Comprehensive and Reliable Crowd Assessment Algorithms

MANAS JOGLEKAR
HECTOR GARCIA-MOLINA
ADITYA PARAMESWARAN

Background

- Crowdsourcing: Tasks such as image tagging
- Workers are often *unreliable*
  - Lack of motivation
  - Lack of skill
- Need to assess worker quality
- Need for Confidence intervals
  - 1/3 errors vs 10/30 errors
- Upper confidence threshold for worker filtering

Problem Setting

- **m** tasks ($t_1 \ldots t_m$)
- **n** workers ($w_1 \ldots w_n$)
- Non-regular data

		t_1	t_2	t_3	t_4	t_5
w_1		r_1	r_2	r_3	-	r_1
w_2		r_1	r_2	-	r_1	r_1
w_3		-	r_3	r_3	r_2	r_1
w_4		r_1	r_2	r_3	-	r_2

Problem Setting

- Binary tasks OR **k** responses ($r_1 \ldots r_k$)
- No gold standard
- Accuracy model
  - Worker $\mathbf{w_i}$ has error rate $\mathbf{p_i}$, or confusion matrix $\mathbf{P_i}$
  - Non-malicious workers (better than random)
  - Worker response **independent** of each other, given true answer
- Goal : Given user-specified confidence level c, find c-confidence interval for p_i or entries in P_i

Warm-up: 3 workers, binary tasks

- Equal false positive and negative error rates
- To find: confidence intervals for p_i, for each i
- Can be found using
  - Mean estimate for p_i
  - Variance of p_i estimate
- Easy if gold standard available
- Agreement rate q_{ij}, probability of worker w_i, w_j agreeing

Warm-up: 3 workers, binary tasks

- Compute agreement rates (q_{ij} for worker w_i, w_j)

	t_1	t_2	t_3	t_4	t_5
w_1	y	y	-	n	n
w_2	n	-	y	n	n
w_3	-	y	y	n	y
True	n	y	y	n	n

$E[q_{12}] = 2/3$

- $q_{ij} \sim p_i p_j + (1-p_i)(1-p_j)$
- So $p_i = f_i(q_{ij}, q_{ik}, q_{jk})$

$$= ½ - ½ \sqrt{((q_{ij} – ½)(q_{ik} – ½)/(q_{jk} – ½))}$$

Warm-up: 3 workers, binary tasks

- To find variance in p_i using q_{ij}'s we use:
- **Theorem**: $Y = f(X_1, X_2, X_3)$
  - X_i's **normal**, f **linear** ($\sim a_1X_1 + a_2X_2 + a_3X_3$)
  - $\mathrm{Var}(Y) = \Sigma_{i,j}\, a_i a_j \,\mathrm{Cov}(X_i, X_j)$
- Works for approximately normal (binomial), locally linear (differentiable)
- Linear coefficients given by partial derivatives

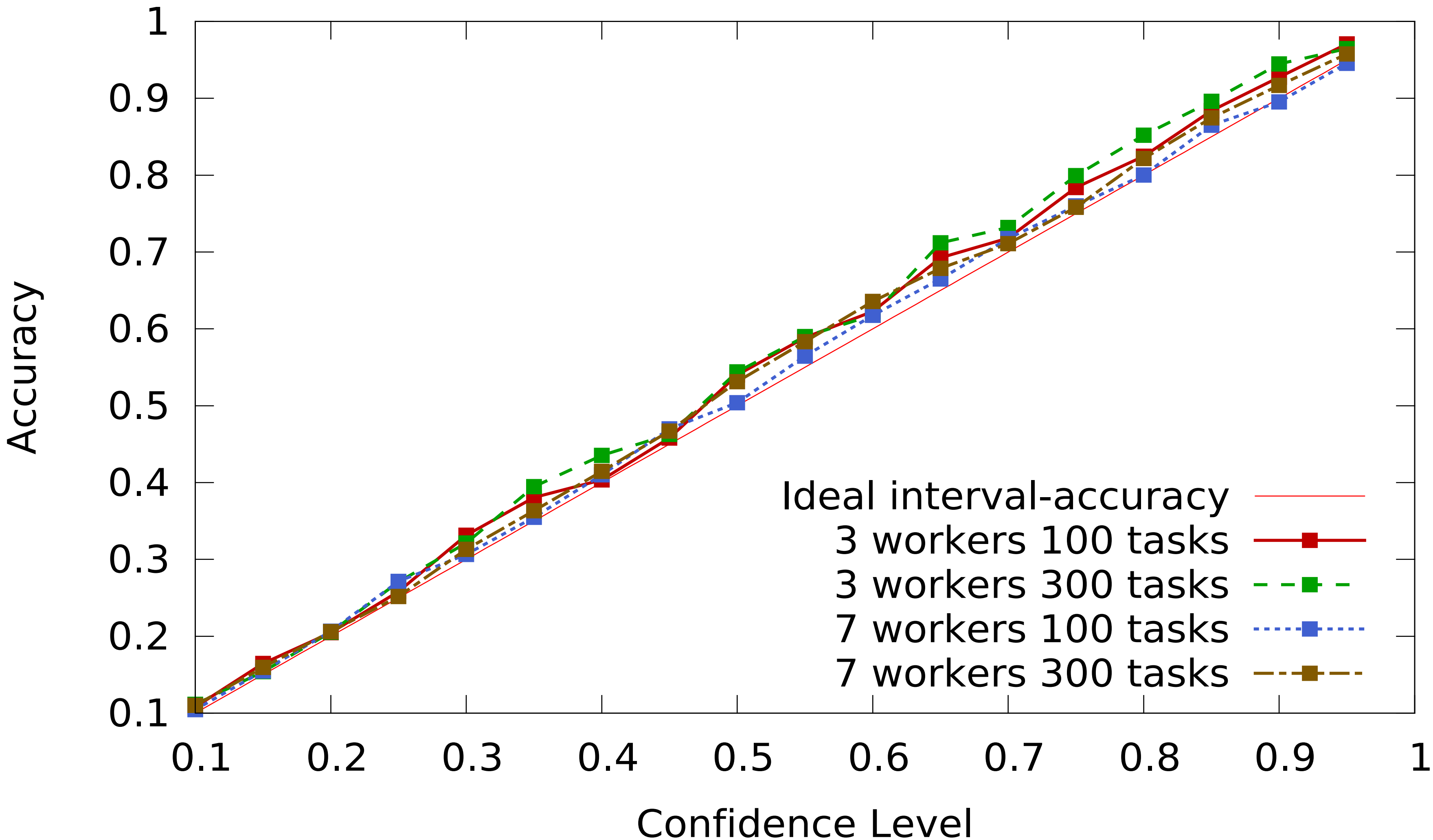
Warm-up: 3 workers, binary tasks

- p_i is Y; q_{ij}, q_{ik}, q_{jk} are X's
- Variance in p estimate depends on
  - Variance in q estimates
  - Covariances between q estimates
  - $\delta f_i/\delta q_{ij}$, $\delta f_i/\delta q_{jk}$
- $Var(p_i) = \Sigma_{q,q'}$ Cov (q,q') x $\delta f_i/\delta q$ x $\delta f_i/\delta q'$
- Estimate variances, derivatives
- Use $E[p_i]$, $Var(p_i)$ to get confidence interval

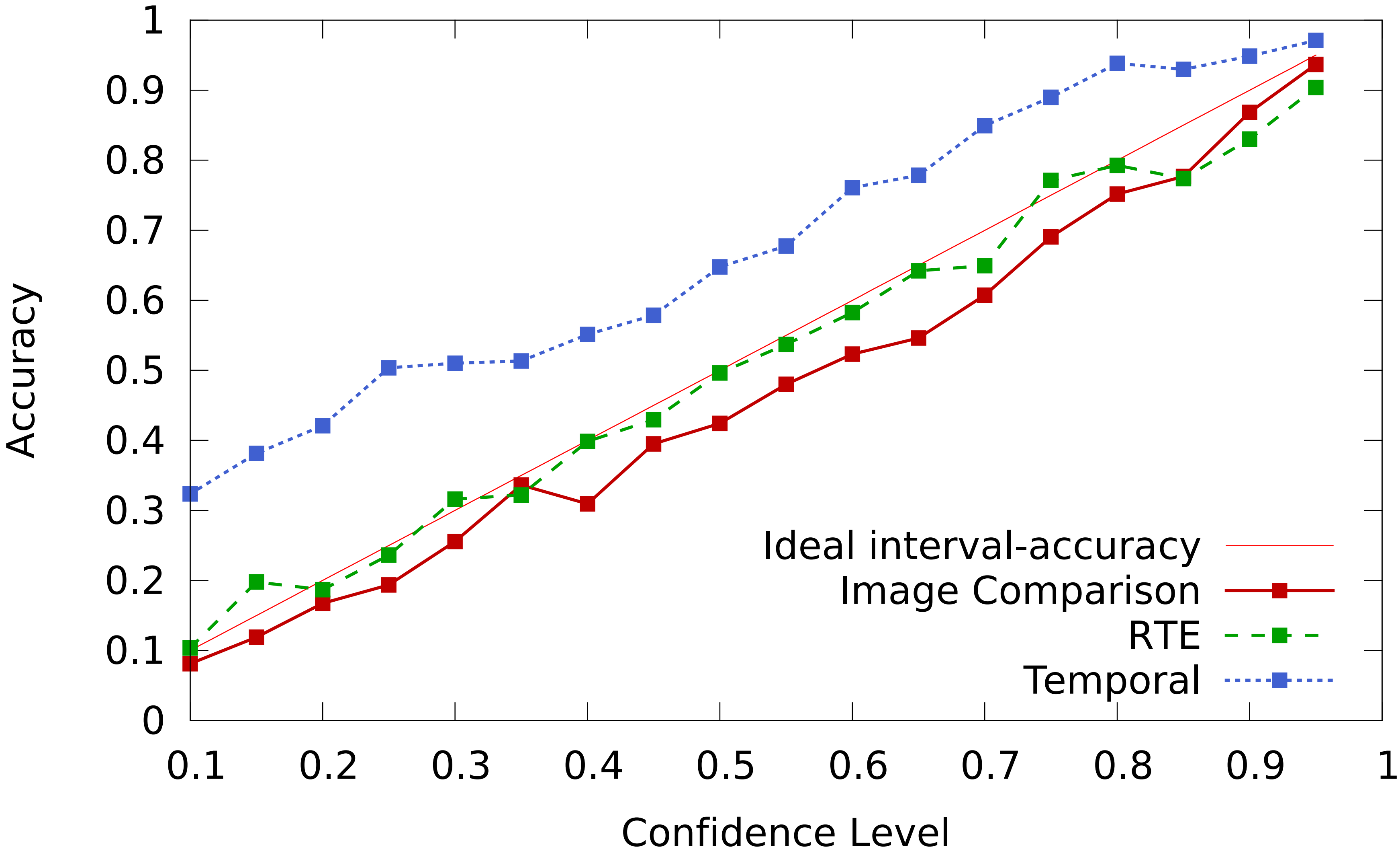
Calibration Experiment

- Accuracy is the fraction of times a confidence interval contains correct value
- Safe: Accuracy of c-confidence interval > c
- Tight: Accuracy <= c

Calibration Results (Synthetic Data)



Calibration Results (Real Data)



Generalizing to many workers

- Example: To evaluate w1
  - w4, w5 `good', `w2, w3' bad
  - groups (w1, w2, w3), (w1, w4, w5)

	t_1	t_2	t_3	t_4	t_5	t_6
w_1	y	y	n	n	y	-
w_2	y	-	n	n	-	n
w_3	y	y	y	n	-	-
w_4	n	y	n	n	y	-
w_5	y	n	n	n	-	n

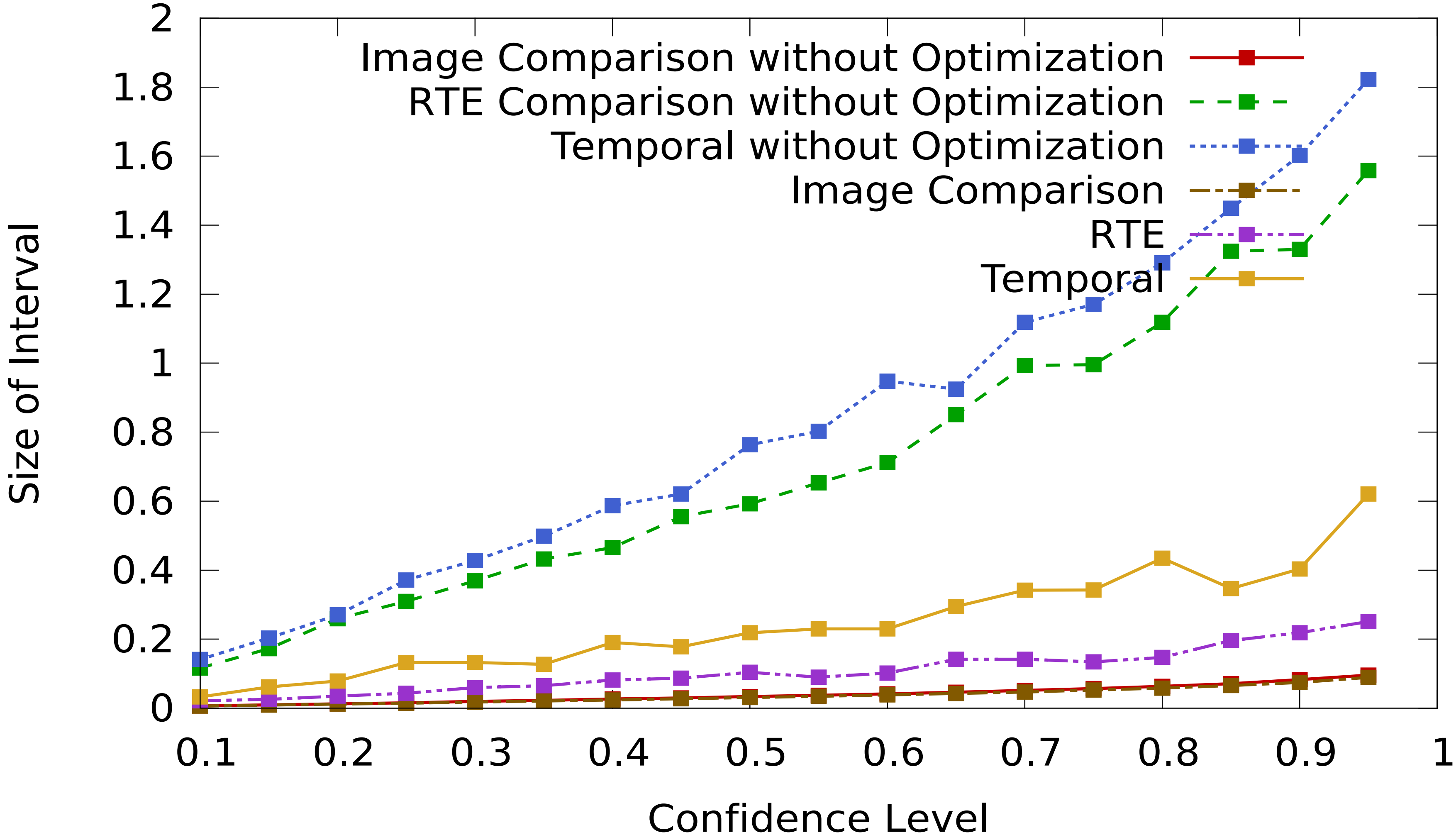
Generalizing to many workers

- Key Idea: Take multiple sets of 3 workers and combine estimates
- For each worker, form N/2 triples.
- Use each triple T_i to compute estimate e_i for p_1
- Compute variances, covariances of estimates
- Use weighted combination of estimates
  - $p = \Sigma_i a_i e_i$

Generalizing to many workers

- Optimum weights for combining
  - Covariance matrix A
  - Given by $A^{-1}L_{N/2}$ where L_k is a k-length vector with values 1/k
- Greedy way of group forming
  - Better to have two good workers in one group than one good worker in two groups, due to weighting
  - Greedily form groups

Results: Weight Optimization



3 Workers Non-binary Tasks

- Confusion matrix P_i , Selectivity vector S, diagonal S^D

e.g.

0.8	0.1	0.3
0.1	0.7	0.1
0.1	0.2	0.6

,

0.4
0.2
0.4

,

0.4	0	0
0	0.2	0
0	0	0.4

- P_i's, S : Column-stochastic, unknown
- Observation probabilities given by $\mathbf{P_i S}$

0.46
0.22
0.32

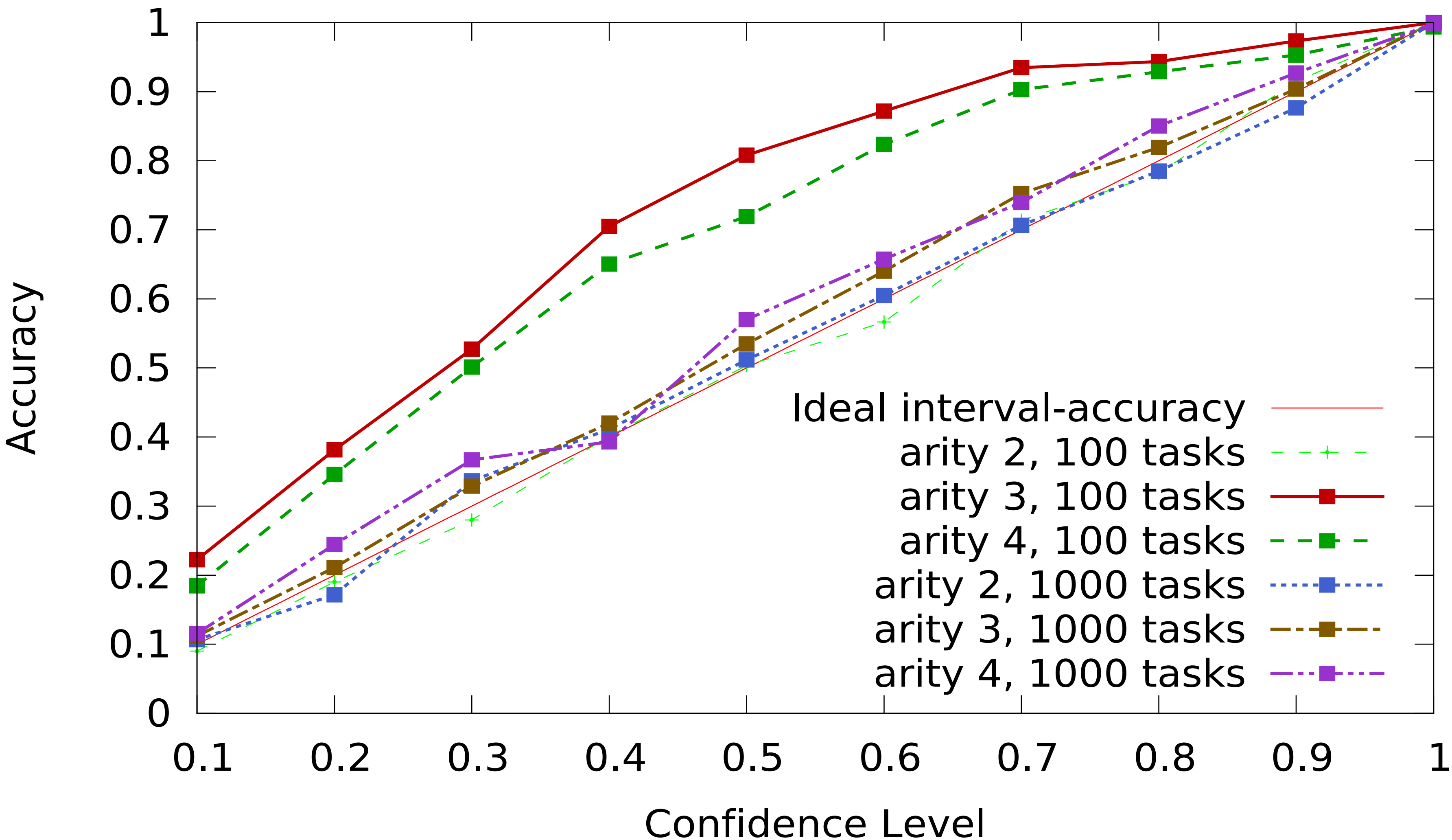
3 Workers Non-binary Tasks

- Compute comparison matrices (frequencies of response-pairs of workers w_i, w_j)
- $C_{ij} \sim P_i S^D P_j^T$
- Compute $D_i = C_{ij}(C_{jk}^T)^{-1} C_{ki} = P_i S^D P_i^T$
- Eigenvalue decomposition of D_i gives $V_i = U \sqrt{S^D}\, P_i$, for a unitary U.

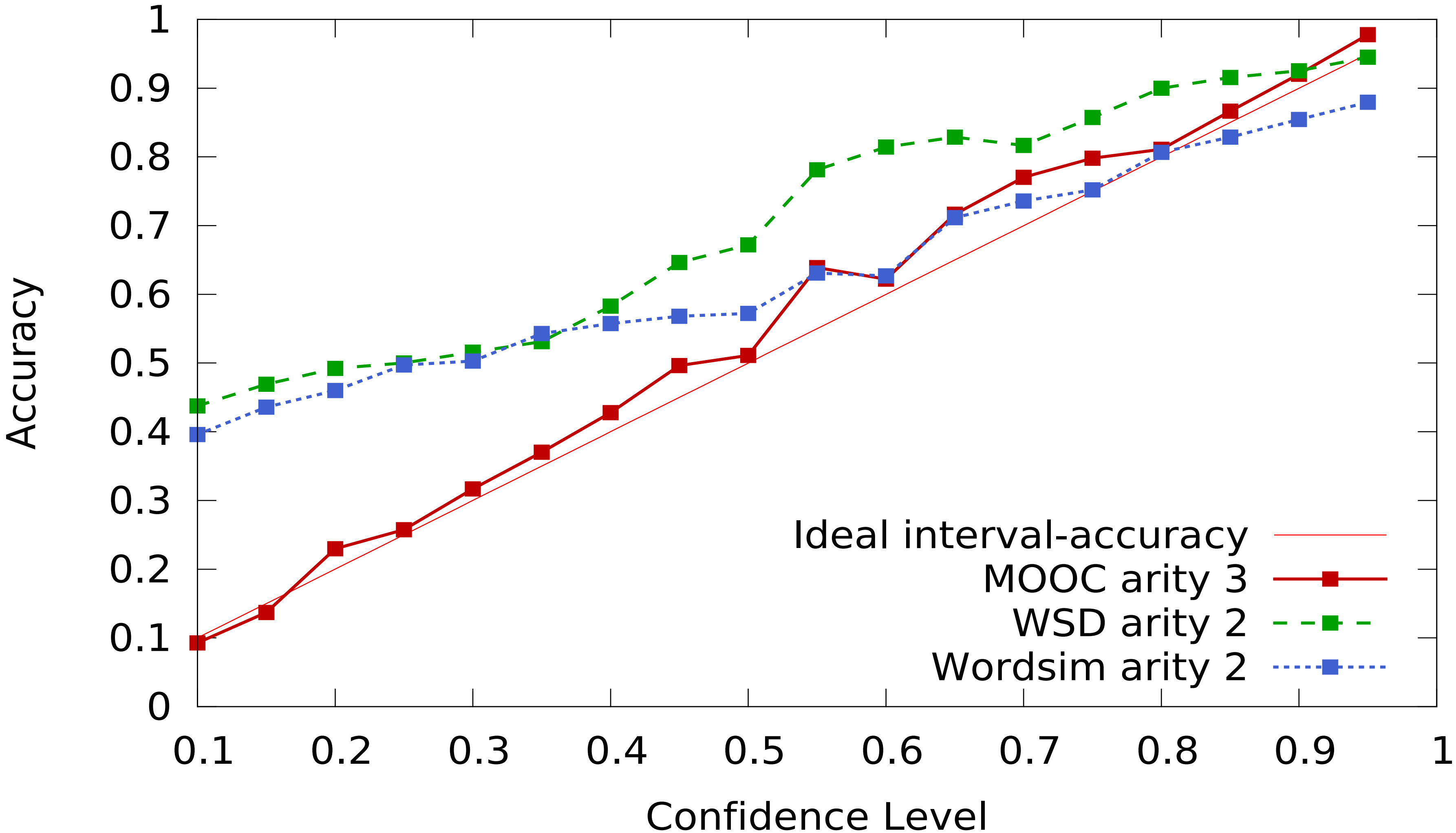
3 Workers Non-binary Tasks

- U can be recovered using 3-way comparisons
- S^D can be recovered using column stochasticity of P^i
- For confidence intervals, use variances/derivatives like in the binary tasks case
- Details in paper

Results: Calibration (Synthetic Data)



Results: Calibration (Real Data)



Conclusion

- We can come up with accurate, tight confidence intervals in very general scenarios
- Confidence Intervals are useful for filtering workers
- Thank You! Questions?