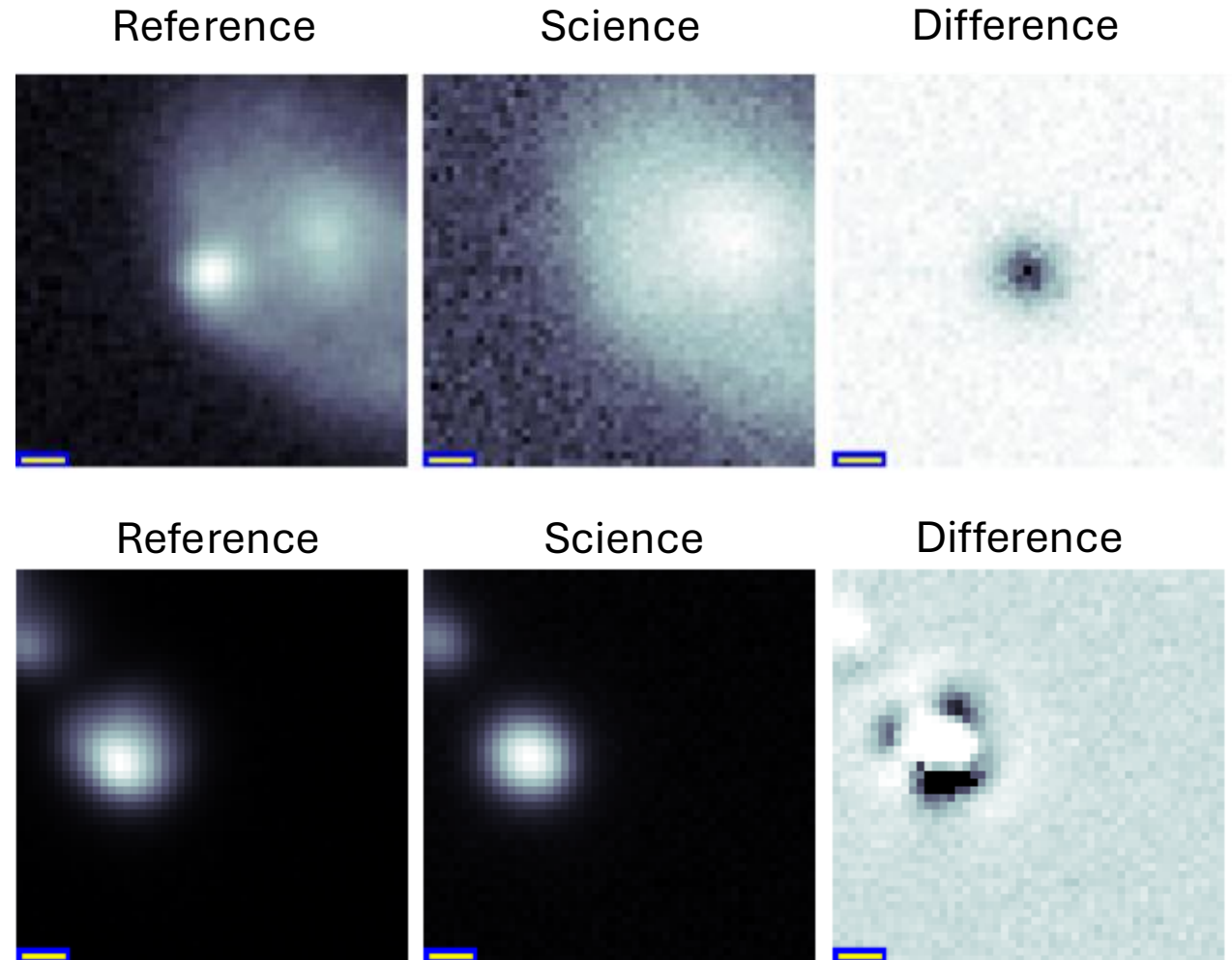
23rd July 2026



Birmingham 2026

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 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.



DMTN-337

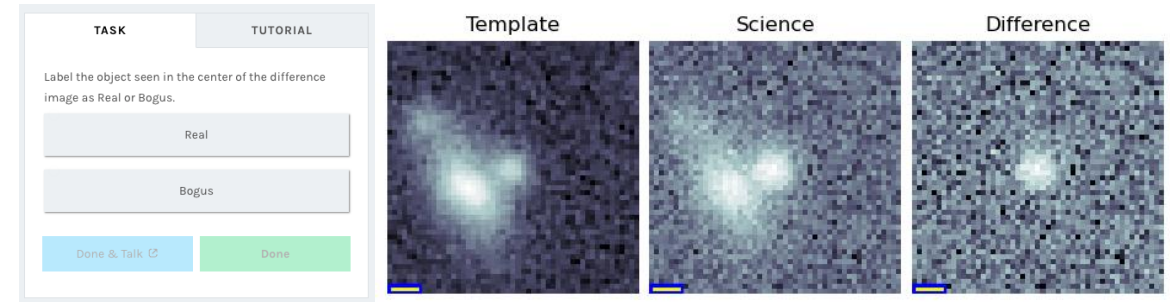


Tatiana Acero-Cuellar

PhD Student, UoD

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.
- For citizen scientists, a subject was retired if the first two labels were in agreement. If they disagreed, five more labels were obtained.
- In parallel, a team of experts labelled **55,161 subjects** over the course of 3 months on a separate project.
- For experts, a subject was also retired if the first two labels were in agreement. If they disagreed, one more label was obtained.



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 $0.5 < \text{RF} < 0.9$.

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:

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

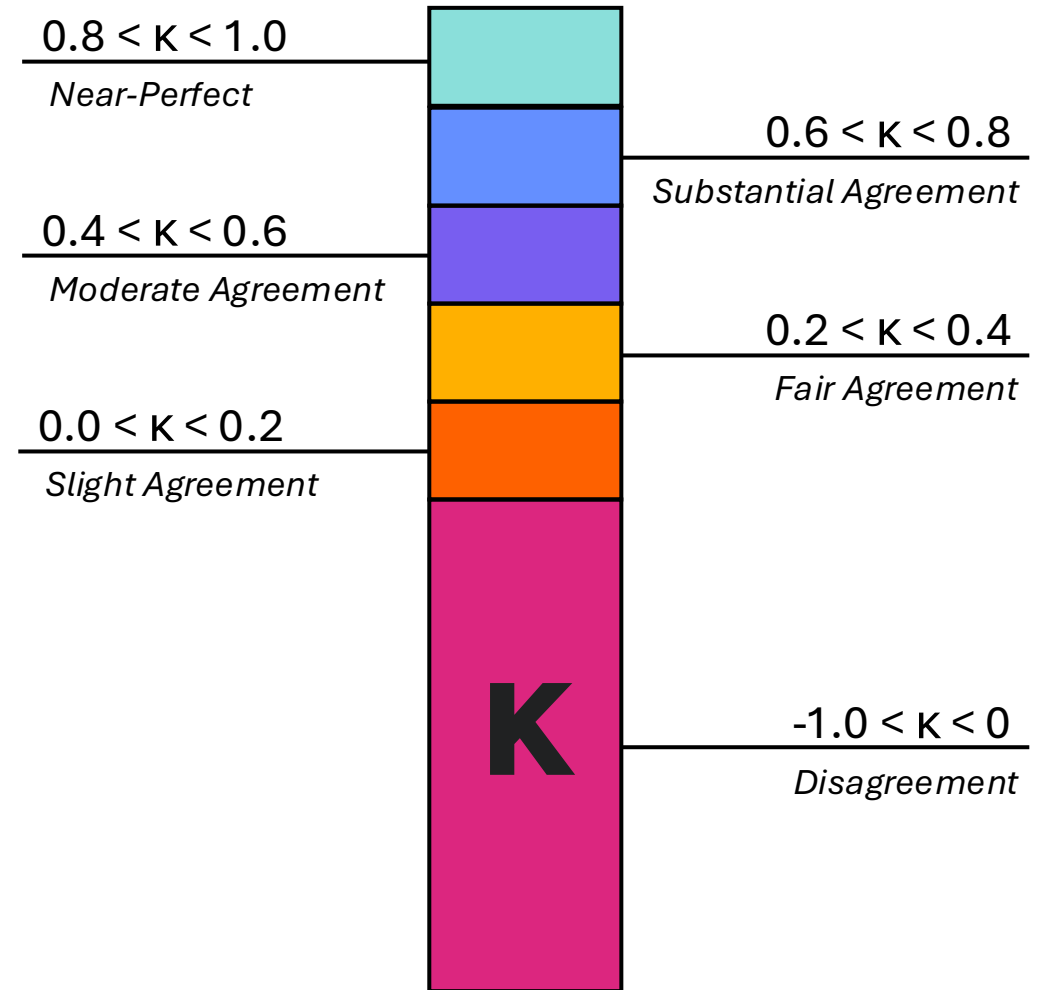
P_o = fraction of items that both labellers agreed upon.

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

Expert Agreement (Silver Standard)

	Alpha	Beta	Gamma	Delta	Epsilon	Zeta	Eta	Theta
Alpha	1.00 (n=40000)	0.73 (n=13662)	0.66 (n=1367)	0.64 (n=3742)	0.67 (n=4406)	0.59 (n=5816)	0.62 (n=4357)	0.54 (n=365)
Beta	0.73 (n=13662)	1.00 (n=40000)	0.62 (n=1185)	0.66 (n=3486)	0.67 (n=3950)	0.59 (n=5984)	0.70 (n=4426)	0.51 (n=334)
Gamma	0.66 (n=1367)	0.62 (n=1185)	1.00 (n=4348)	0.59 (n=310)	0.75 (n=346)	0.60 (n=476)	0.53 (n=333)	0.46 (n=28)
Delta	0.64 (n=3742)	0.66 (n=3486)	0.59 (n=310)	1.00 (n=14138)	0.55 (n=1133)	0.45 (n=1747)	0.64 (n=1300)	0.49 (n=119)
Epsilon	0.67 (n=4406)	0.67 (n=3950)	0.75 (n=346)	0.55 (n=1133)	1.00 (n=15101)	0.65 (n=1871)	0.49 (n=1365)	0.31 (n=123)
Zeta	0.59 (n=5816)	0.59 (n=5984)	0.60 (n=476)	0.45 (n=1747)	0.65 (n=1871)	1.00 (n=22224)	0.38 (n=2312)	0.30 (n=170)
Eta	0.62 (n=4357)	0.70 (n=4426)	0.53 (n=333)	0.64 (n=1300)	0.49 (n=1365)	0.38 (n=2312)	1.00 (n=16933)	0.68 (n=149)
Theta	0.54 (n=365)	0.51 (n=334)	0.46 (n=28)	0.49 (n=119)	0.31 (n=123)	0.30 (n=170)	0.68 (n=149)	1.00 (n=1398)

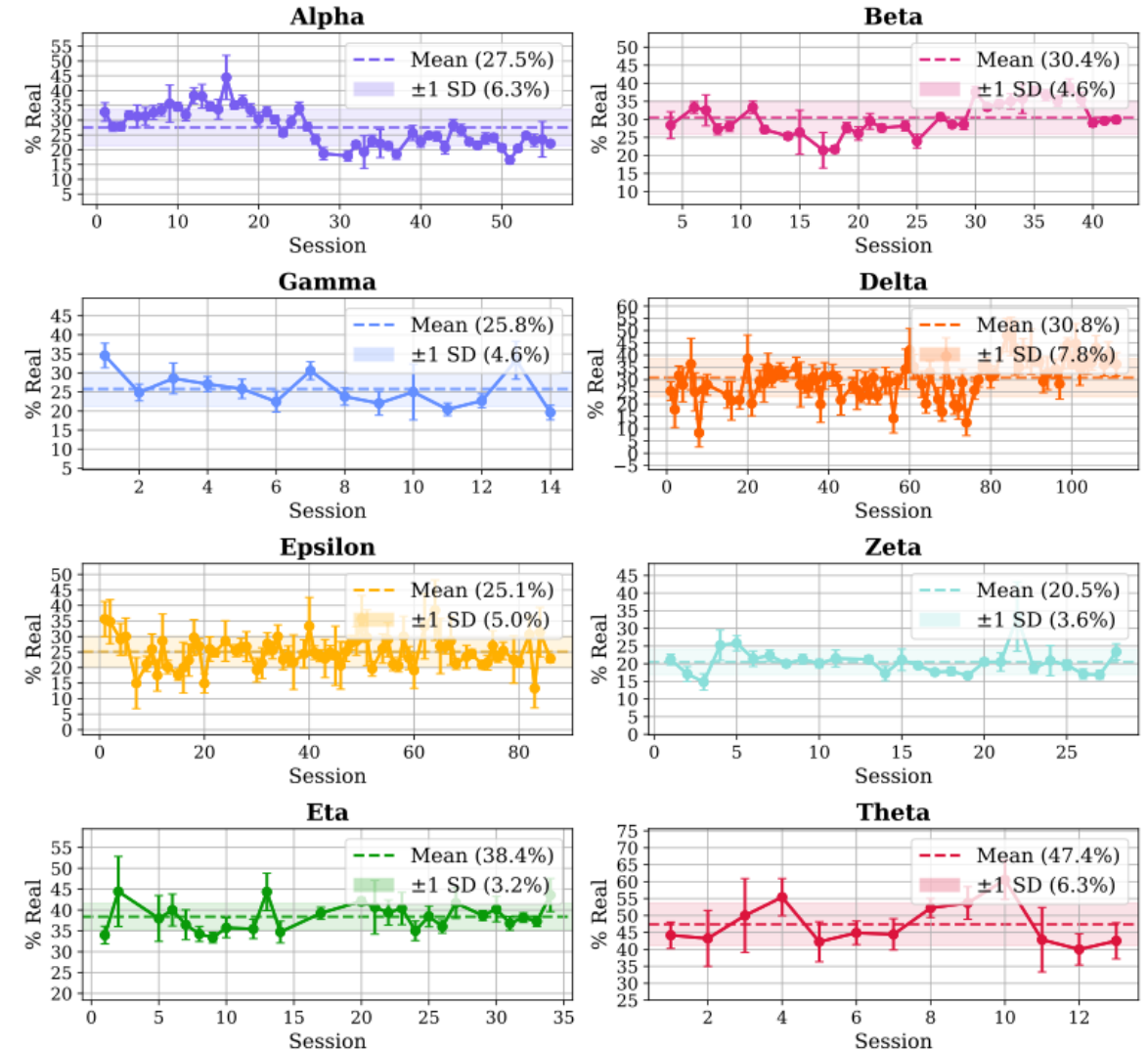
- There is a high level of agreement across experts.
- The least 'agreeable' expert agreed on a classification **73%** of the time. The most agreeable expert agreed on a classification **85%** of the time.
- The most common label pairings see a high level of agreement. Users A and B contribute the most 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.



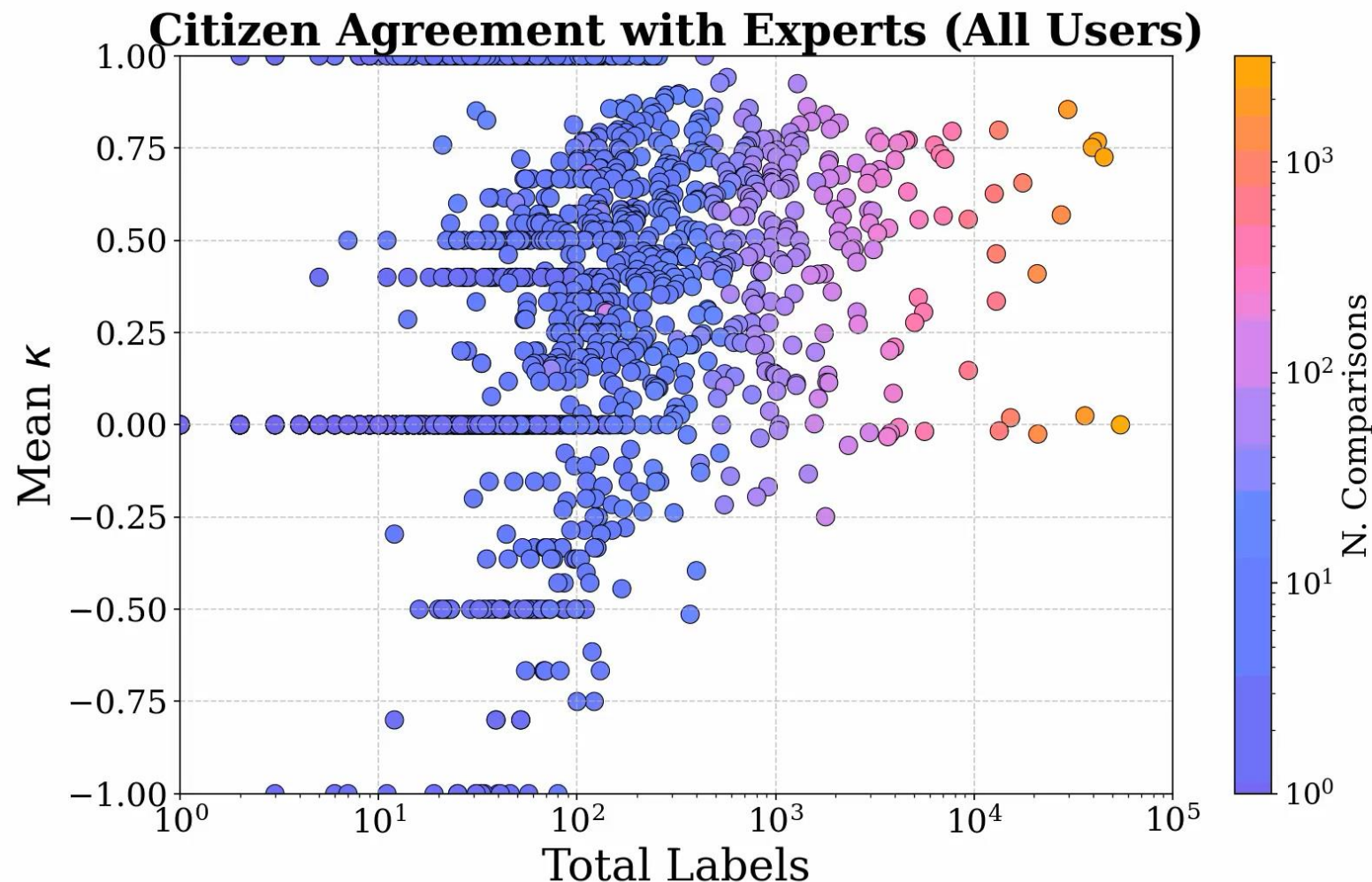
Expert Consistency

- 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.

Expert Optimism by Session (N. labels > 20)



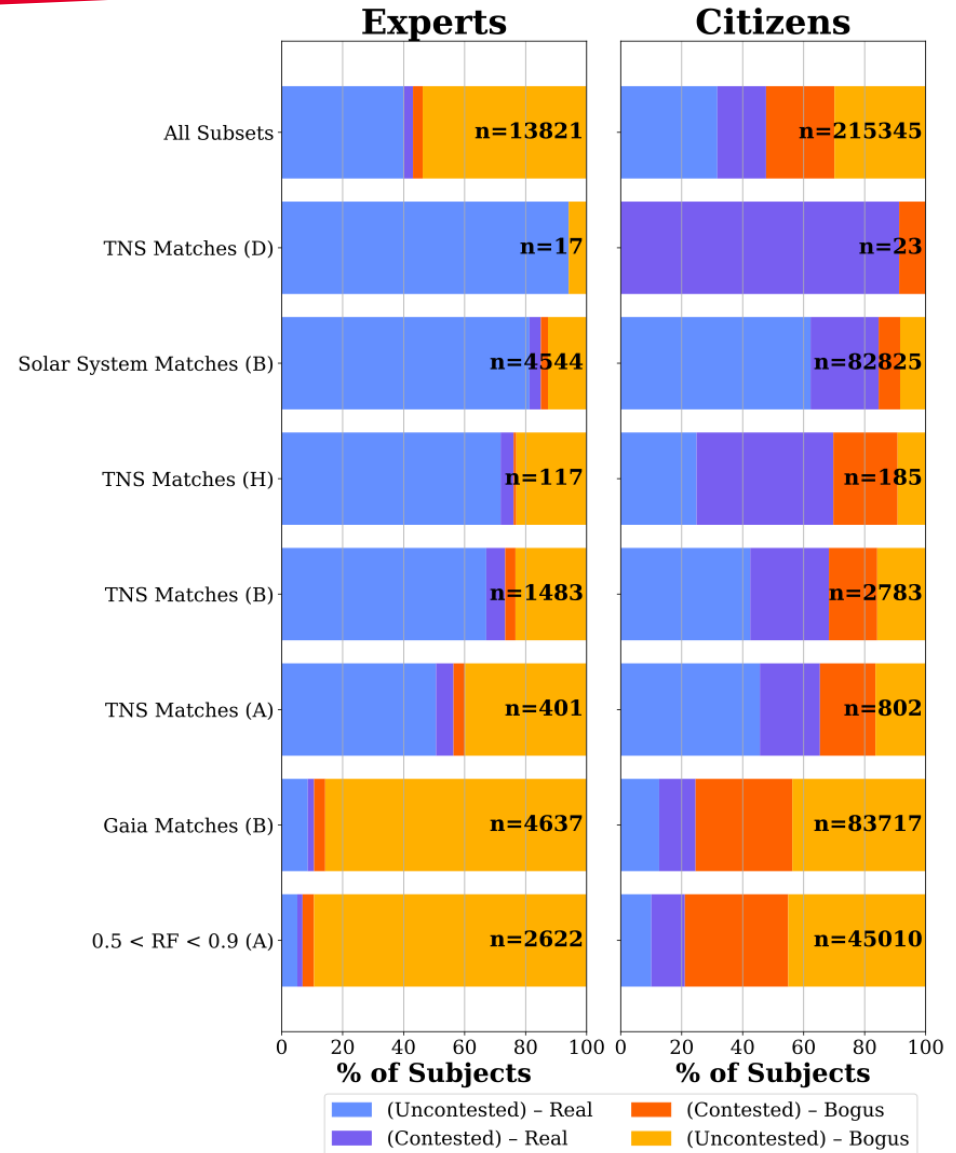
Citizen Science Quality



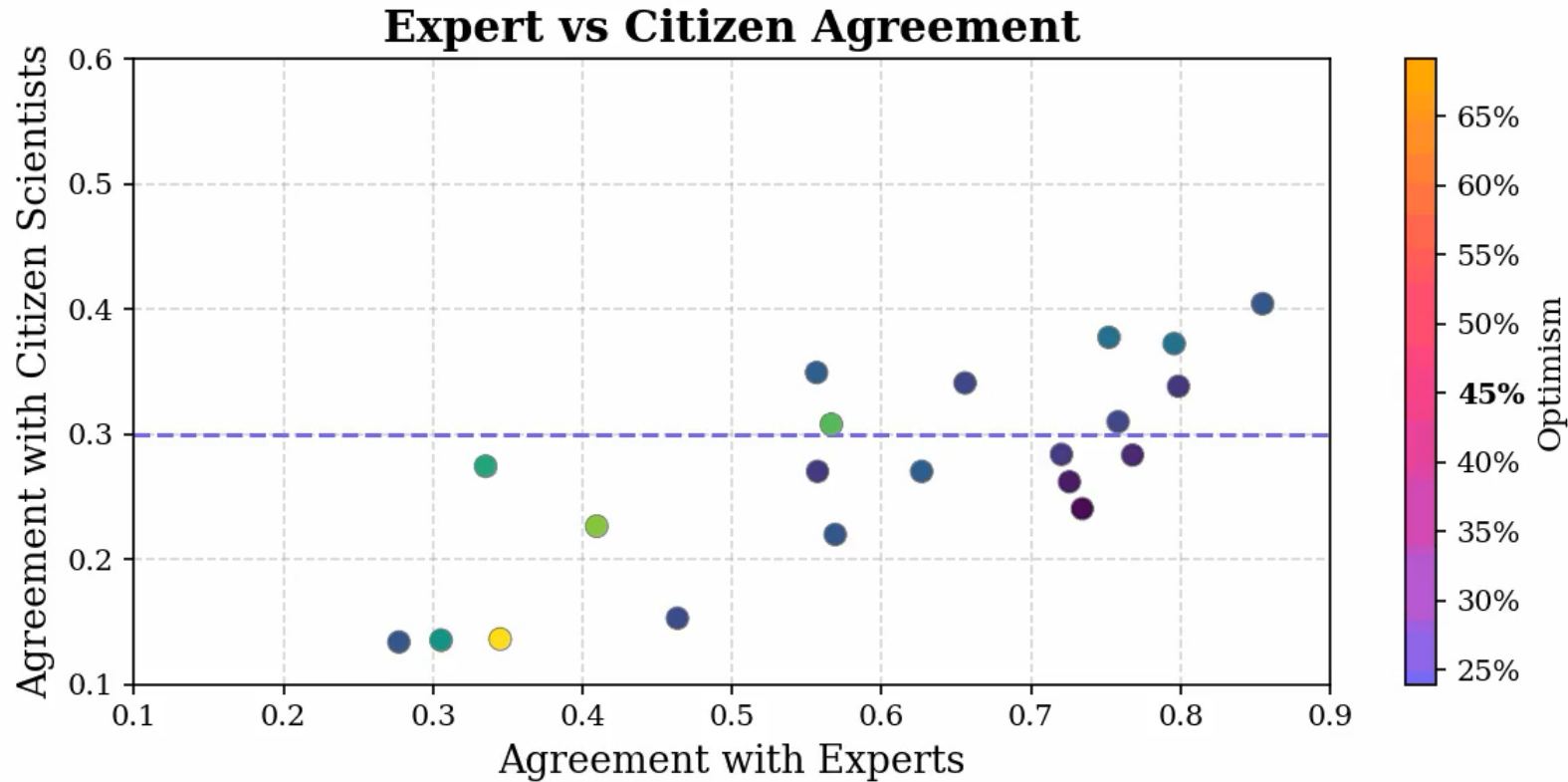
- Accuracy of citizen science labelling tends to increase with experience.
- Horizontal banding for low-volume users can be attributed to the limited number of subjects available for kappa calculation.
- Noise in the population is largely filtered out above 1,000 total labels. Users with a high overlap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

Citizen Science Labelling

- 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.



Which Citizens?



- Agreement with expertise correlates with agreement among other citizen scientists.
- Recalculating agreement between the 'top' citizen scientists increases the number of accurate volunteers above a given threshold.
- We can filter for conformity to identify reliable labelling in bulk.

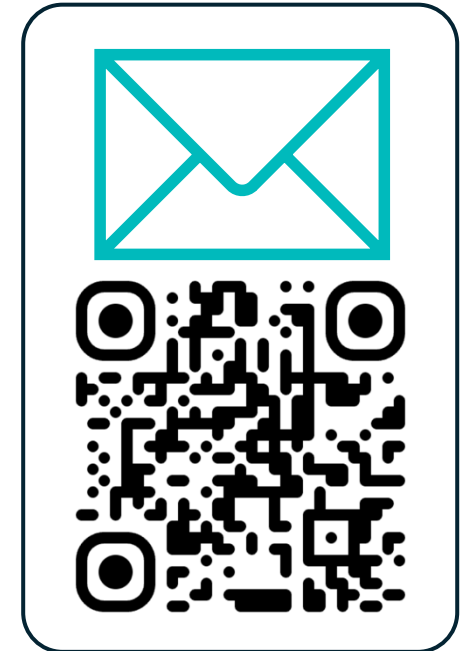


Conclusions

- **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.



DMTN-337



Contact Me