

CLARA

Confidence of Labels and Raters

Viet-An Nguyen, Peipei Shi, Jagdish Ramakrishnan, Udi Weinsberg, Henry C. Lin, Steve Metz, Neil Chandra, Jane Jing

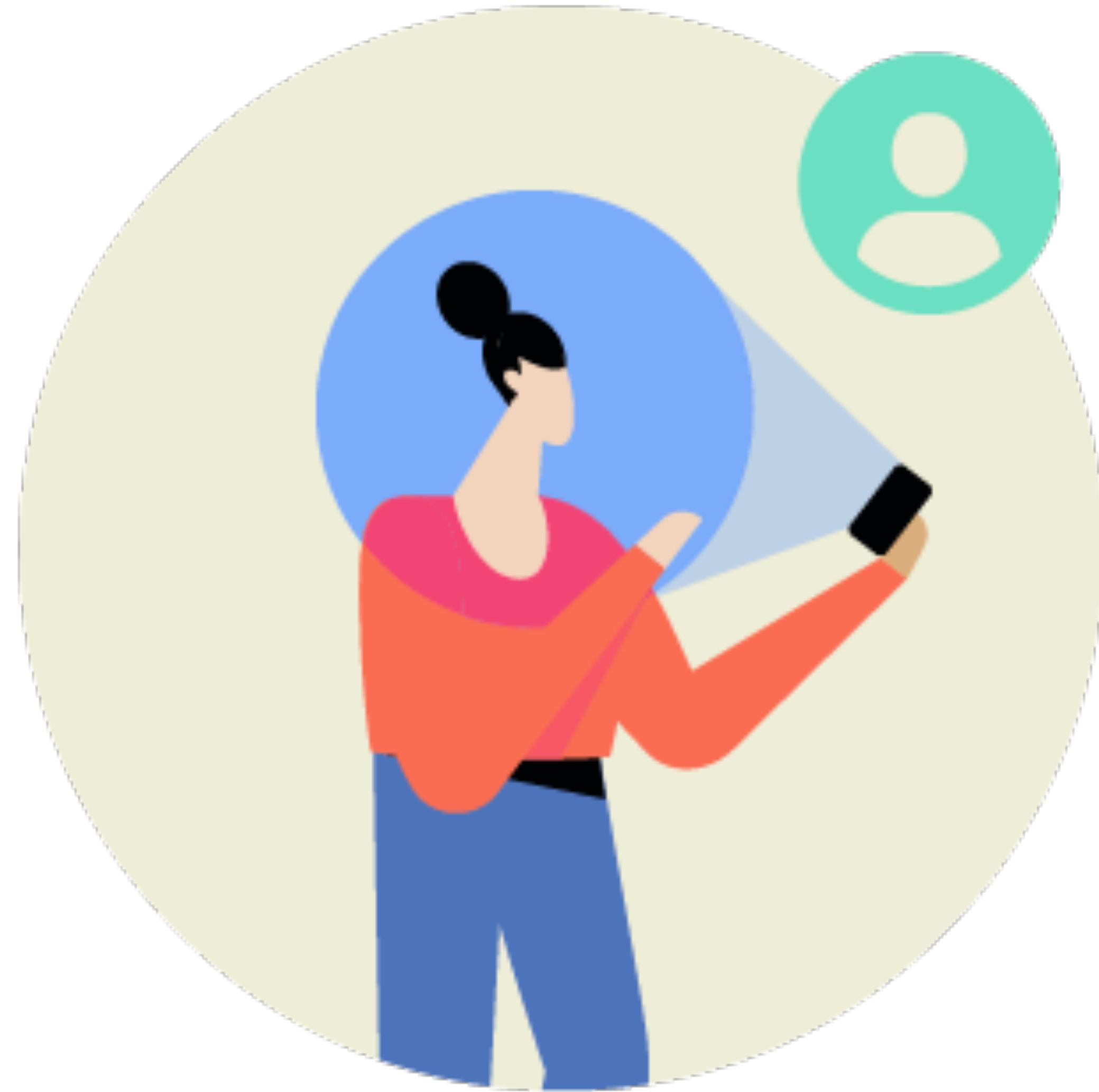
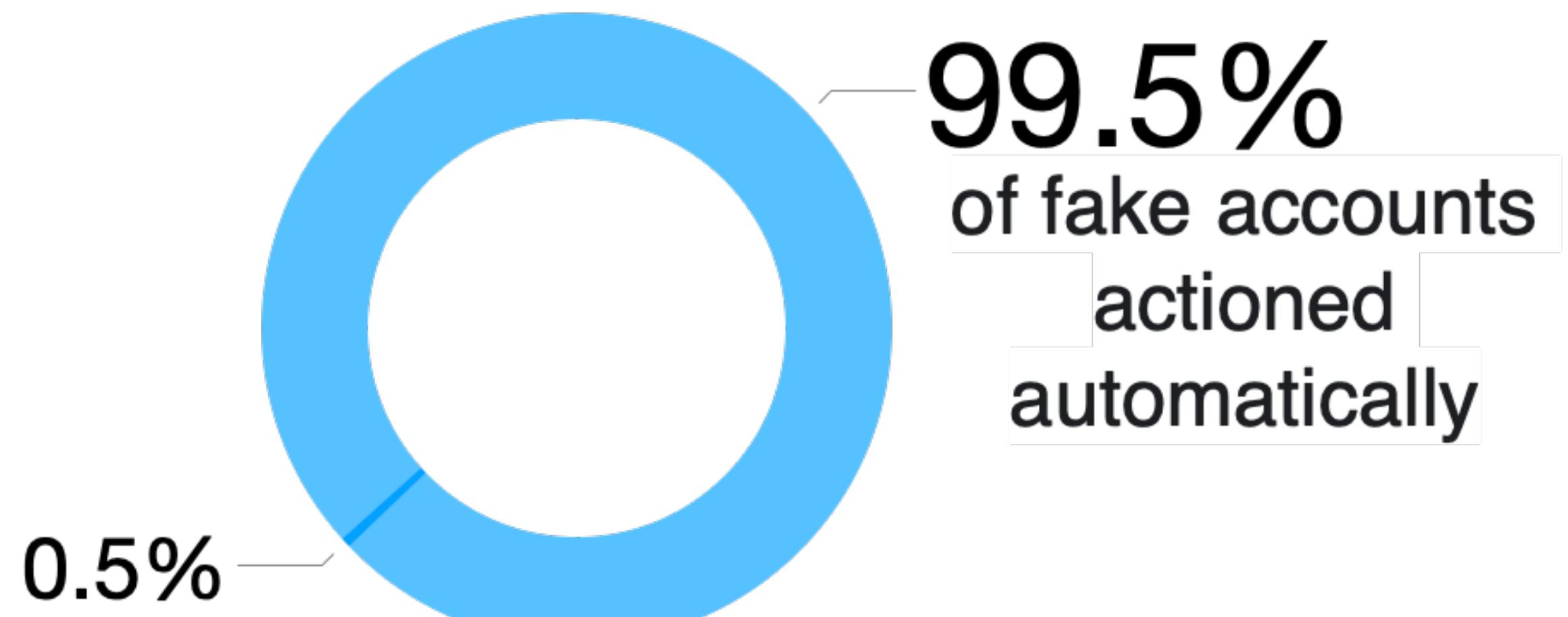
Facebook

Dimitris Kalimeris

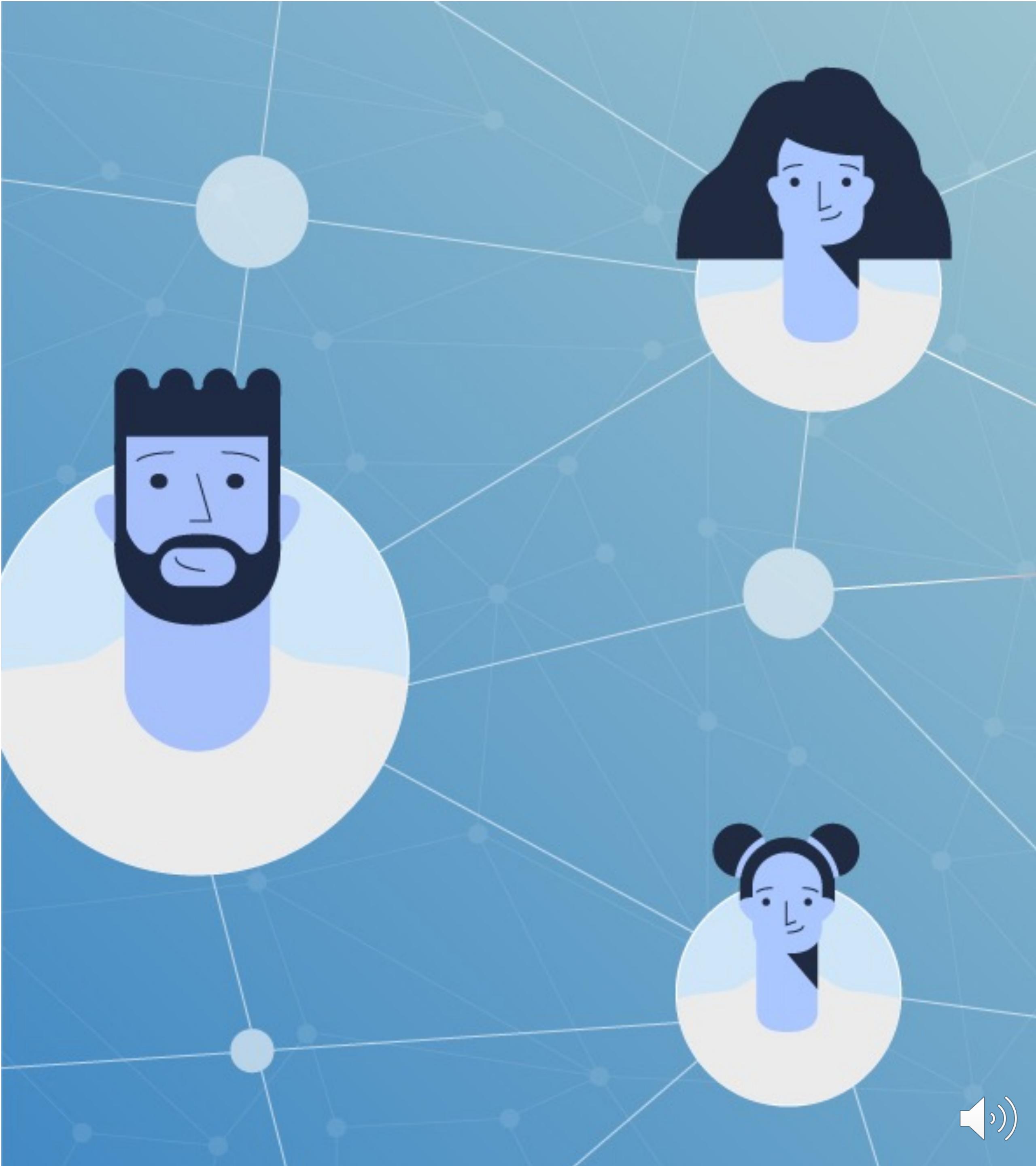
Harvard University



Automation is essential for large-scale systems



**However, there
are many areas
where human
decisions are
needed.**





**It is sometimes
hard for humans
to make
decisions**

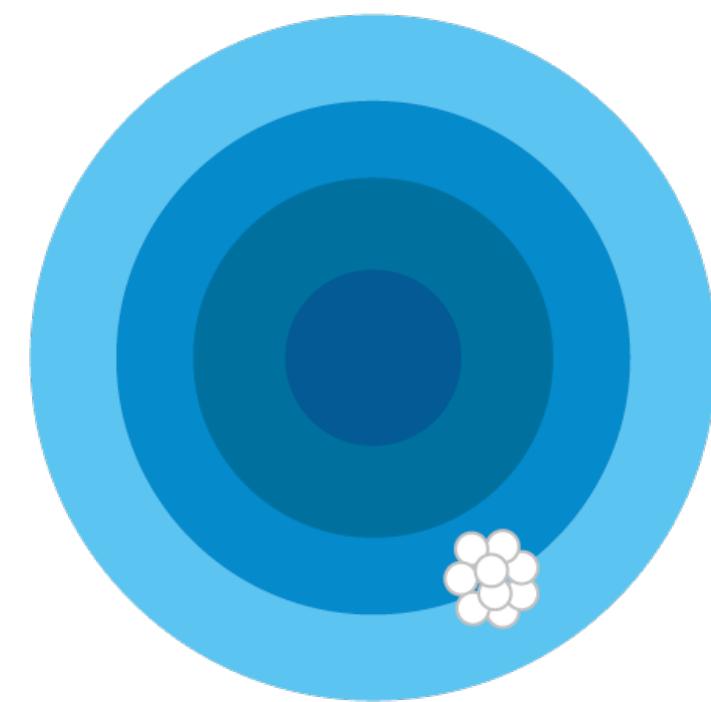


What makes people noisy decision makers?

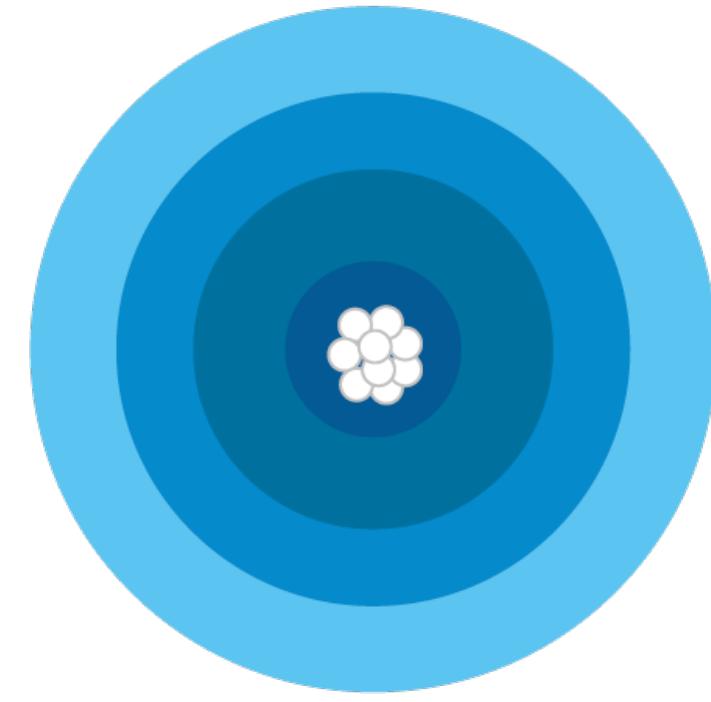


Individual bias

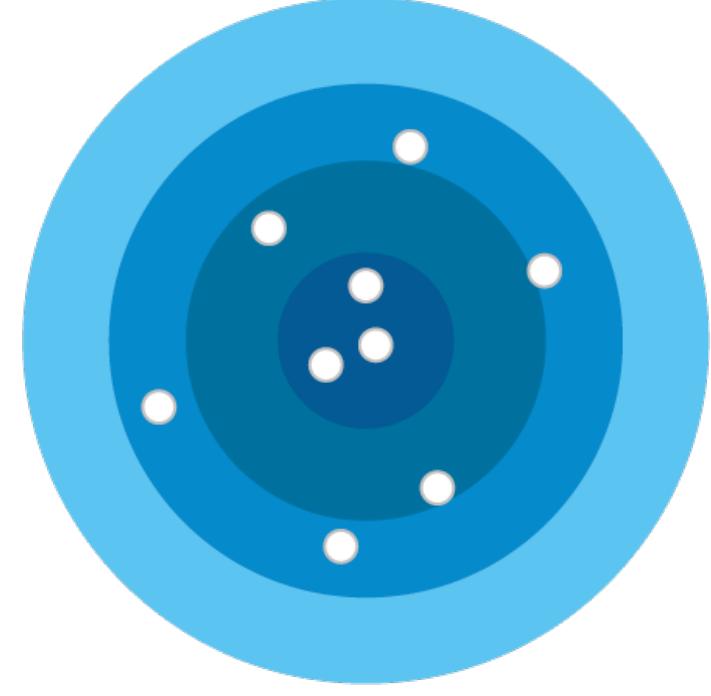
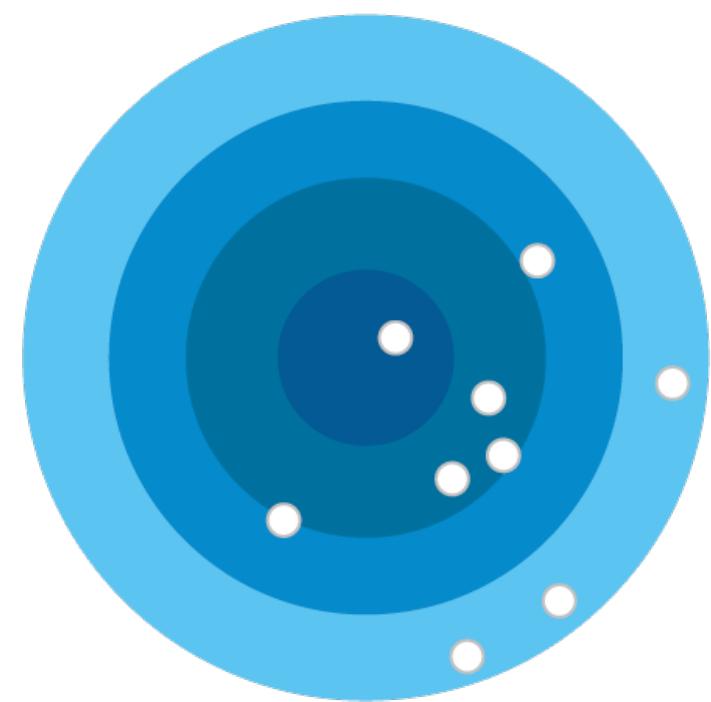
More Bias



Less Bias



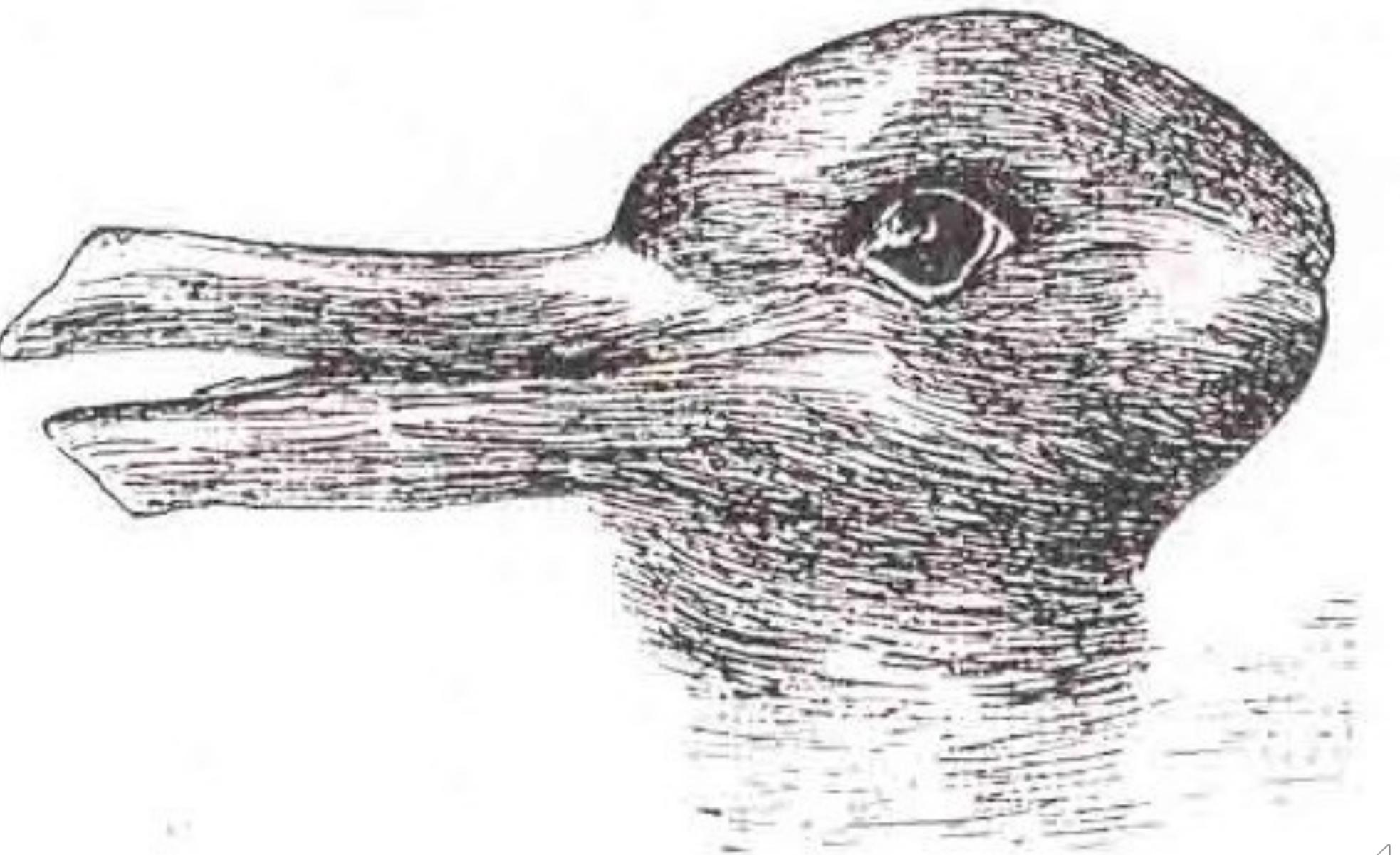
Less Variance



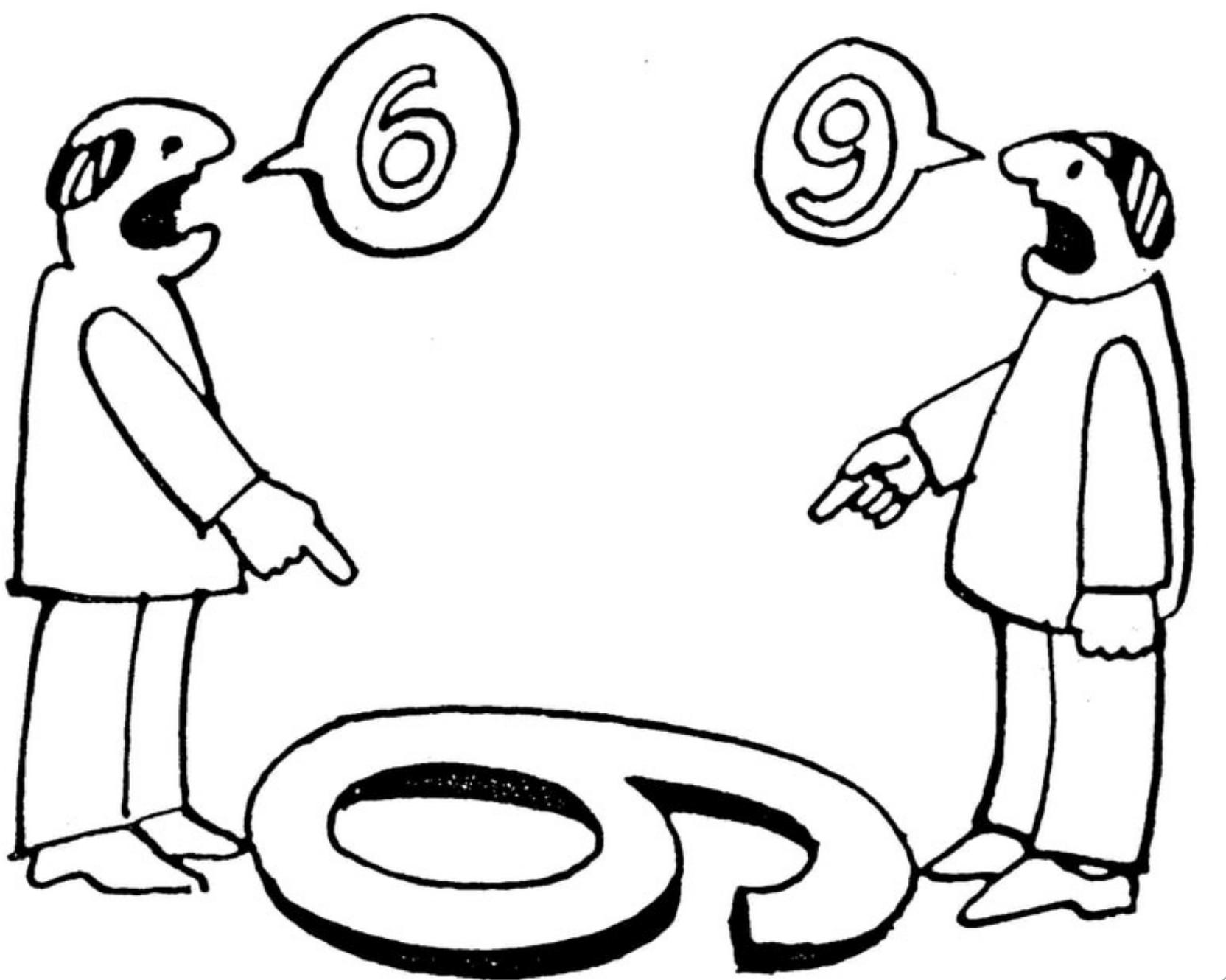
More Variance



Ambiguity of the guidelines



Subjectivity in the decision



Simple mistakes

mistake



**Let's consider a few examples where
noisy decisions have negative outcomes**



PREVALANCE

The percentage of policy-violating content out of all content seen by Facebook users.



ENFORCEMENT

Taking down content
or entities that violate
the community
standards.



TRAINING MODELS

Using human-generated labels as “ground truth” for training ML models.



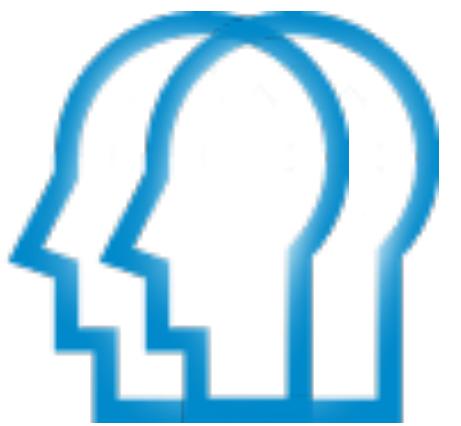
**So how should we deal with noisy
decision making?**



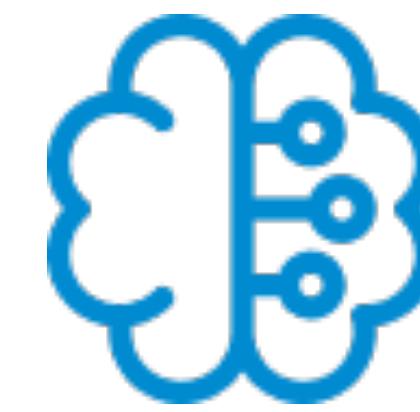
Hire
experts



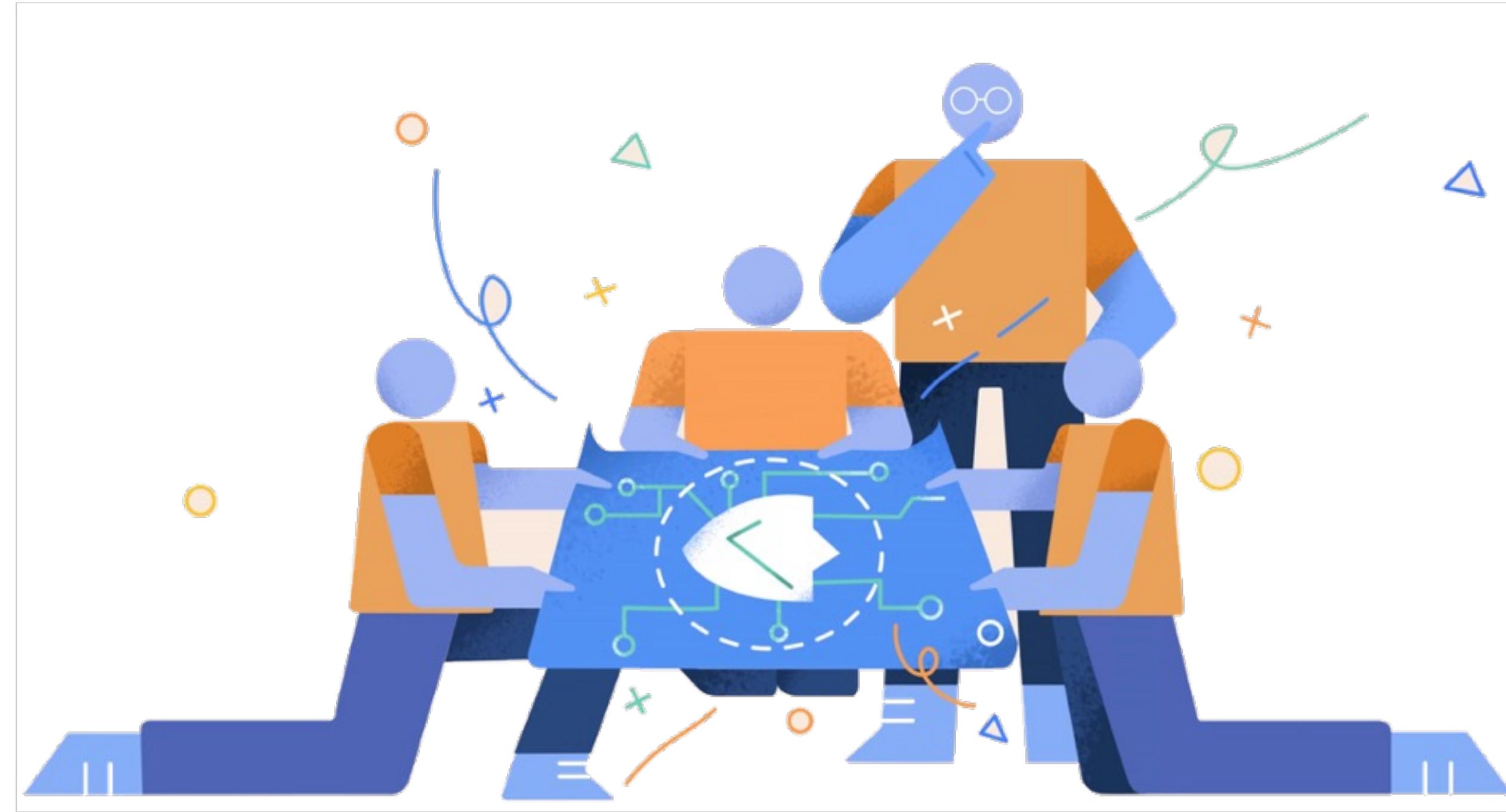
Ask
several people



Leverage
Machine Learning



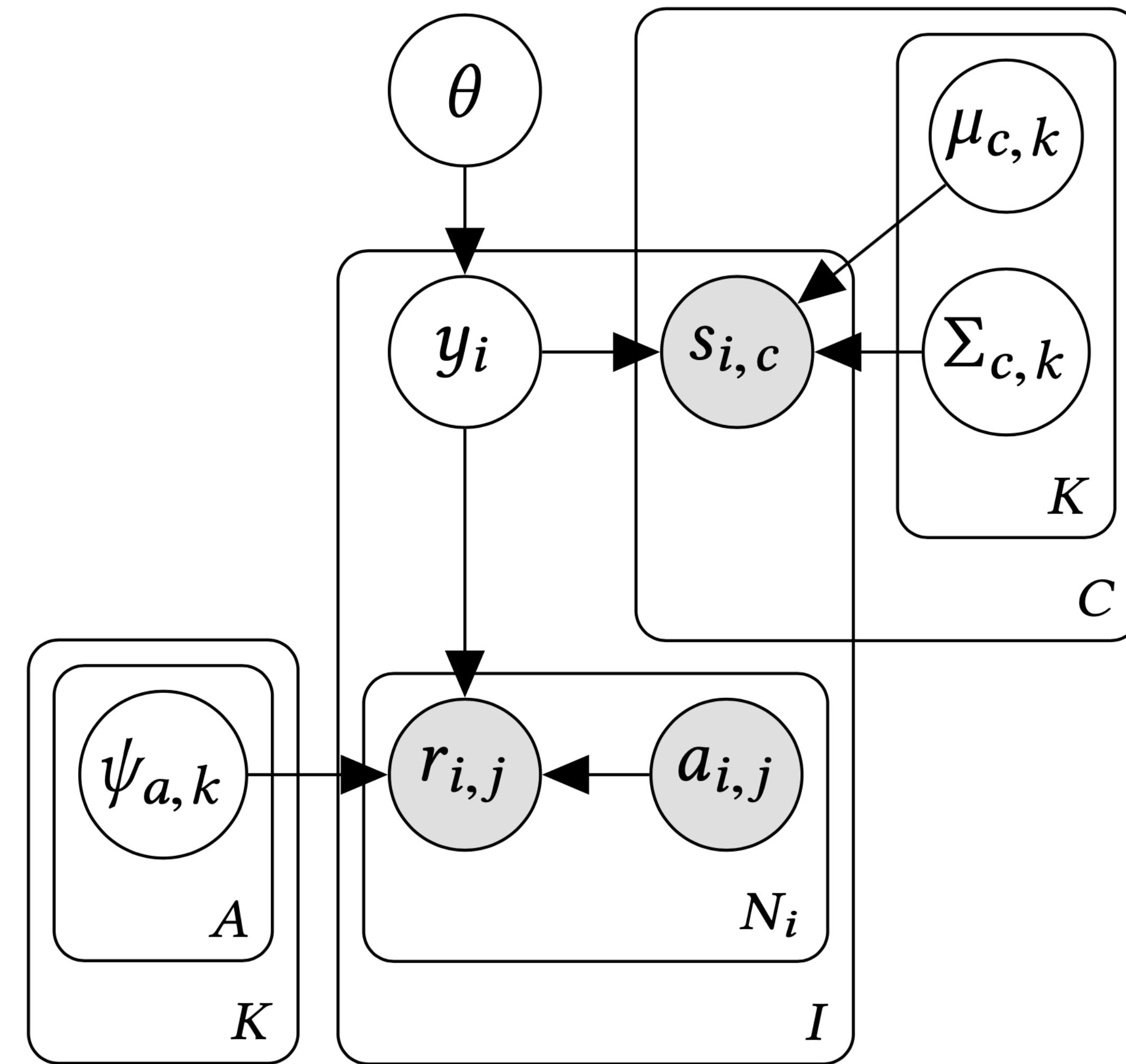
But how do we aggregate multiple labels into a single decision?



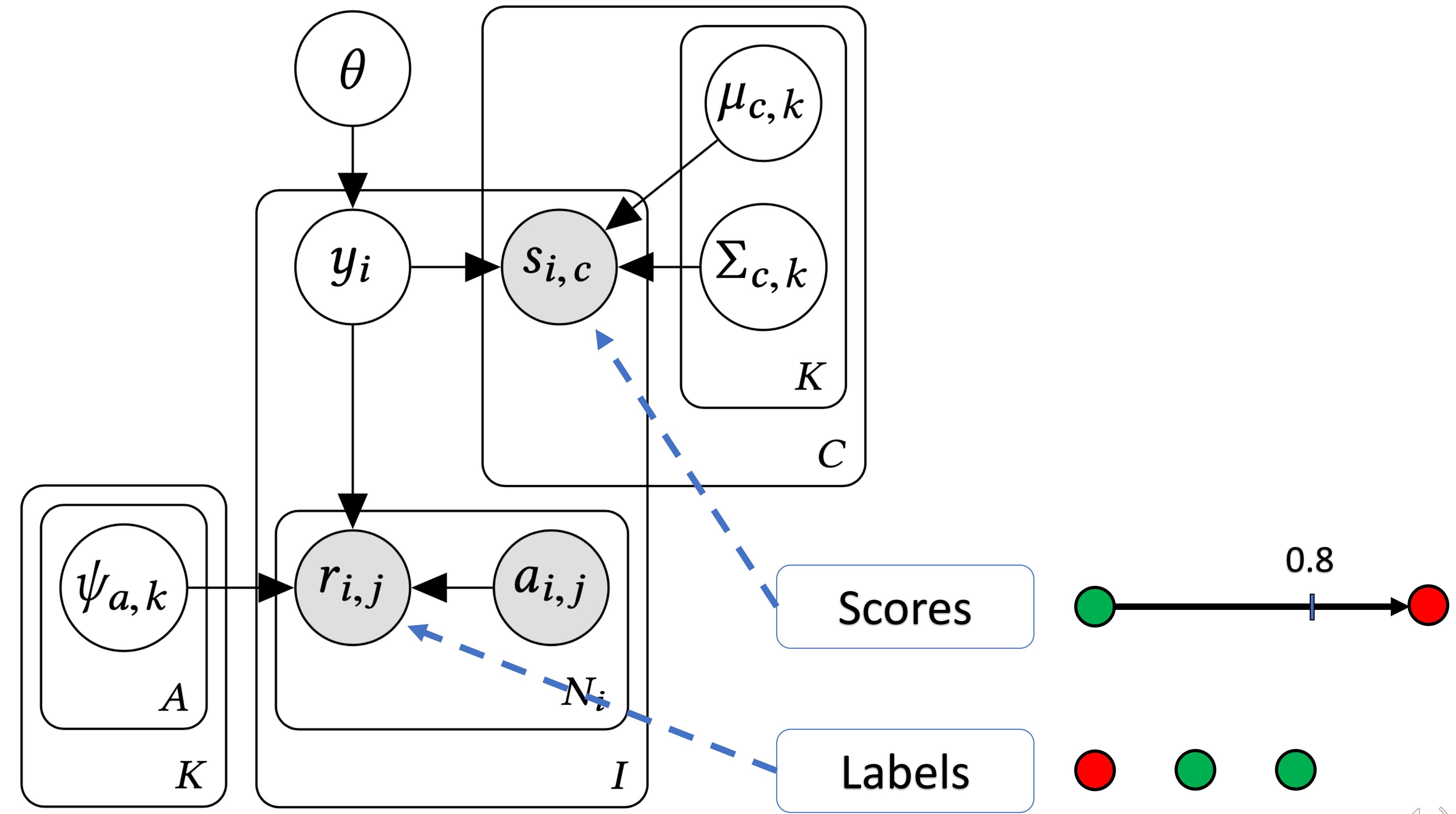
...and can we quantify the certainty of the decision?



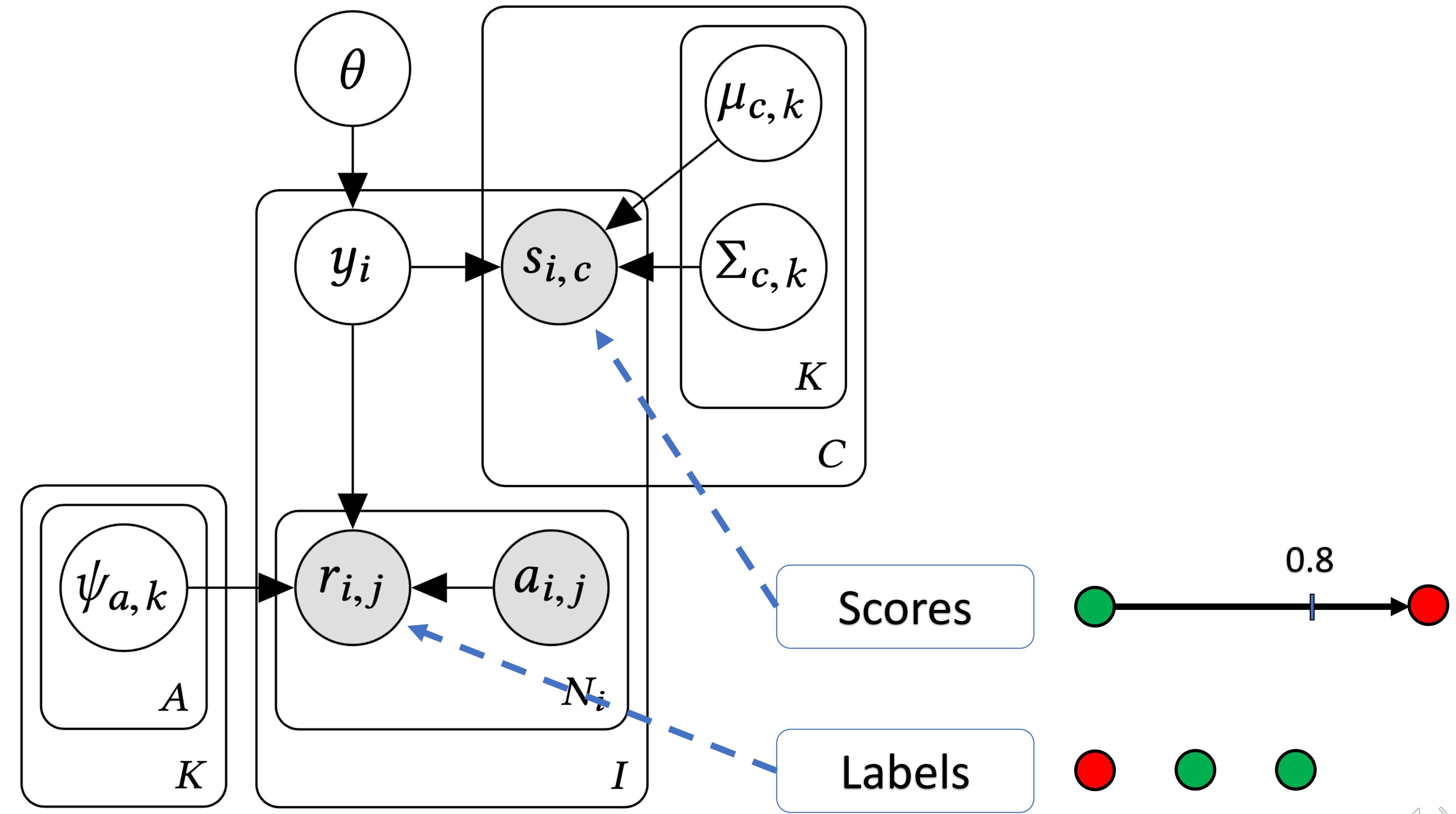
CLARA: Confidence of Labels and Raters



CLARA: Confidence of Labels and Raters



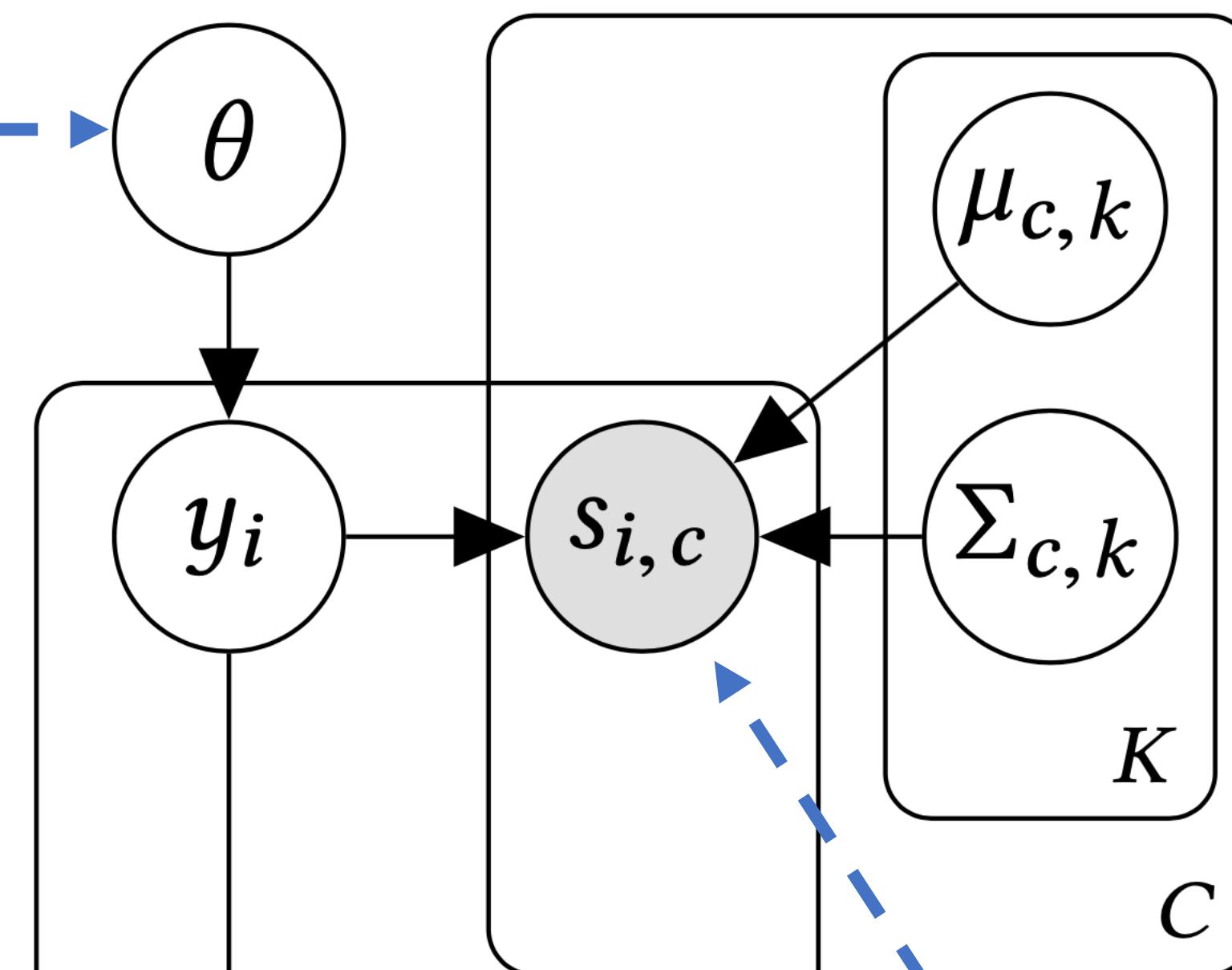
CLARA: Confidence of Labels and Raters



CLARA: Confidence of Labels and Raters

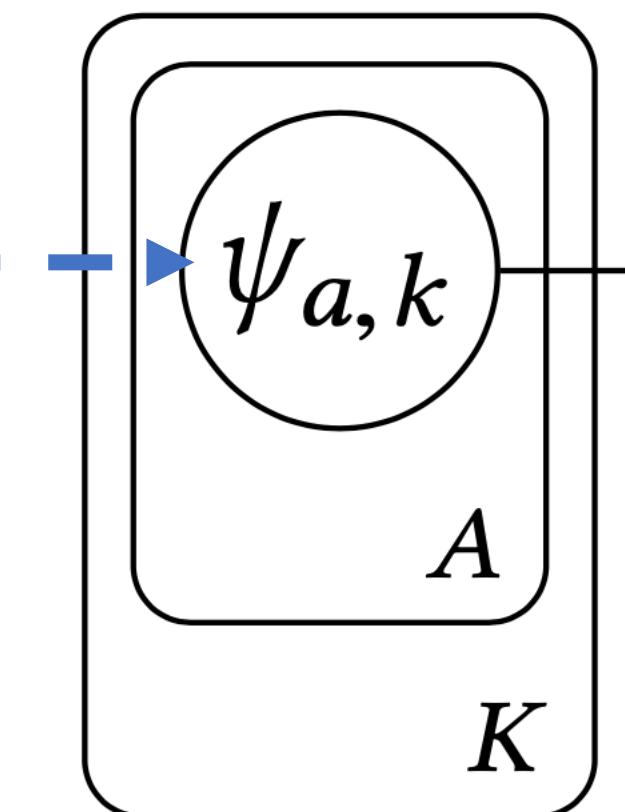
	0.9
	0.1

Overall
Prevalence

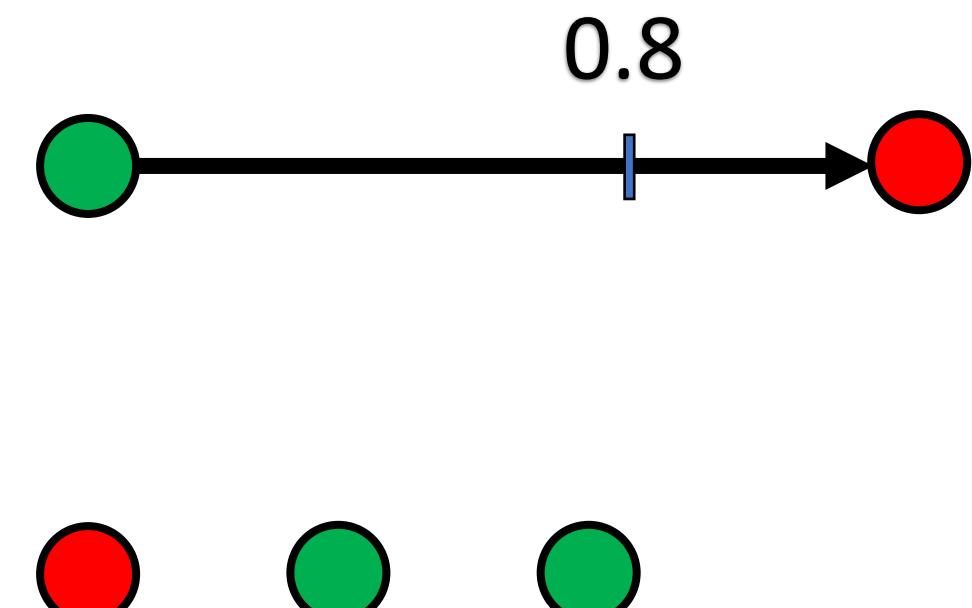


	0.95	0.05
	0.2	0.8

Reviewer
Confusion
Matrix



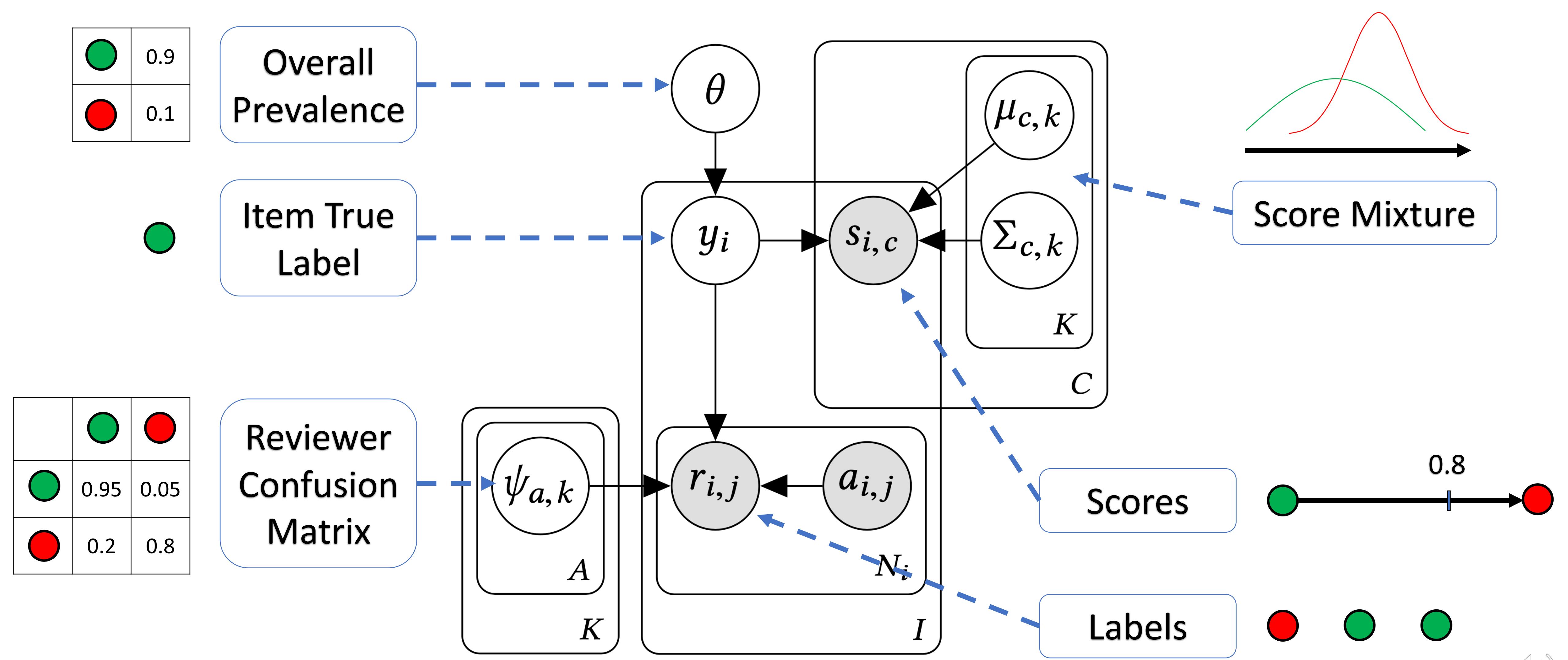
Scores



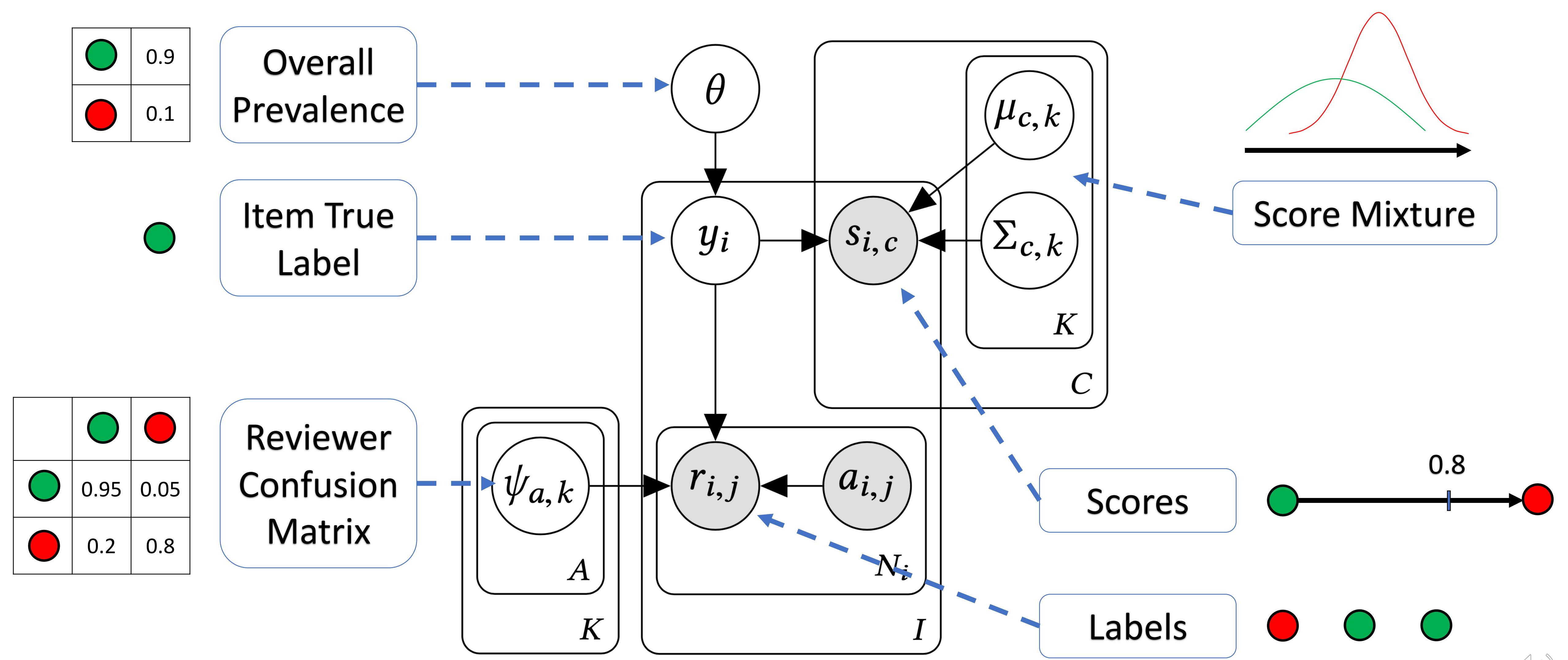
Labels



CLARA: Confidence of Labels and Raters



CLARA: Confidence of Labels and Raters





Applications

1. Prevalence Measurement
2. Reviewer Performance Measurement
3. Labeling Efficiency

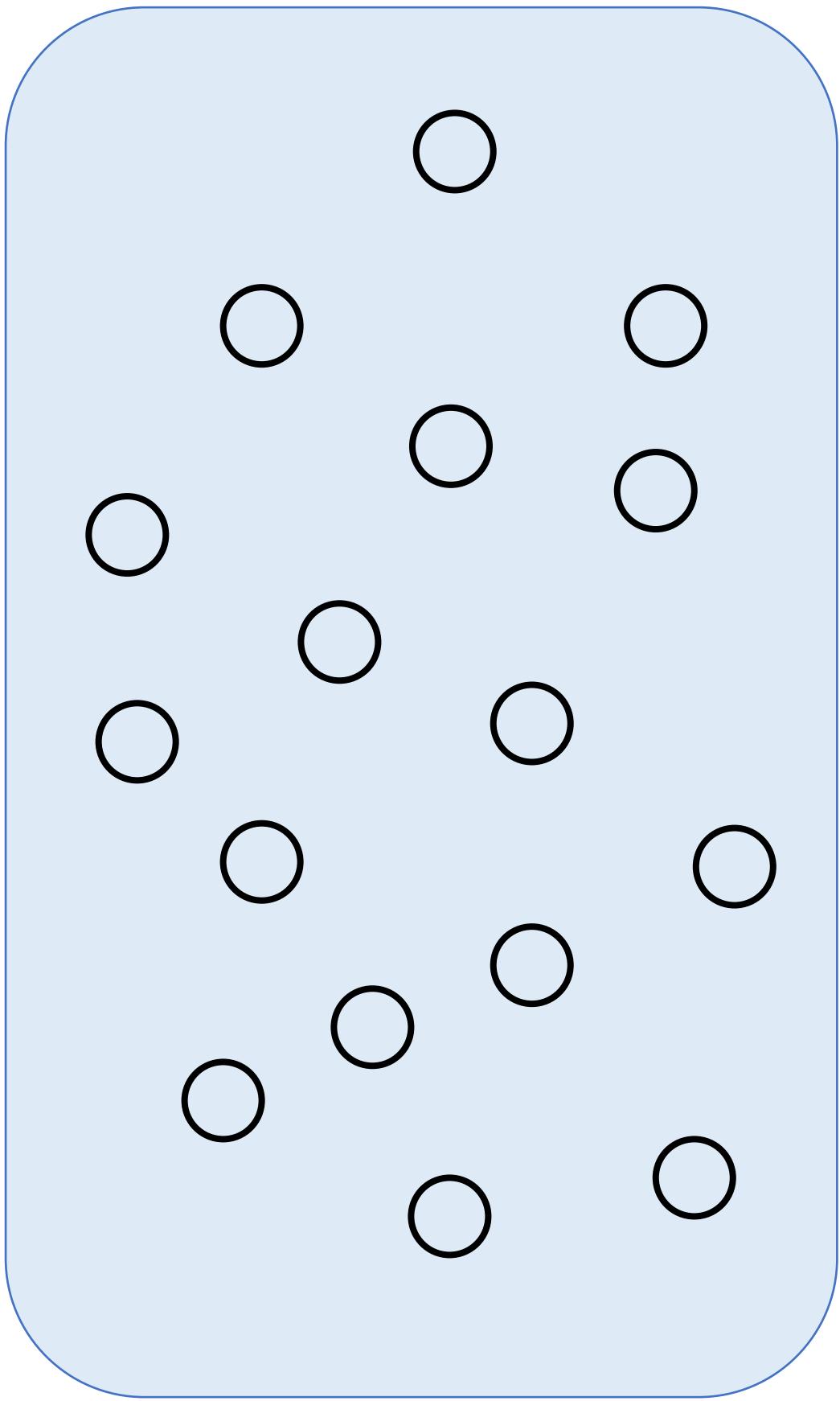


Prevalence Measurement

Measure the percentage of policy-violating content out of all content seen by Facebook users



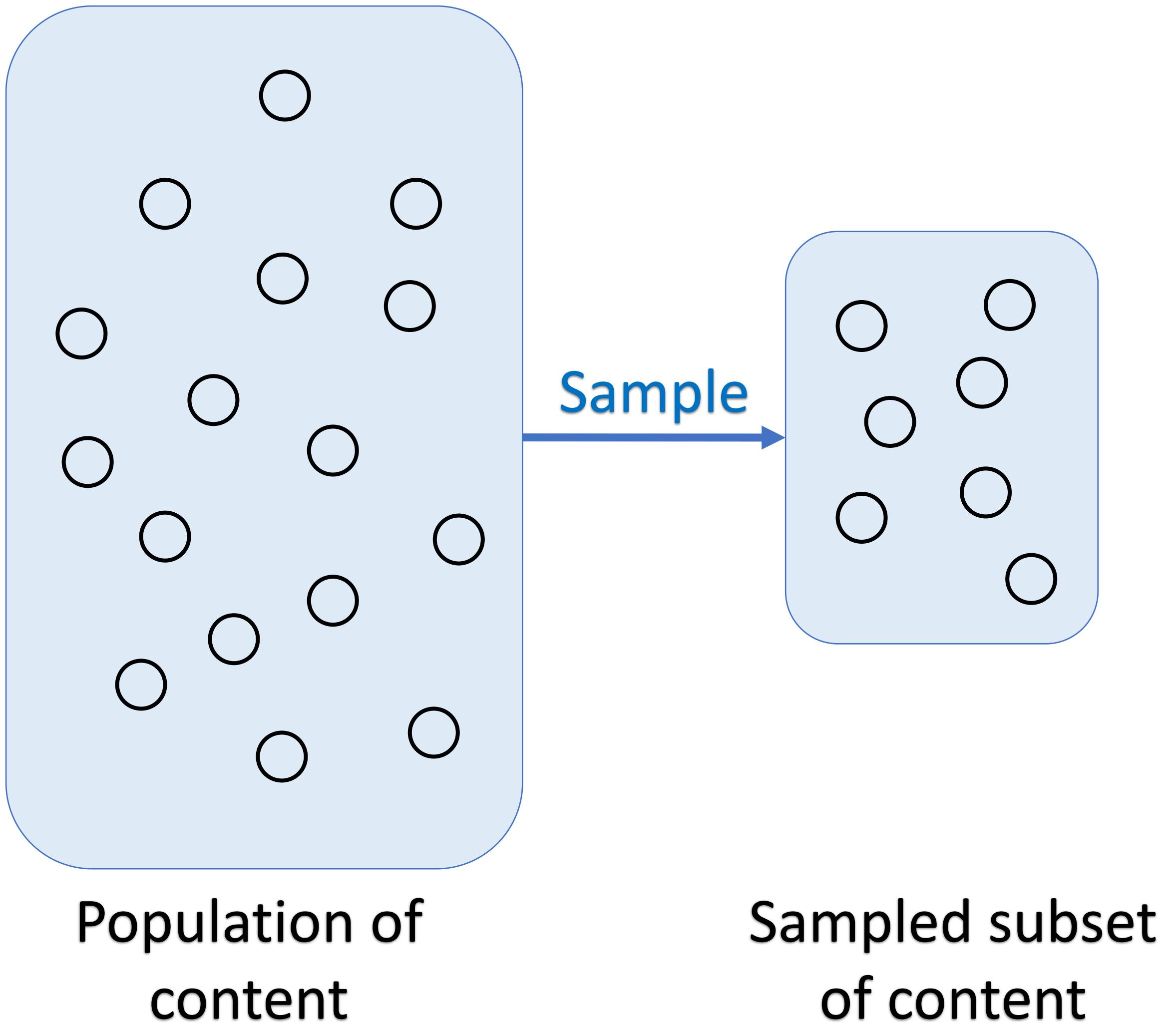
Prevalence Measurement



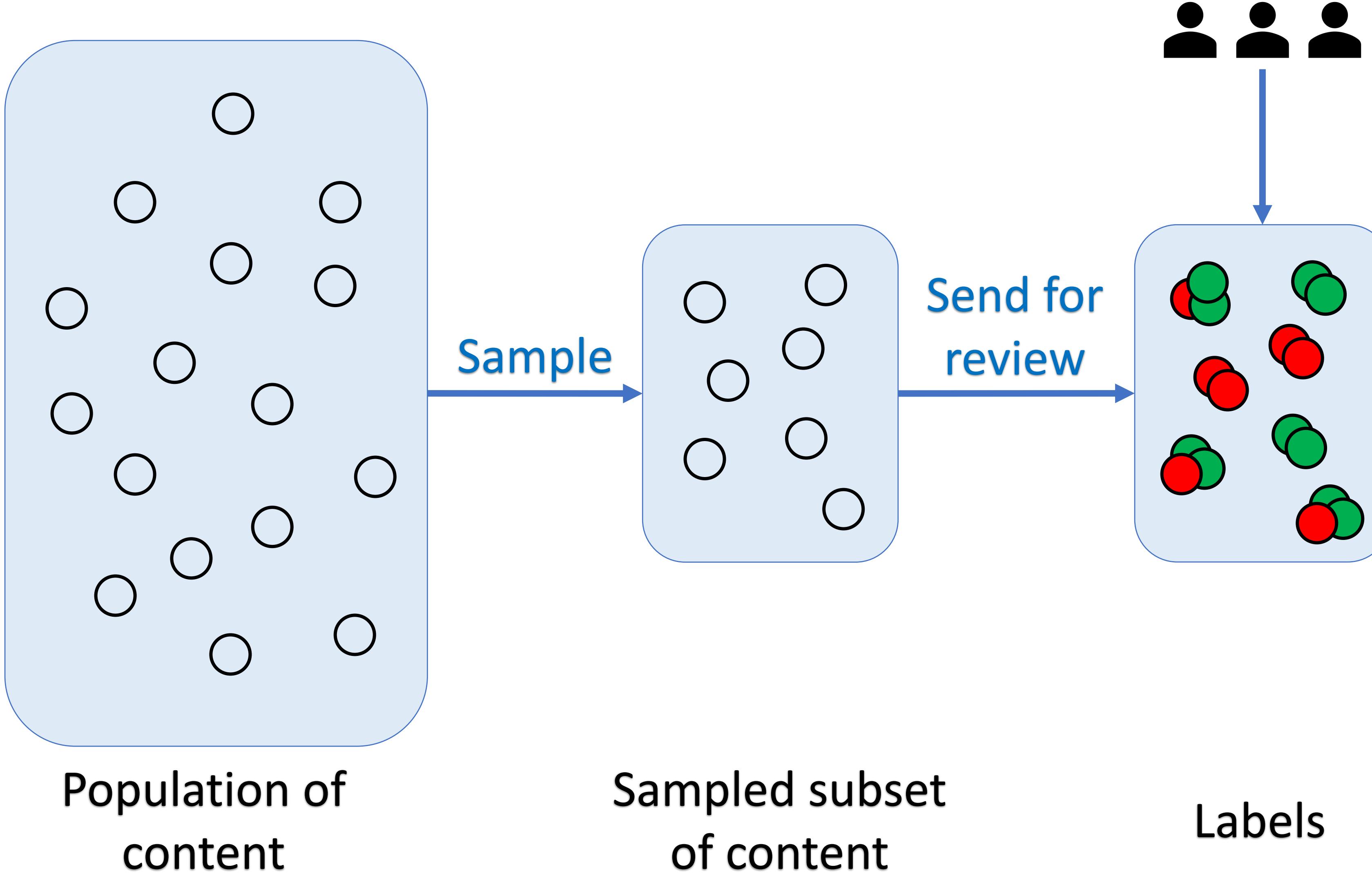
Population of
content



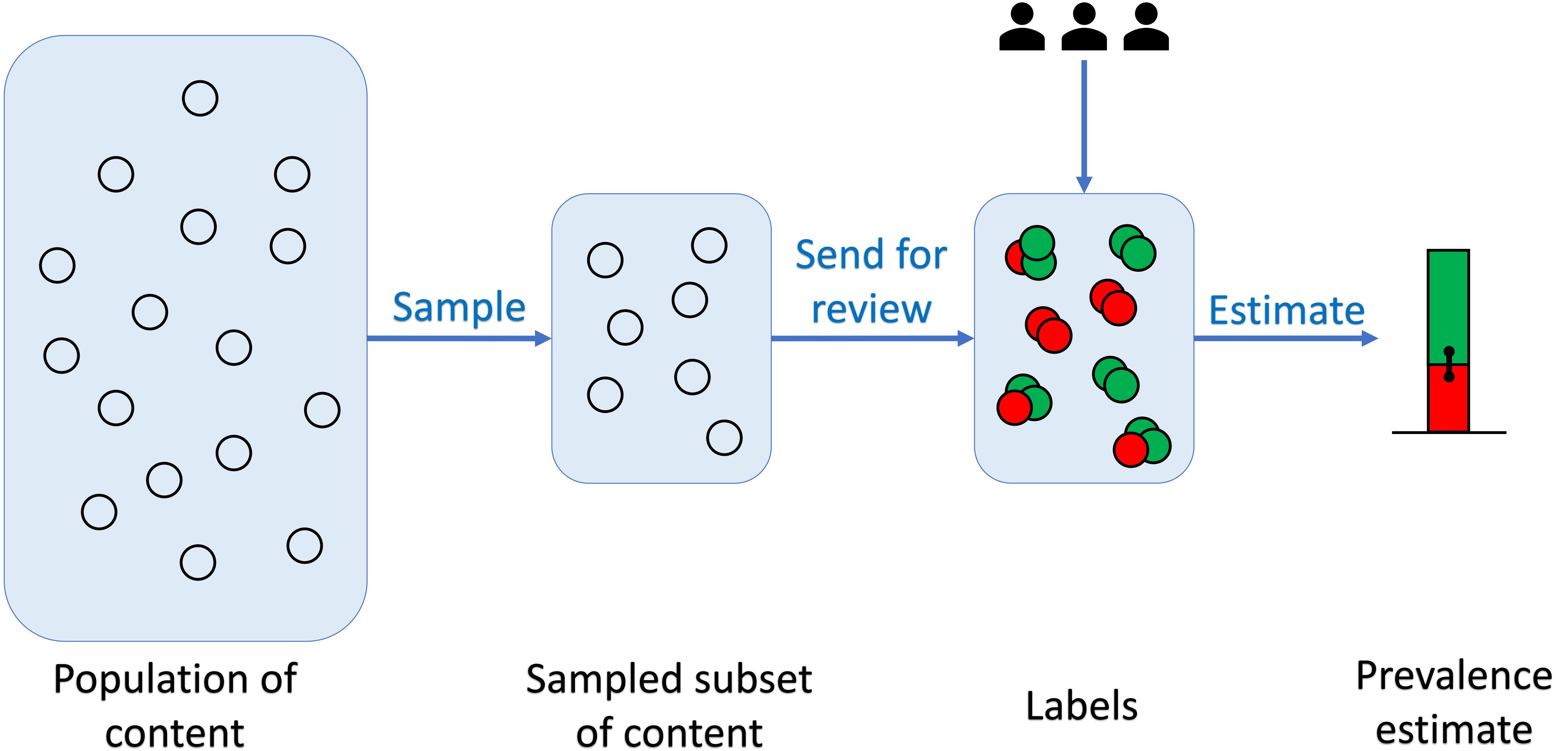
Prevalence Measurement



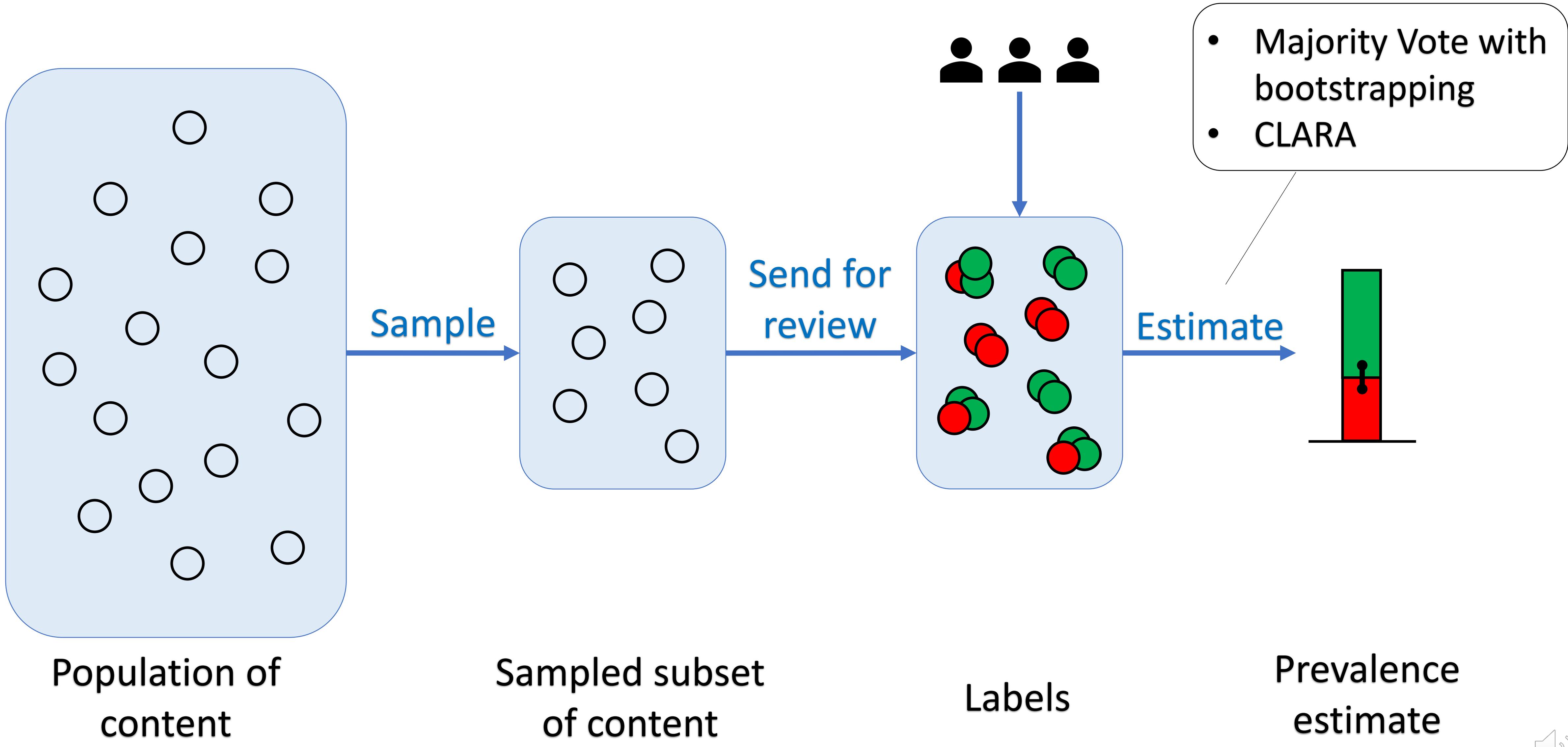
Prevalence Measurement



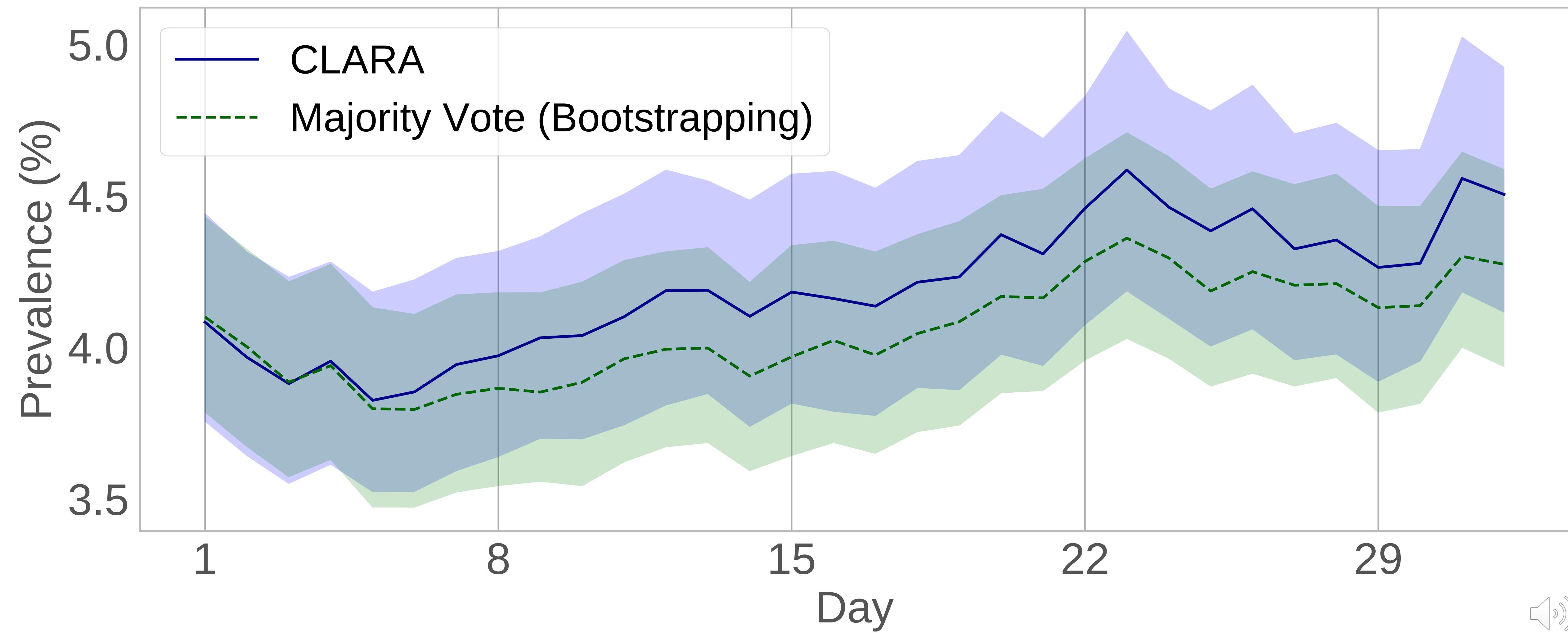
Prevalence Measurement



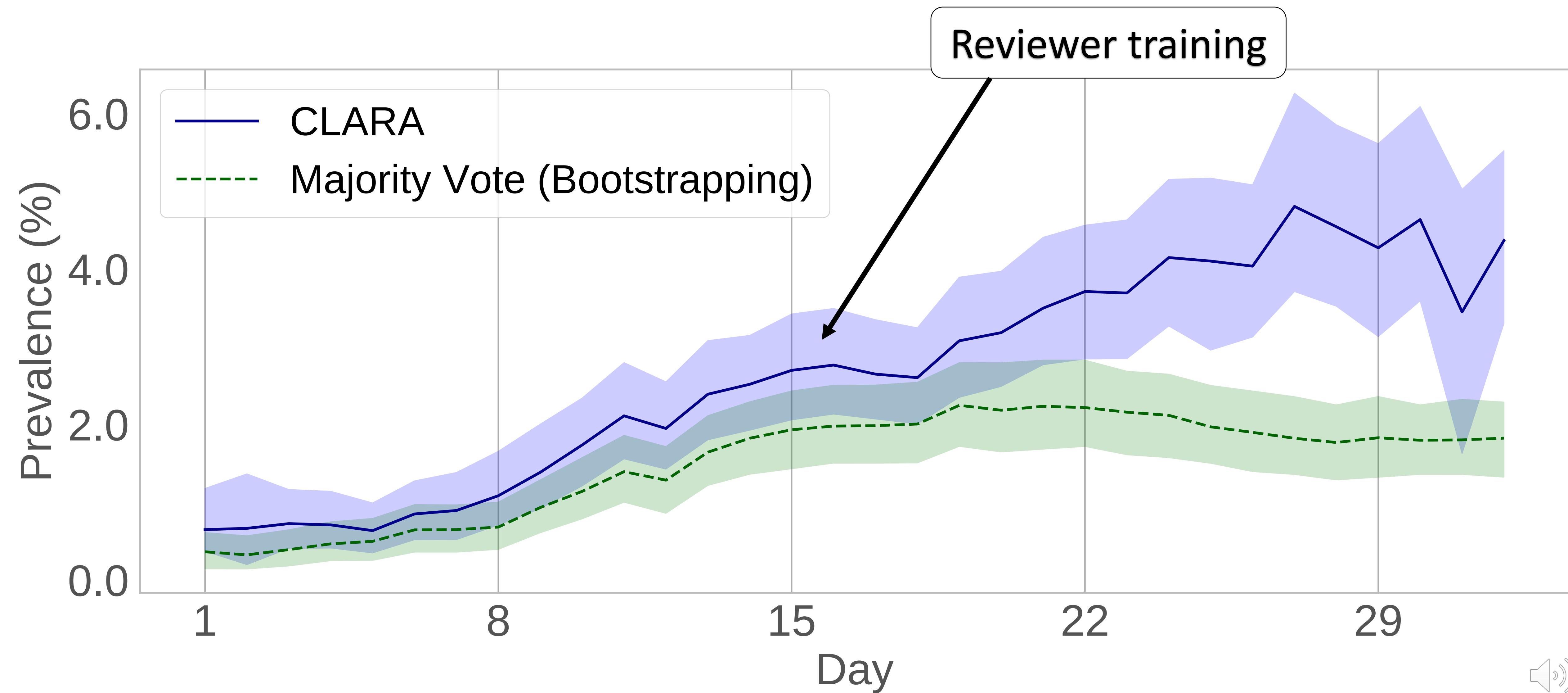
Prevalence Measurement



Prevalence Estimates (Violation Type A)



Prevalence Estimates (Violation Type B)



Reviewer Performance Measurement

Measure the performance of reviewers in labeling violating content



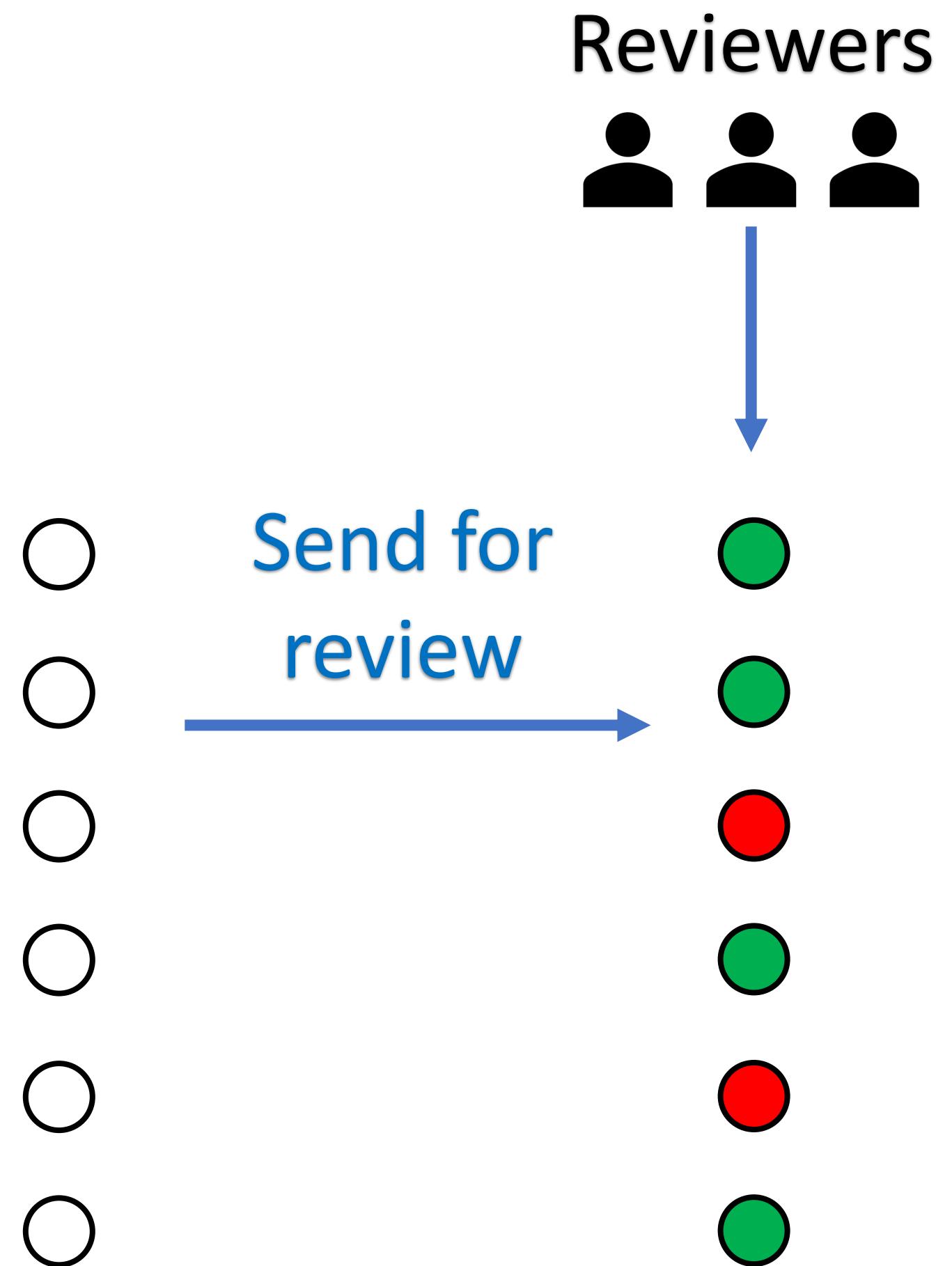
Measuring Reviewer Performance

-
-
-
-
-
-

Content for
review



Measuring Reviewer Performance

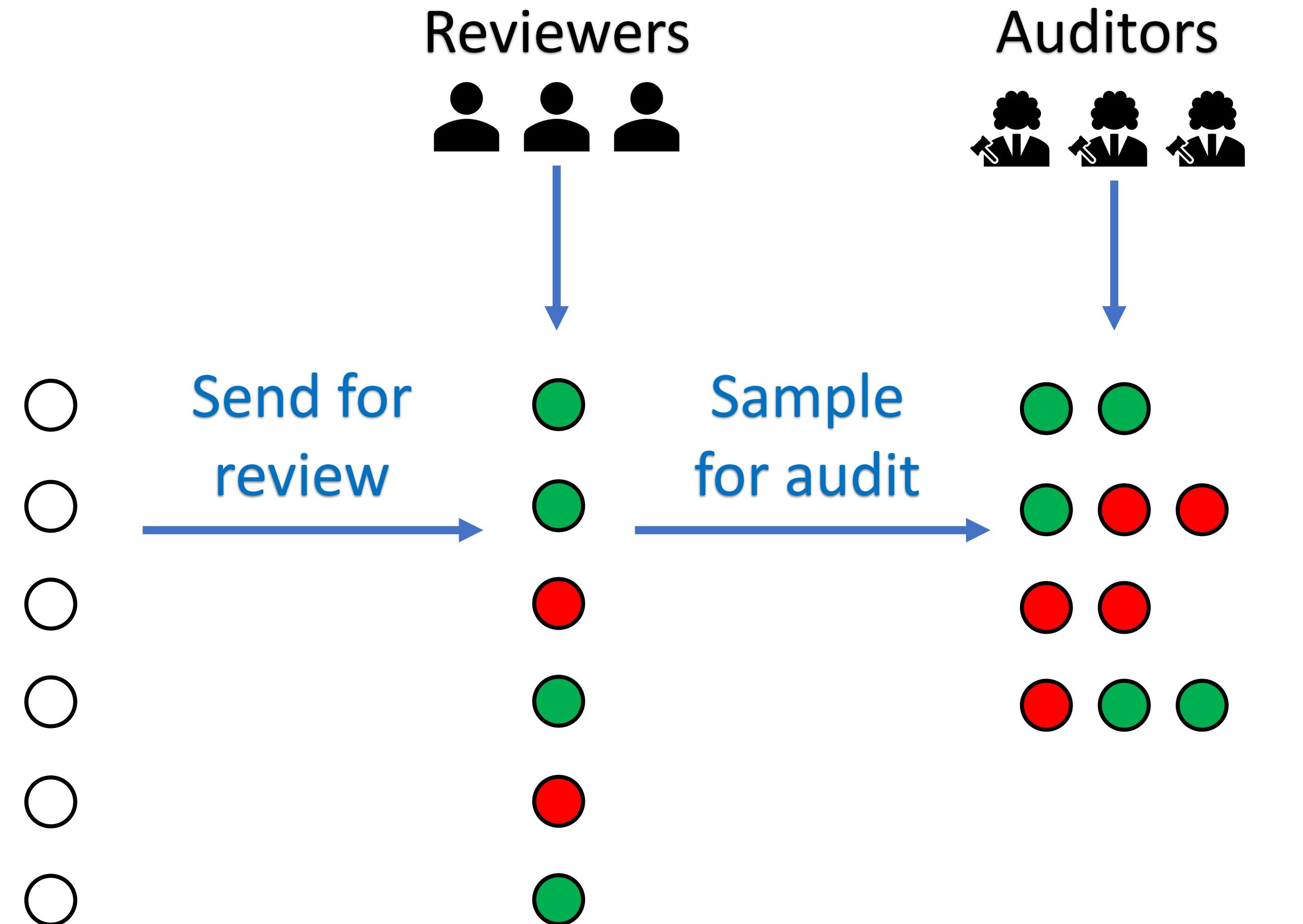


Content for
review

Single labels
by reviewers



Measuring Reviewer Performance



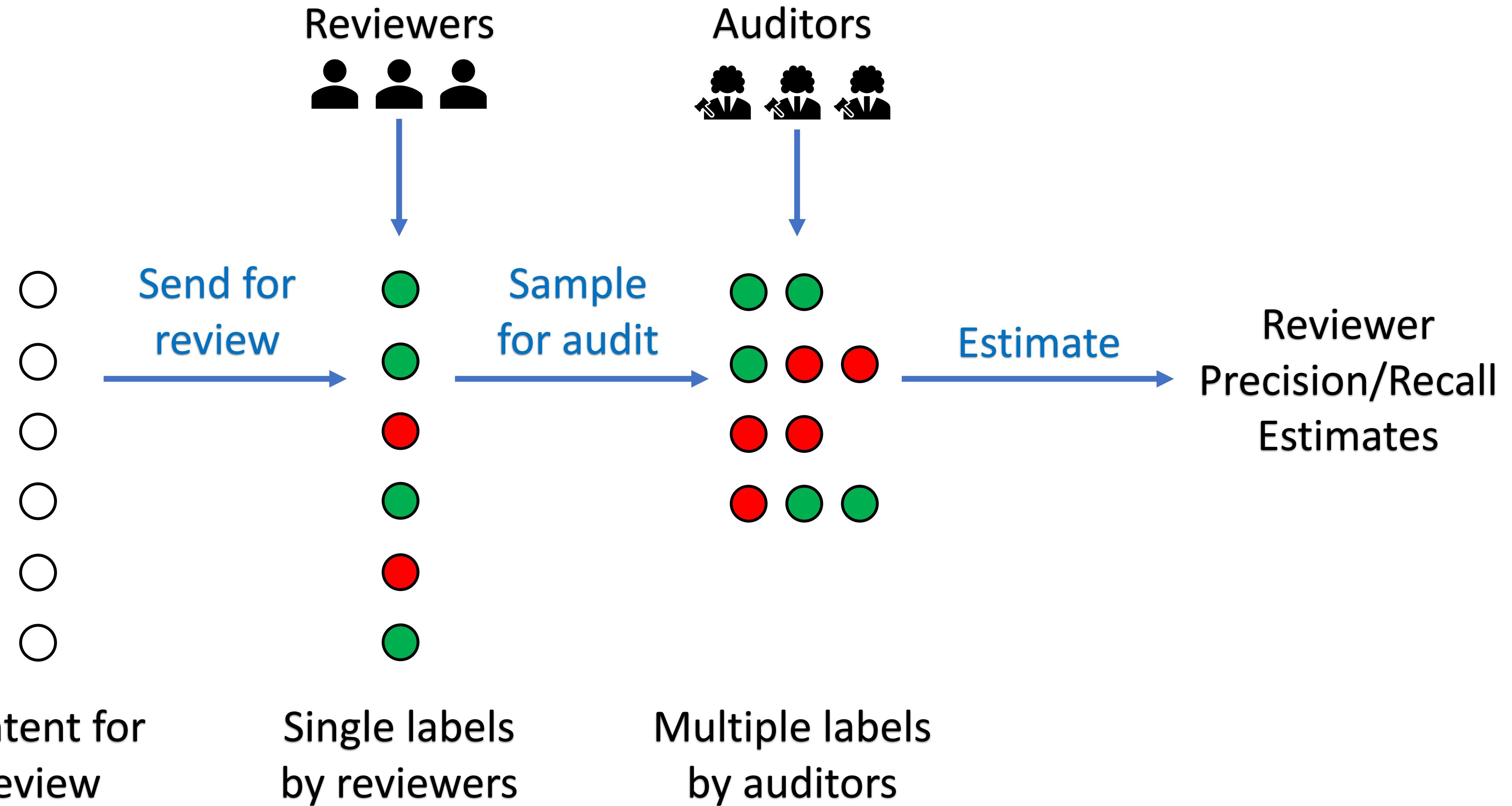
Content for
review

Single labels
by reviewers

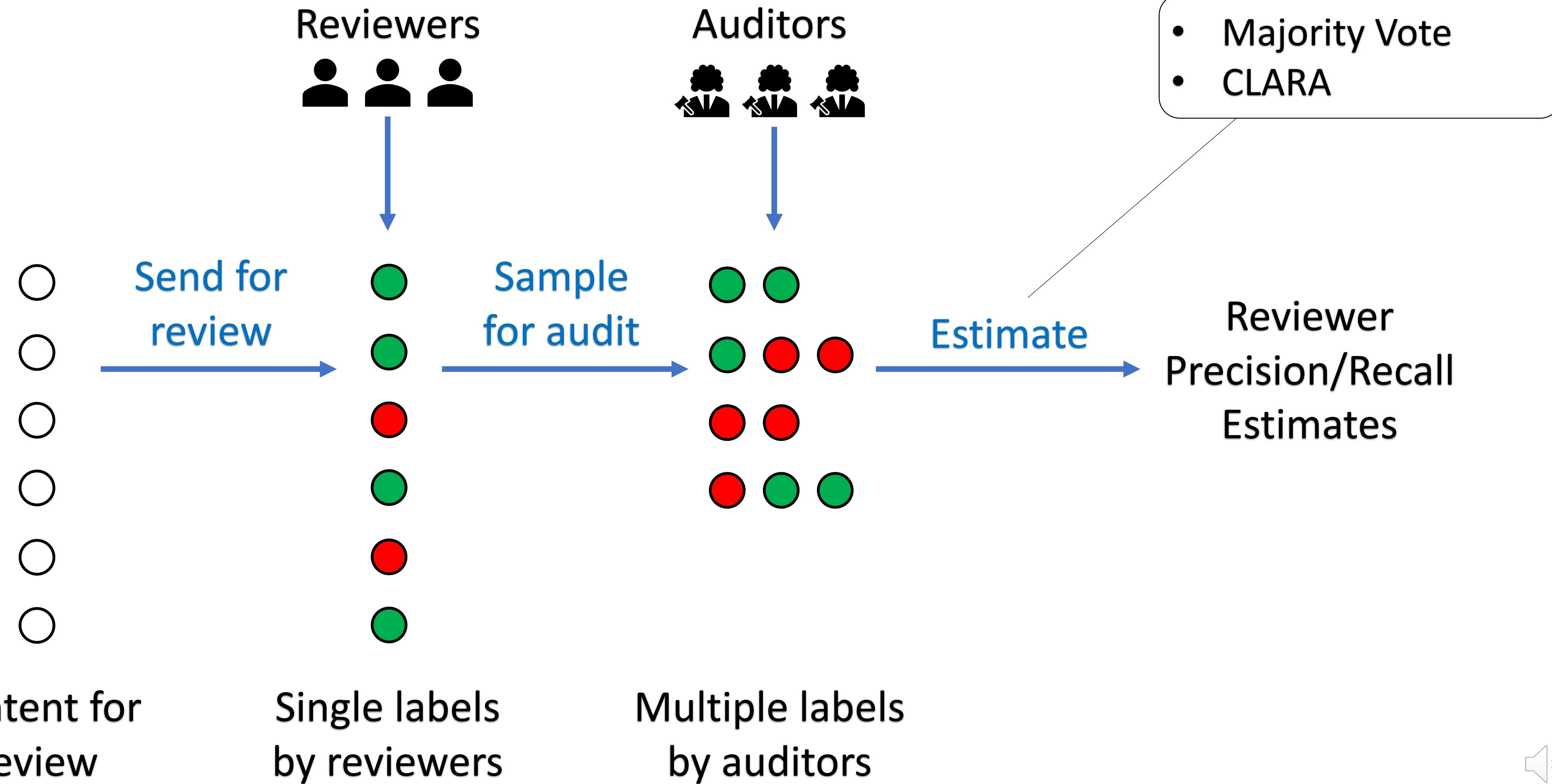
Multiple labels
by auditors



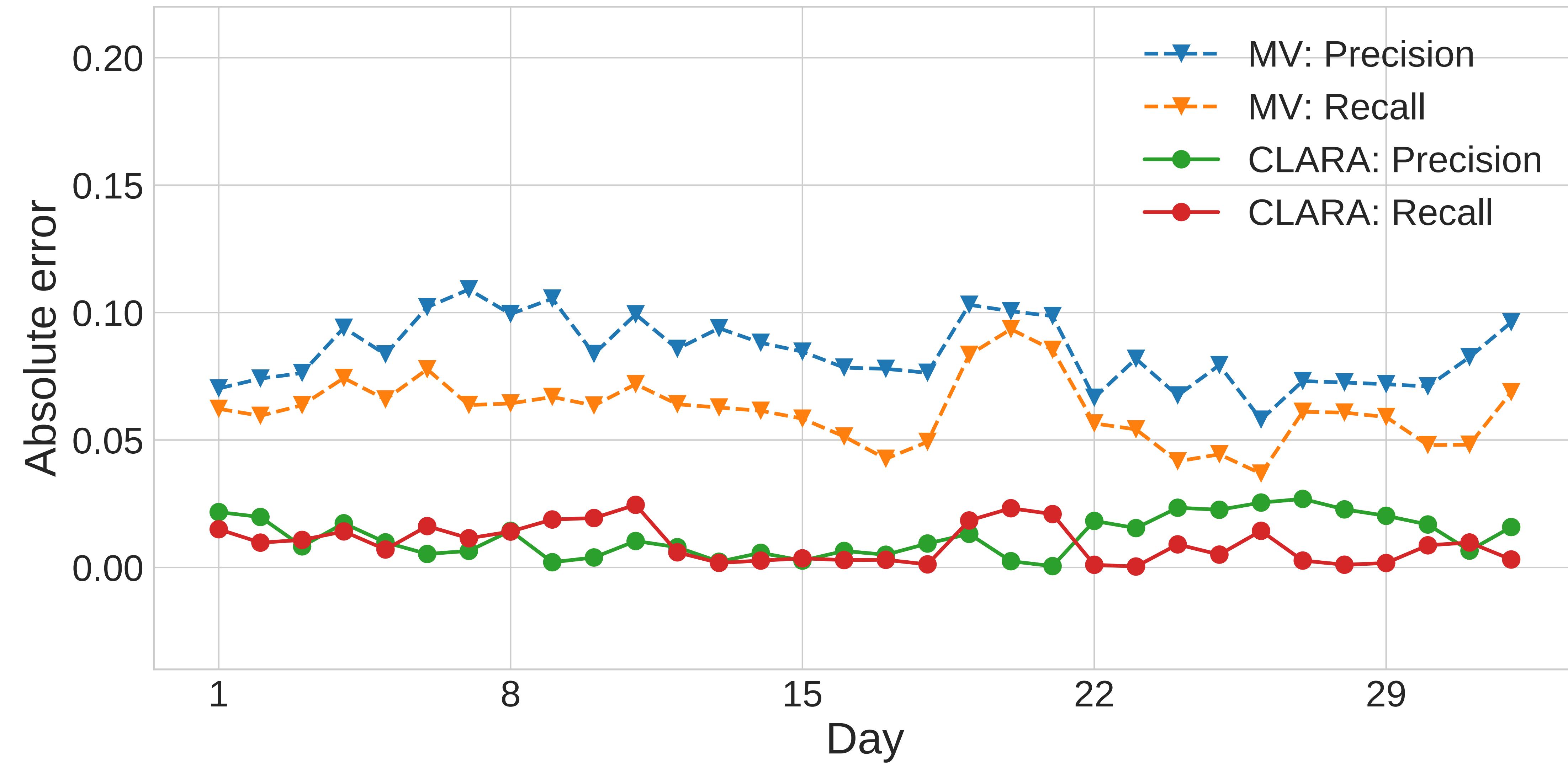
Measuring Reviewer Performance



Measuring Reviewer Performance



Reviewer Precision/Recall Estimates



Labeling Efficiency

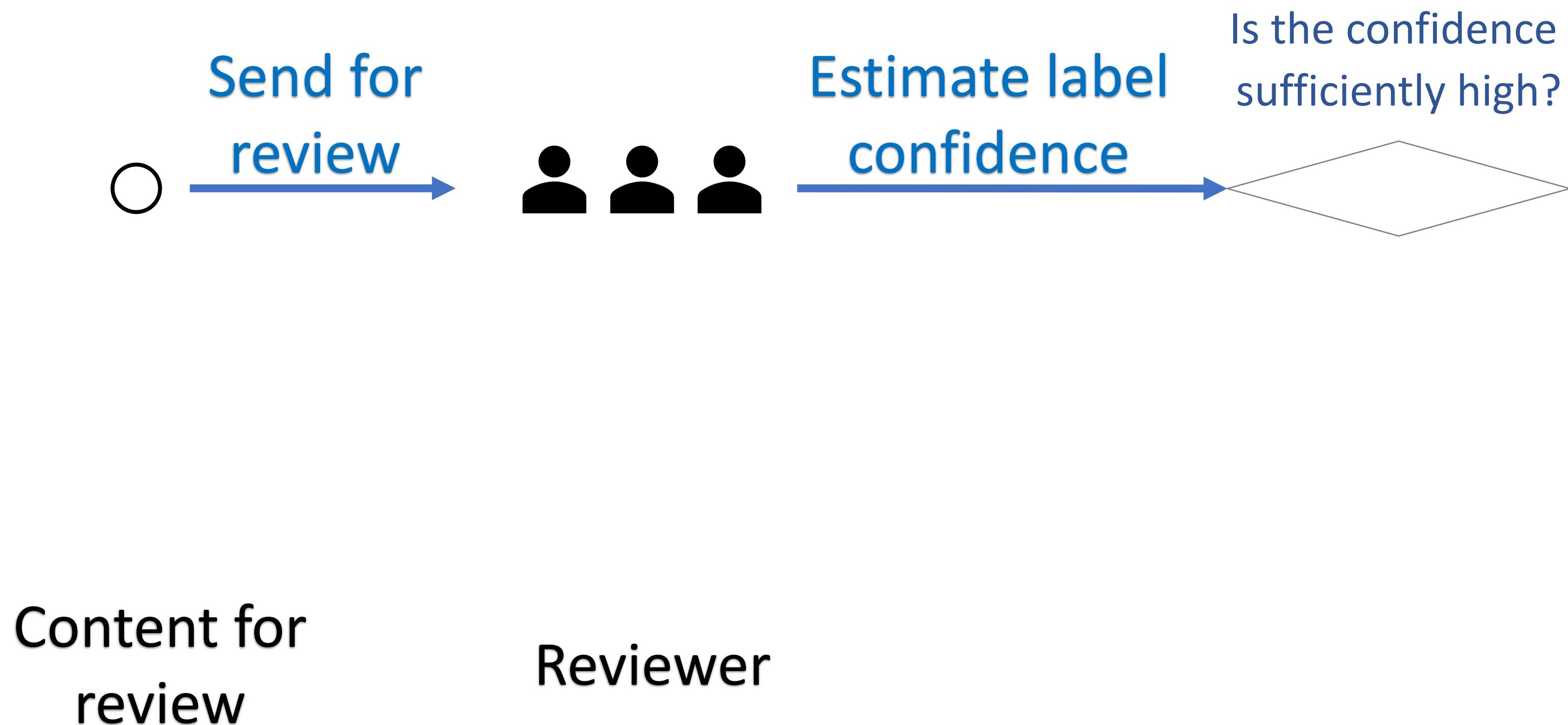
Improve labeling efficiency by only sending content for additional review if the confidence is low



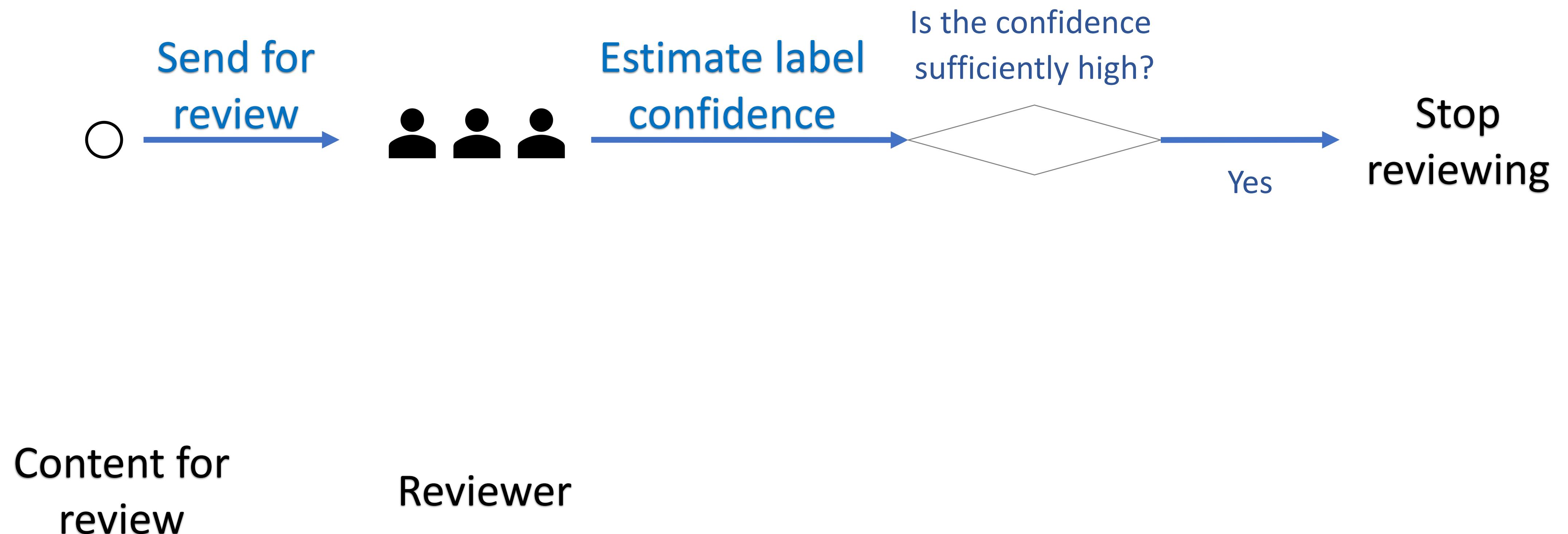
Improving Labeling Efficiency



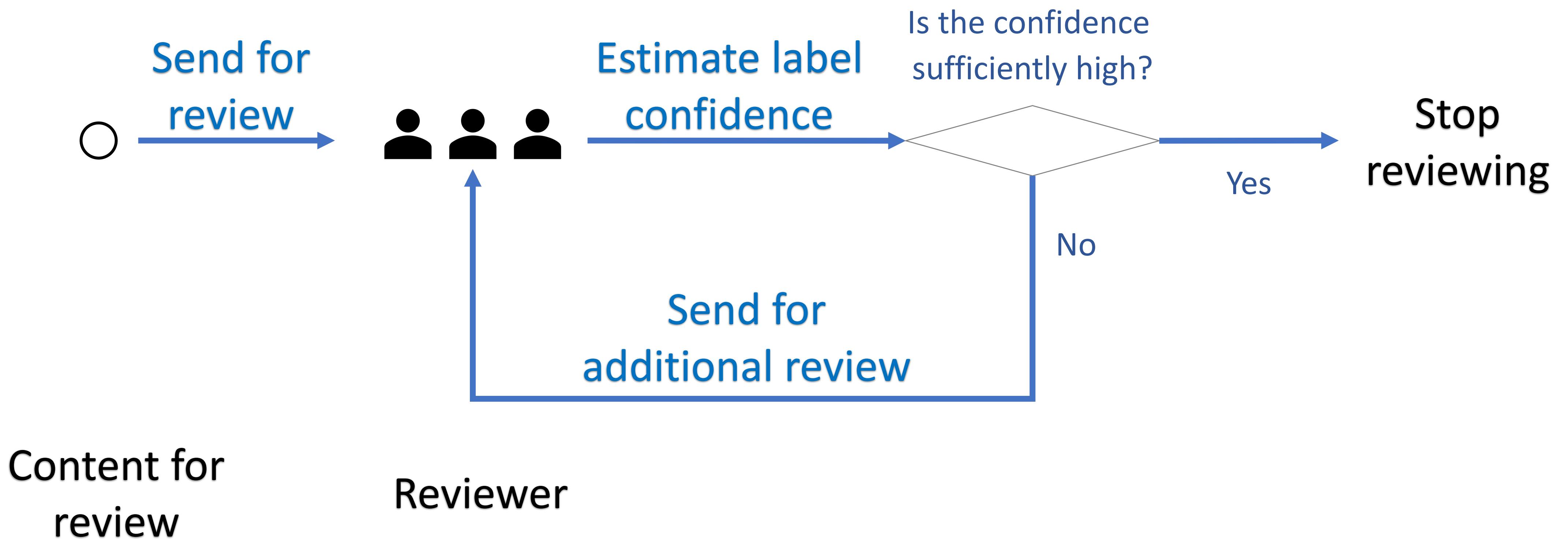
Improving Labeling Efficiency



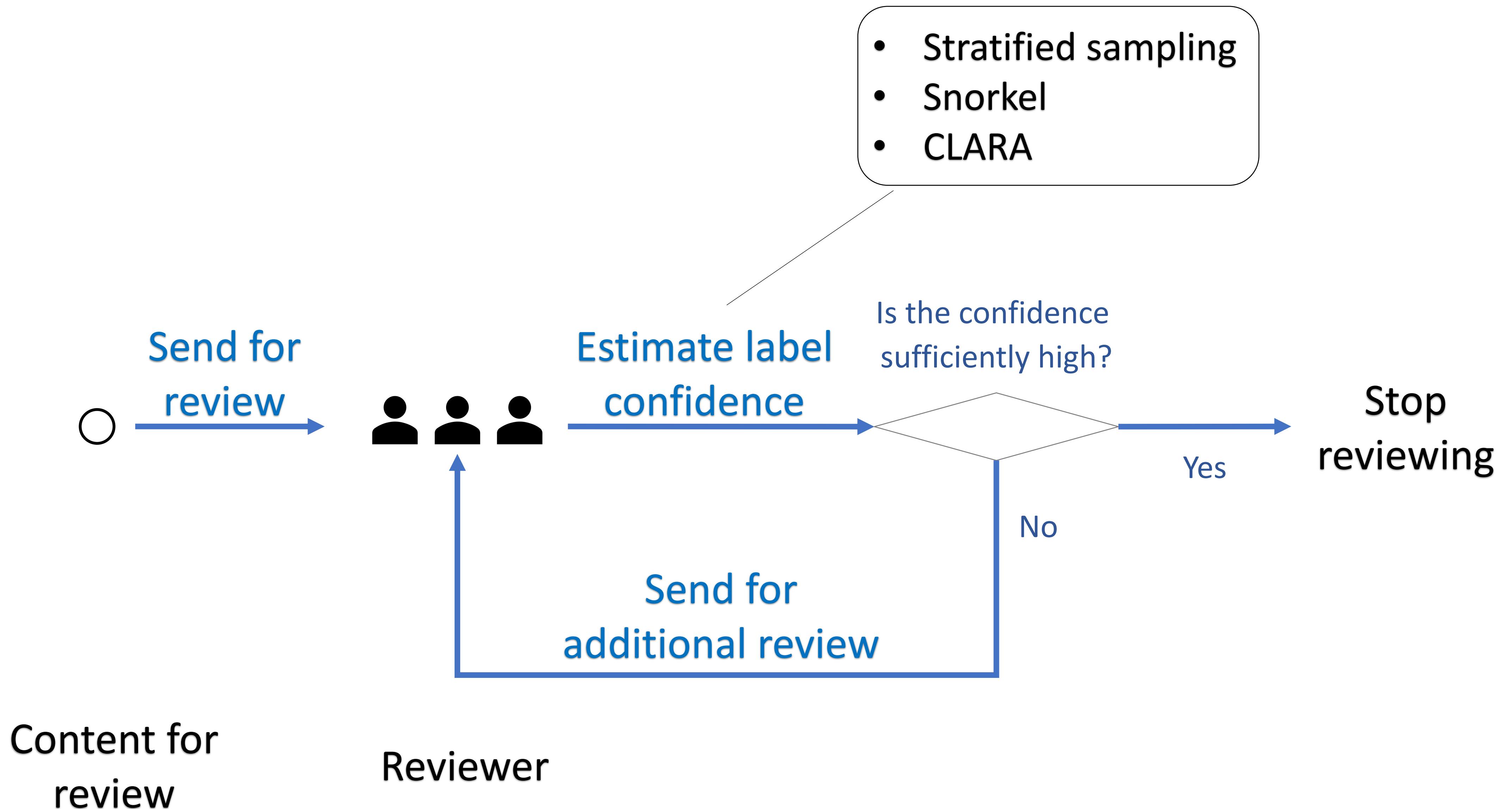
Improving Labeling Efficiency



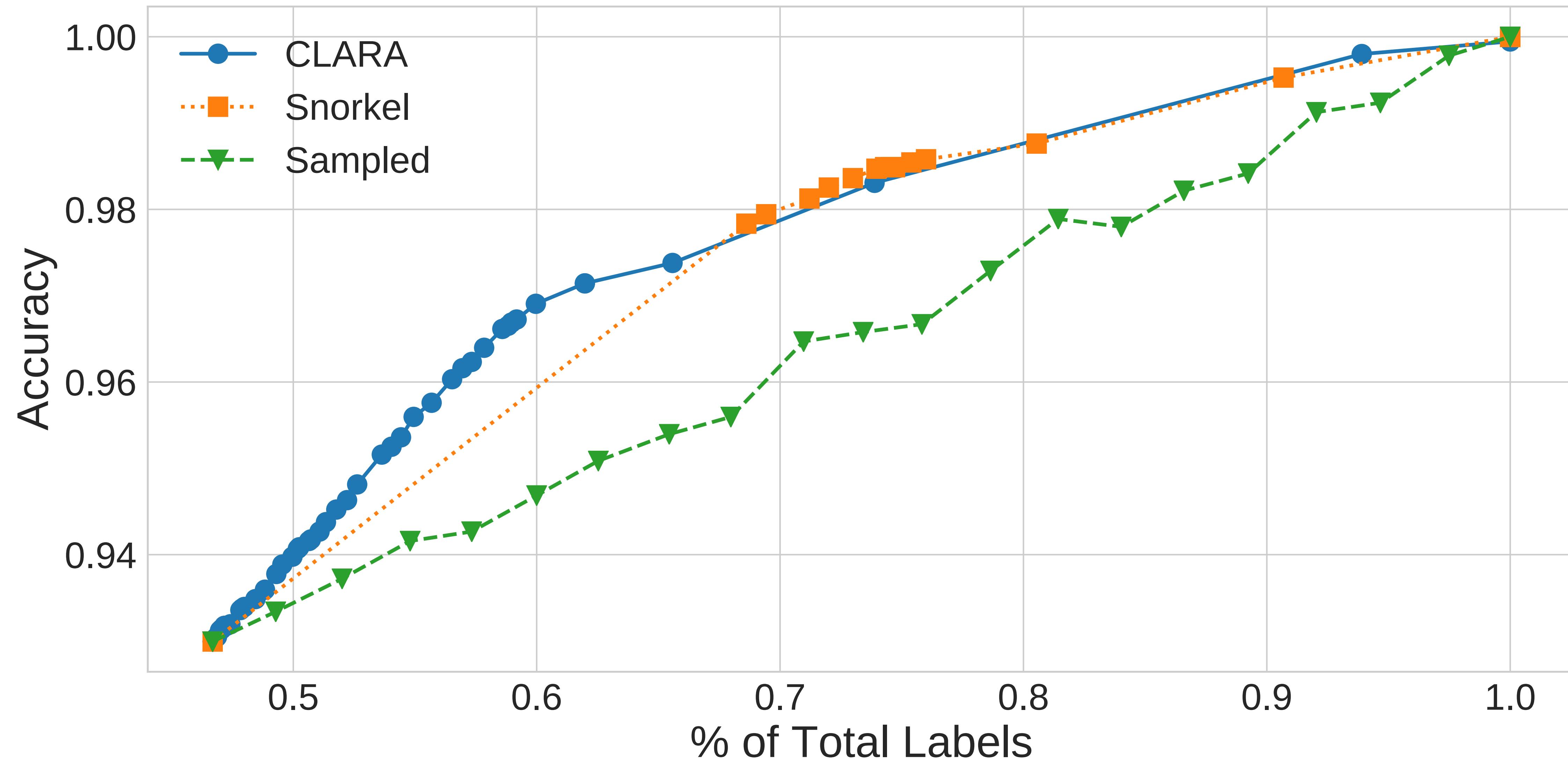
Improving Labeling Efficiency



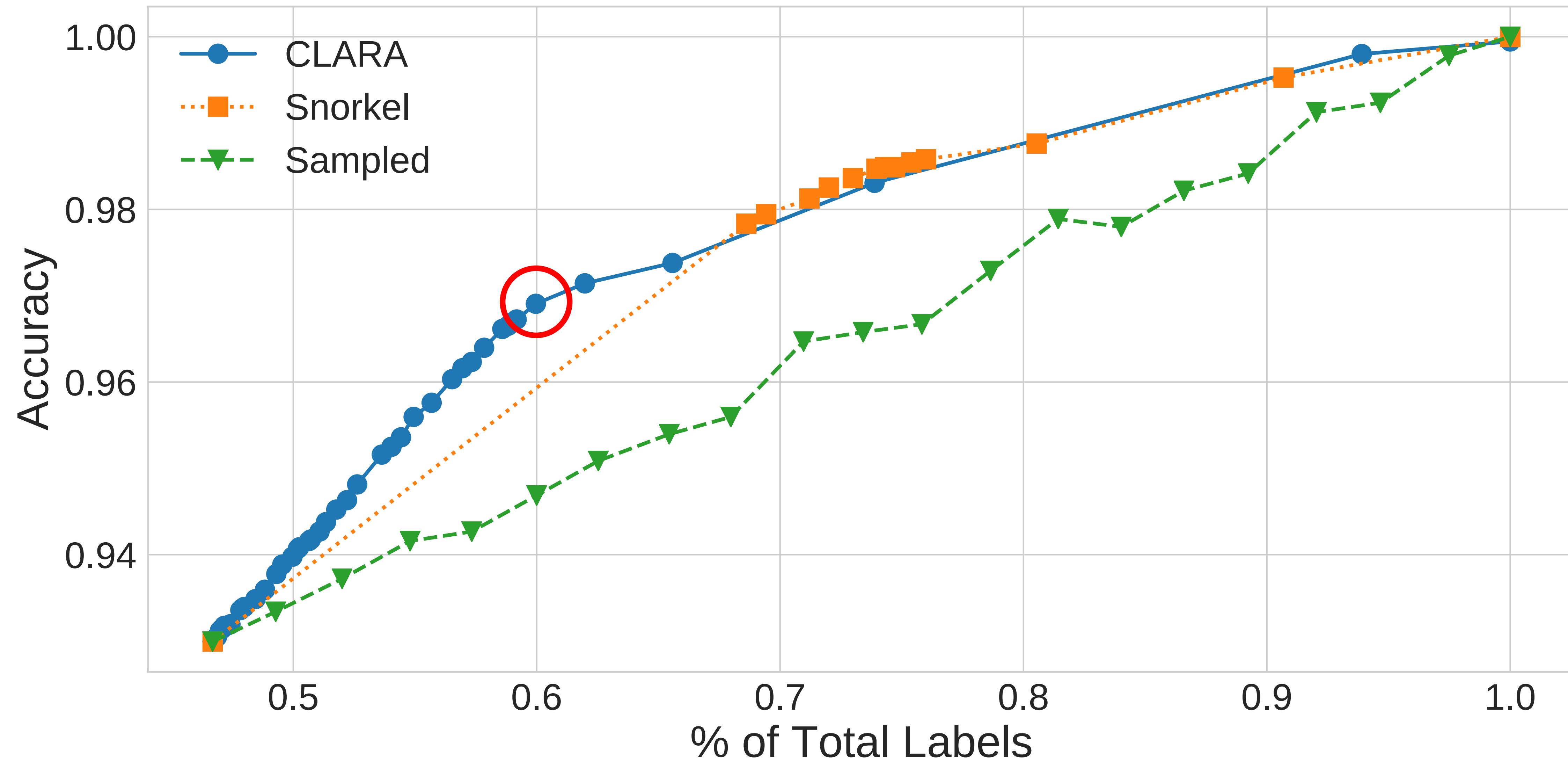
Improving Labeling Efficiency



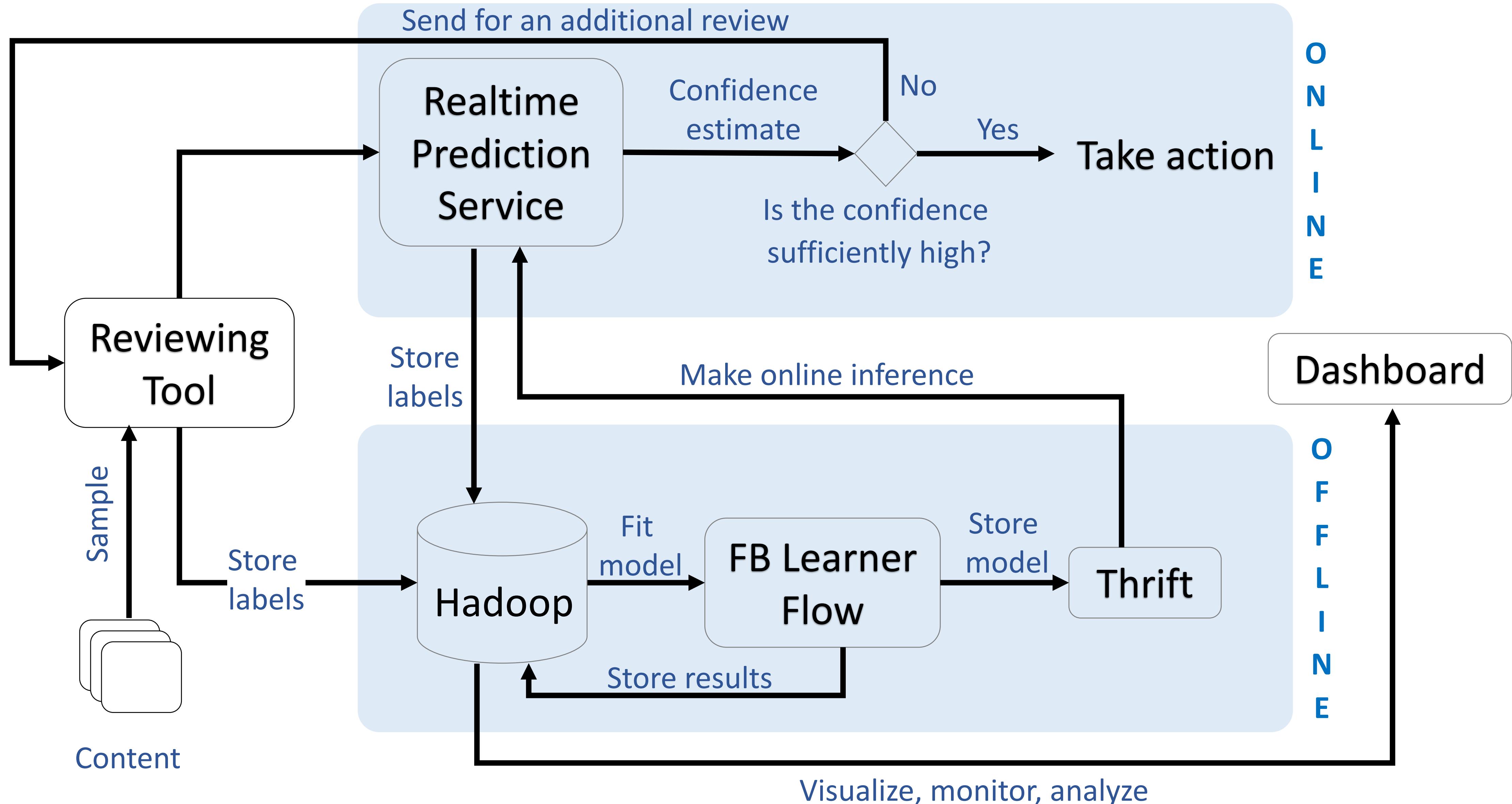
Cost/Accuracy Tradeoff Curve



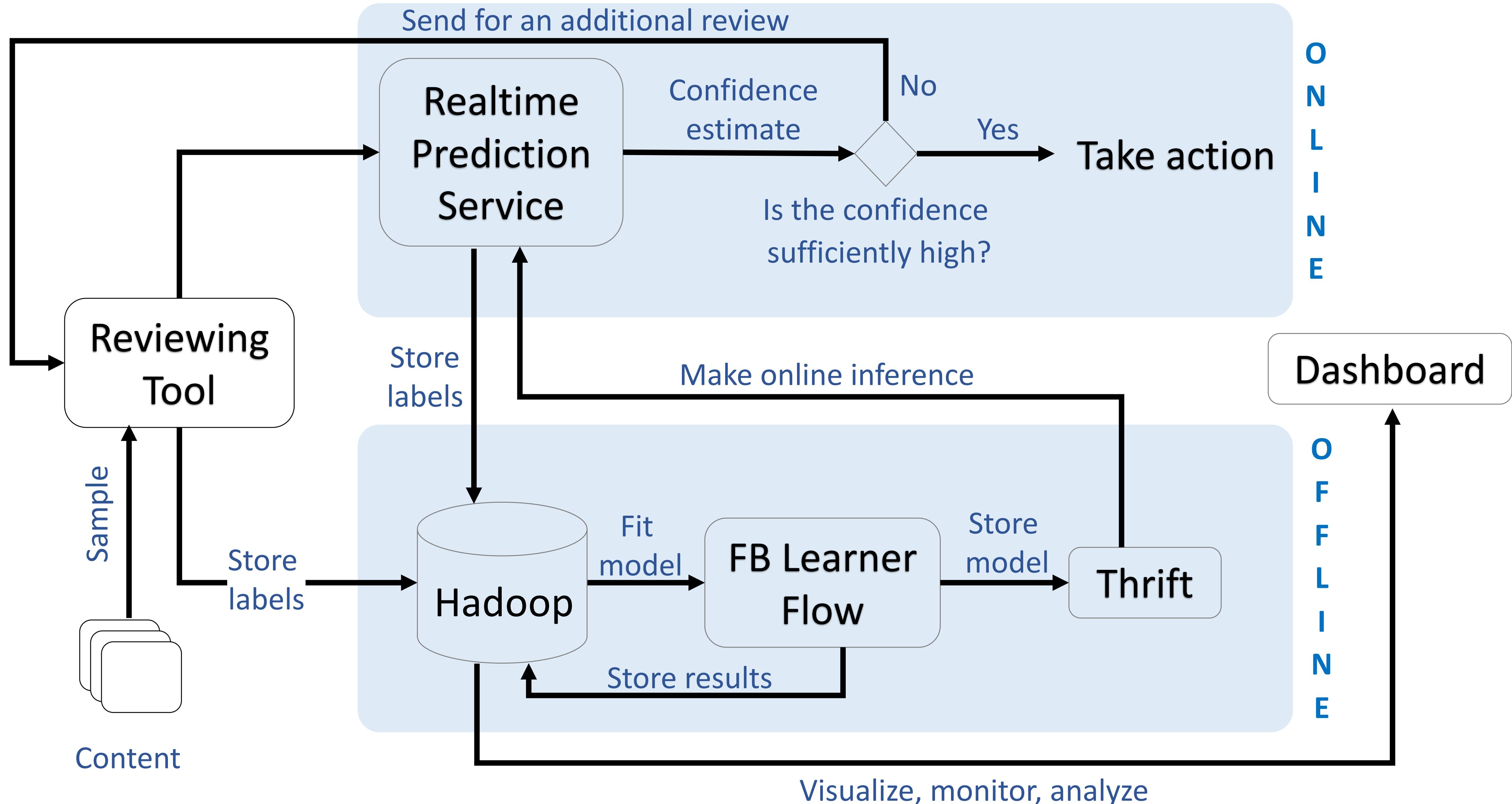
Cost/Accuracy Tradeoff Curve



System Overview



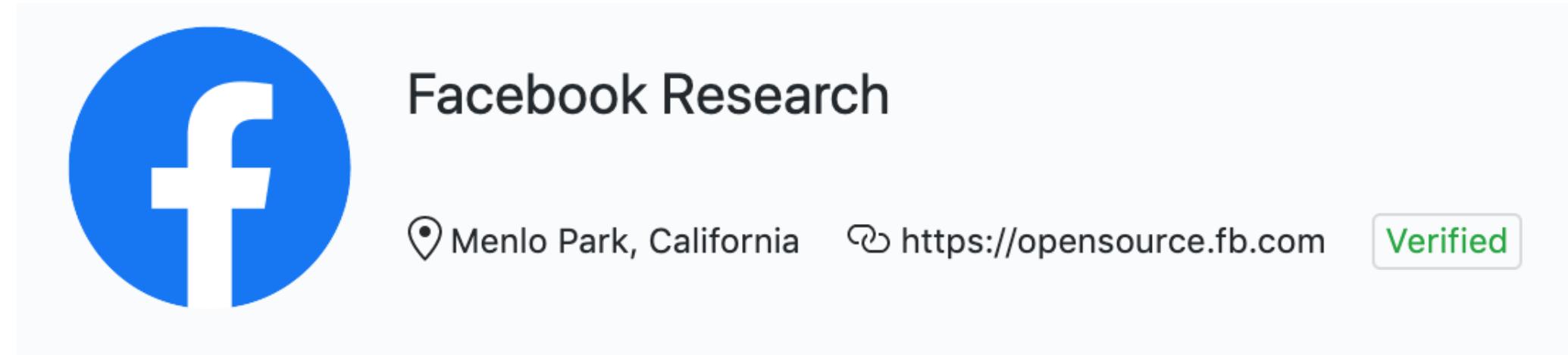
System Overview



Conclusion

- CLARA, a system developed and deployed at Facebook to estimate the uncertainty of human labels
- Extensive simulations and comparison with state-of-the-art
- Results on real Facebook deployment

And the source code



<https://github.com/facebook/clara>

