Using Replicates in Information Retrieval Evaluation

Ellen Voorhees

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Cranfield

- Set of documents
  - a document is a “package of information”
  - representative of documents in target application
- Set of topics
  - topic: statement of information need
  - assumed to be a(n independent) sample of the universe of user questions
  - large samples needed since there is known to be a large variance in retrieval results across queries
- Relevance judgments
  - which docs should be retrieved for each topic
  - foundation of evaluation measures

Cranfield

- Each system (variant) produces a ranked list of documents for each topic. This is a ‘run’.
- A score is computed for each topic in each run based on the ranks at which relevant documents are retrieved.
- Retrieval results are reported as averages over a set of topics.

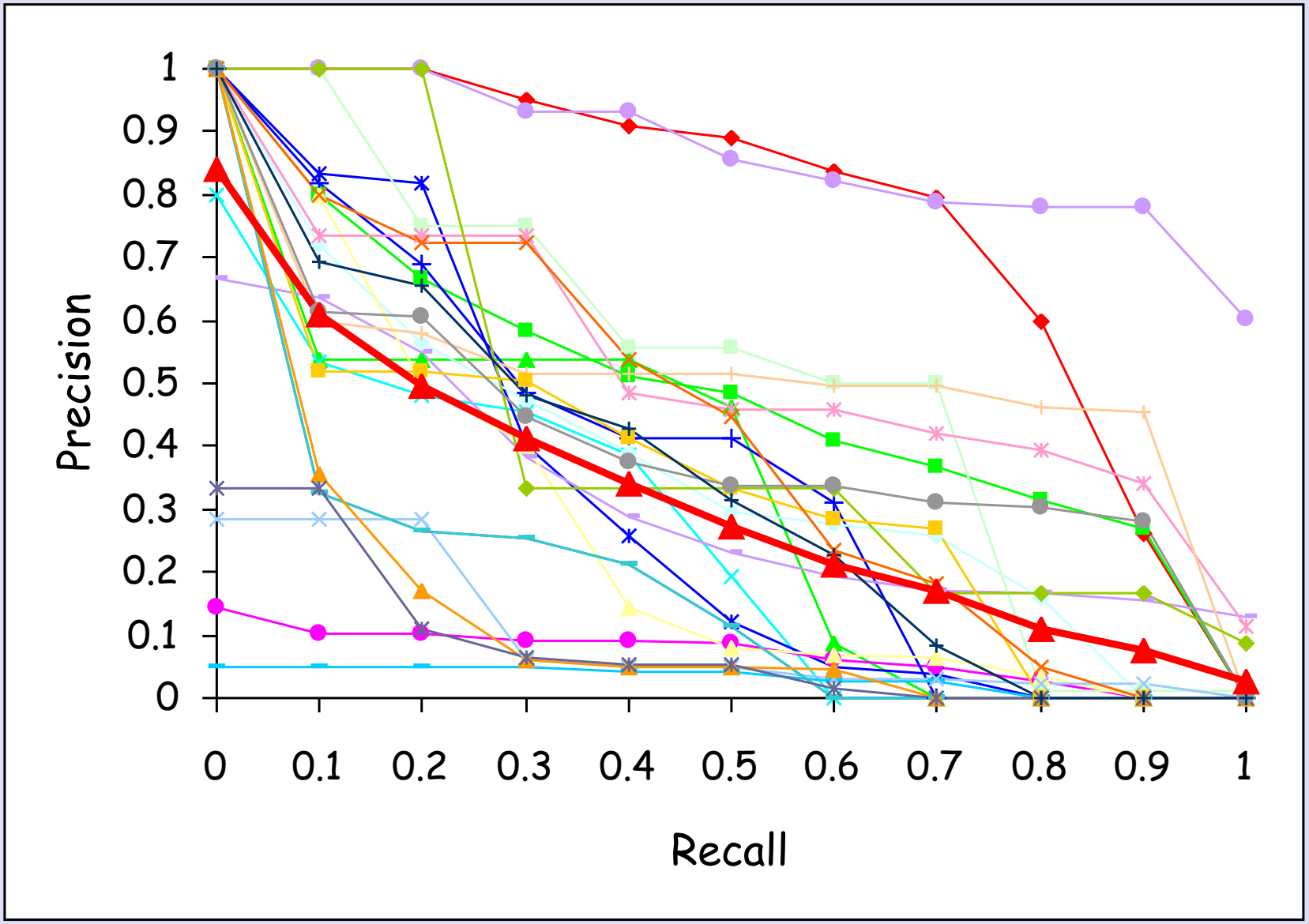
ANOVA Model of Retrieval Score

$$y_{ij} = \mu + t_i + s_j + \varepsilon_{ij}$$

where

- y_{ij} is the value of the measure of the *j* th system on the *i* th topic,
- μ is the true mean,
- t_i is the effect due to topic *i*,
- s_j is the effect due to system *j*, and
- ε_{ij} is all other variation (the error)

Cranfield


Precision
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Recall

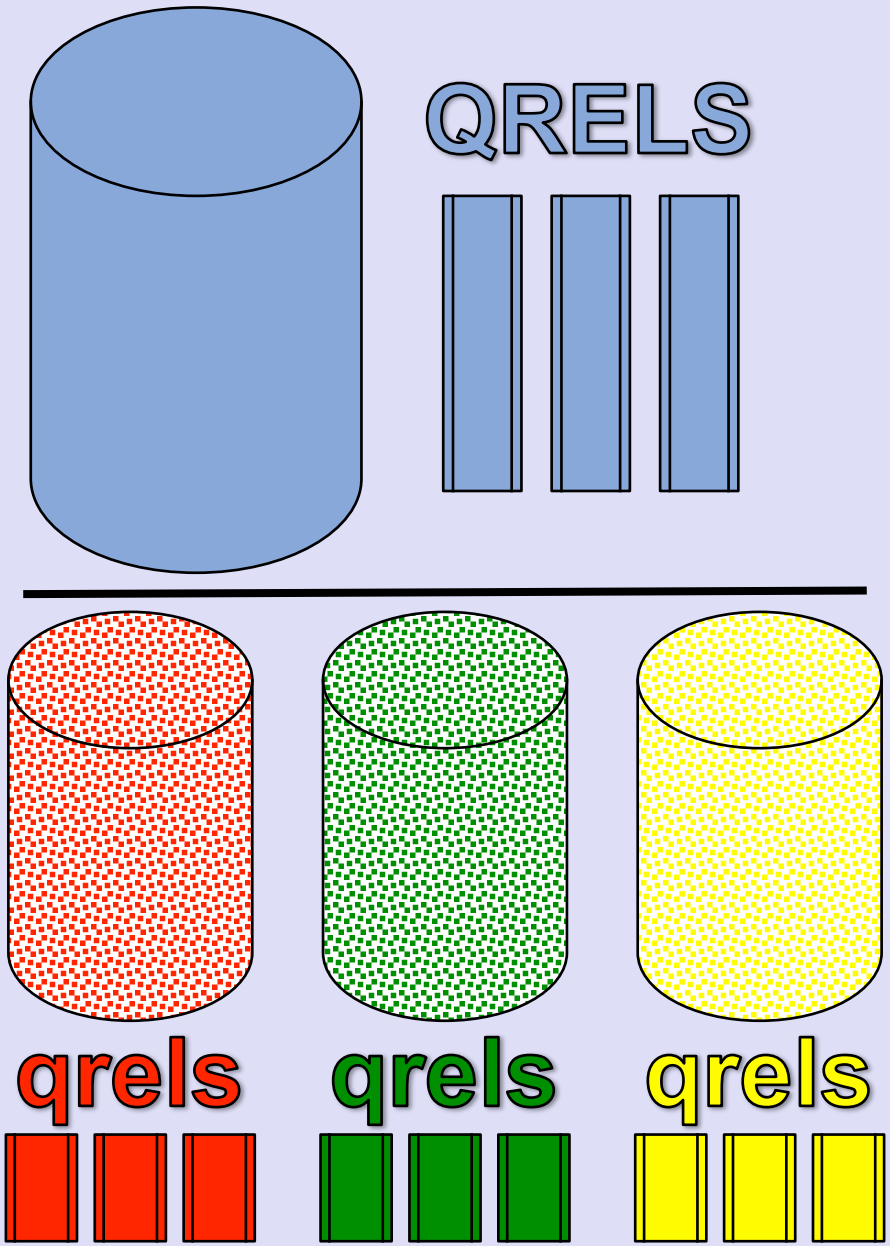
Cranfield

- We know from previous work (e.g., Banks et al., 1999) that system effect, topic effect, and interaction effects are all significant and large
- So, more accurate model would include interaction term:

$$y_{ij} = \mu + t_i + s_j + (ts)_{ij} + \varepsilon_{ij}$$

- But we have only one score per topic-system. How can we get multiple scores and thus capture interaction effect?

Replicating Scores



- Randomly partition document set into *x* segments
- Create *x* subsets of relevance judgments (‘qrels’), where each subset is restricted to docs from a single partition
- For a given run, create *x* subruns, where each subrun is restricted to docs from a single partition
- Calculate score for each topic in each subrun using corresponding judgment subset

Partitions

- Critical assumption is that scores computed from partitions are representative of that run's score in original collection
  - for random assignment of documents and small *x*, very likely to be true [see Sanderson et al. 2012]…
  - …assuming number of relevant per partition roughly balanced
    - subsequent steps require at least one relevant per topic per partition

Bootstrap ANOVA

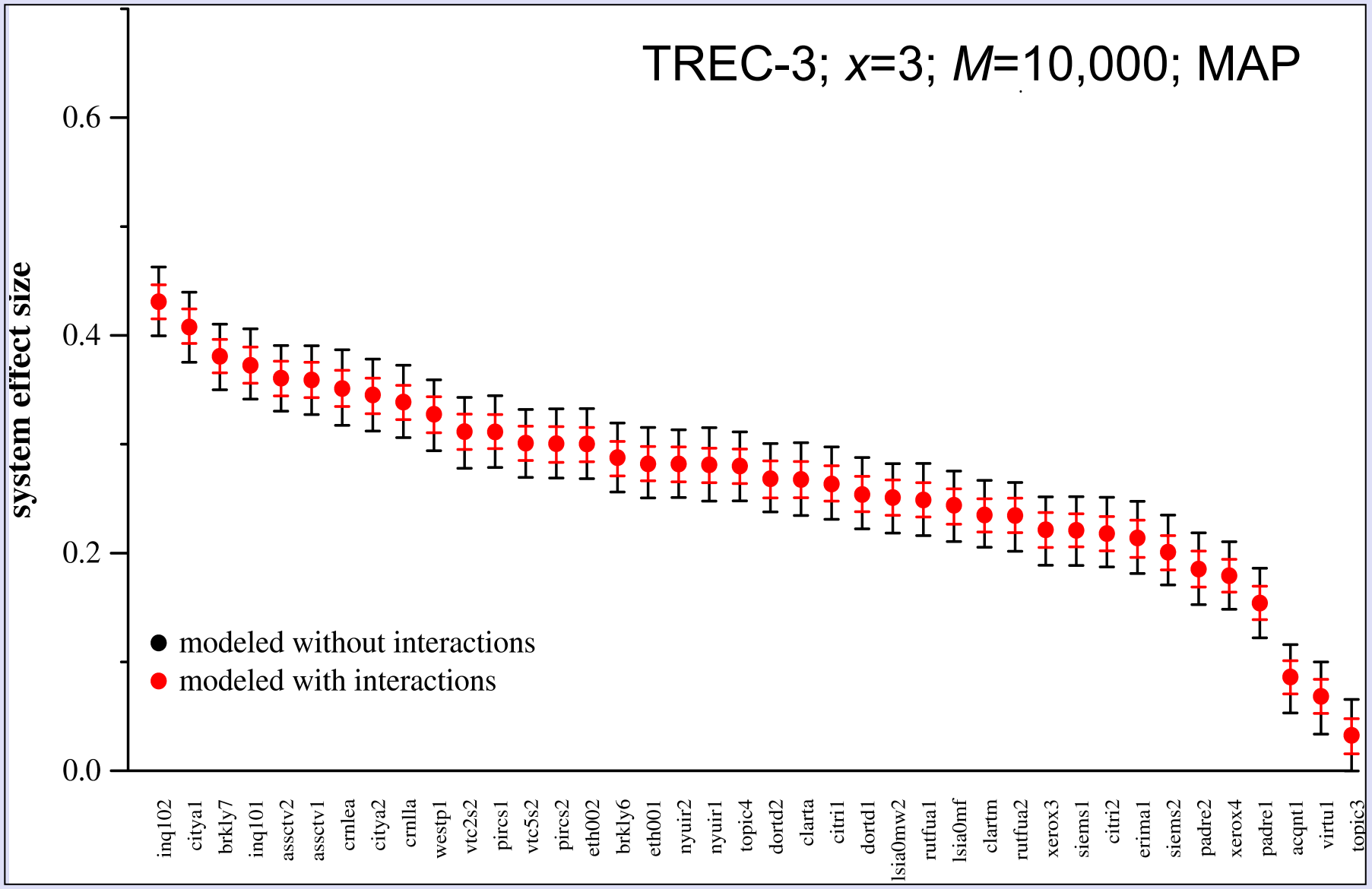
- Now have *x* scores per system-topic combo
- Use bootstrap ANOVA to get distribution of estimated parameters of the model
  - use least-squares fit of data to estimate model parameters
    - have fully balanced design since all runs have scores for all topics
  - for *M* times, randomly assign the residuals of the initial model fit across all the system-topic combos and compute new estimates of the model parameters using these values

Confidence Intervals on S_j

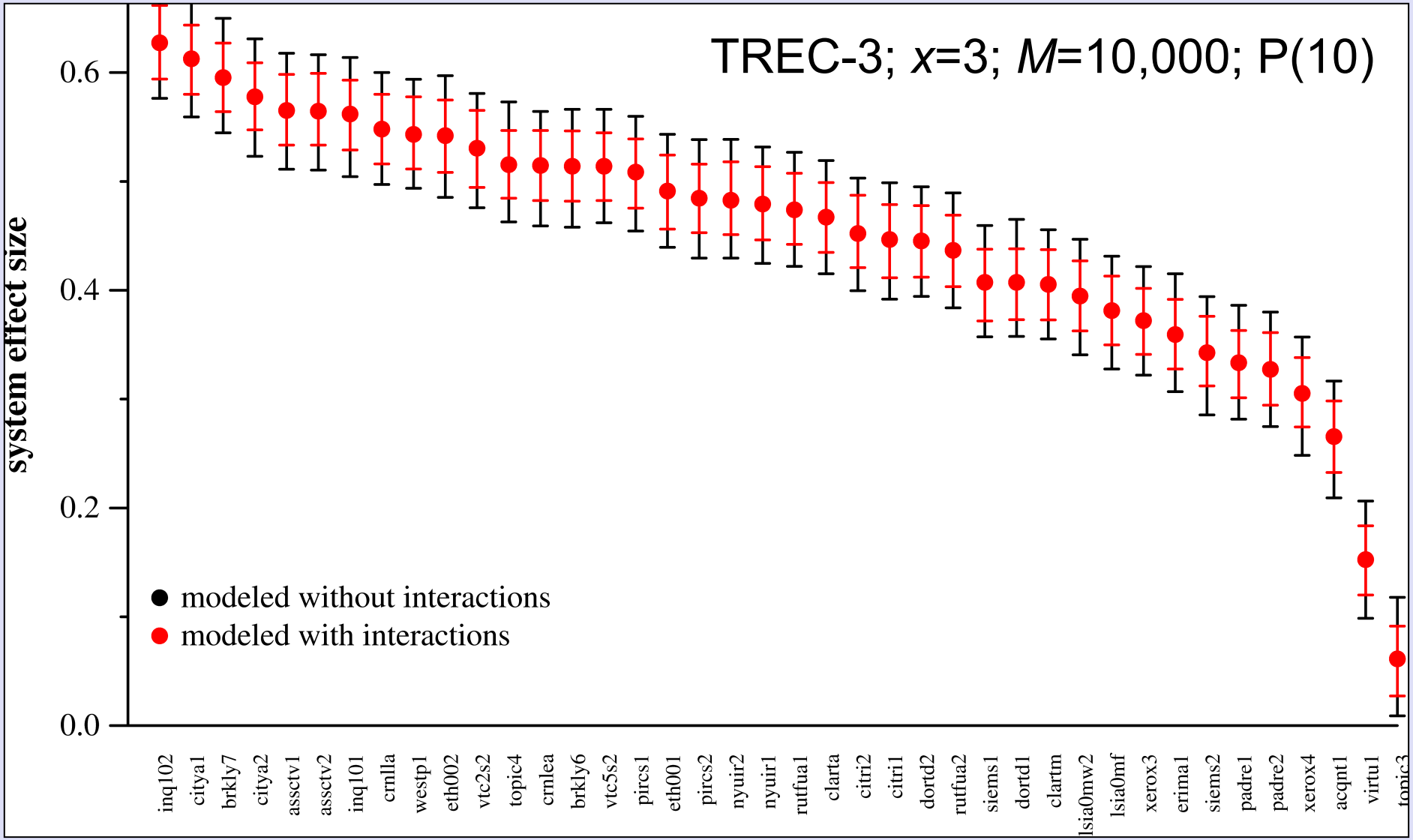
- Result of the bootstrap process is *M* estimates of all the model parameters, and thus, in particular, of the system effect s_j for all j

- Use 2.5% and 97.5% quantiles over the *M* estimates of s_j for the 95% confidence interval of the system effect for system j

Confidence Intervals on S_j



Confidence Intervals on S_j



Confidence Intervals on S_j

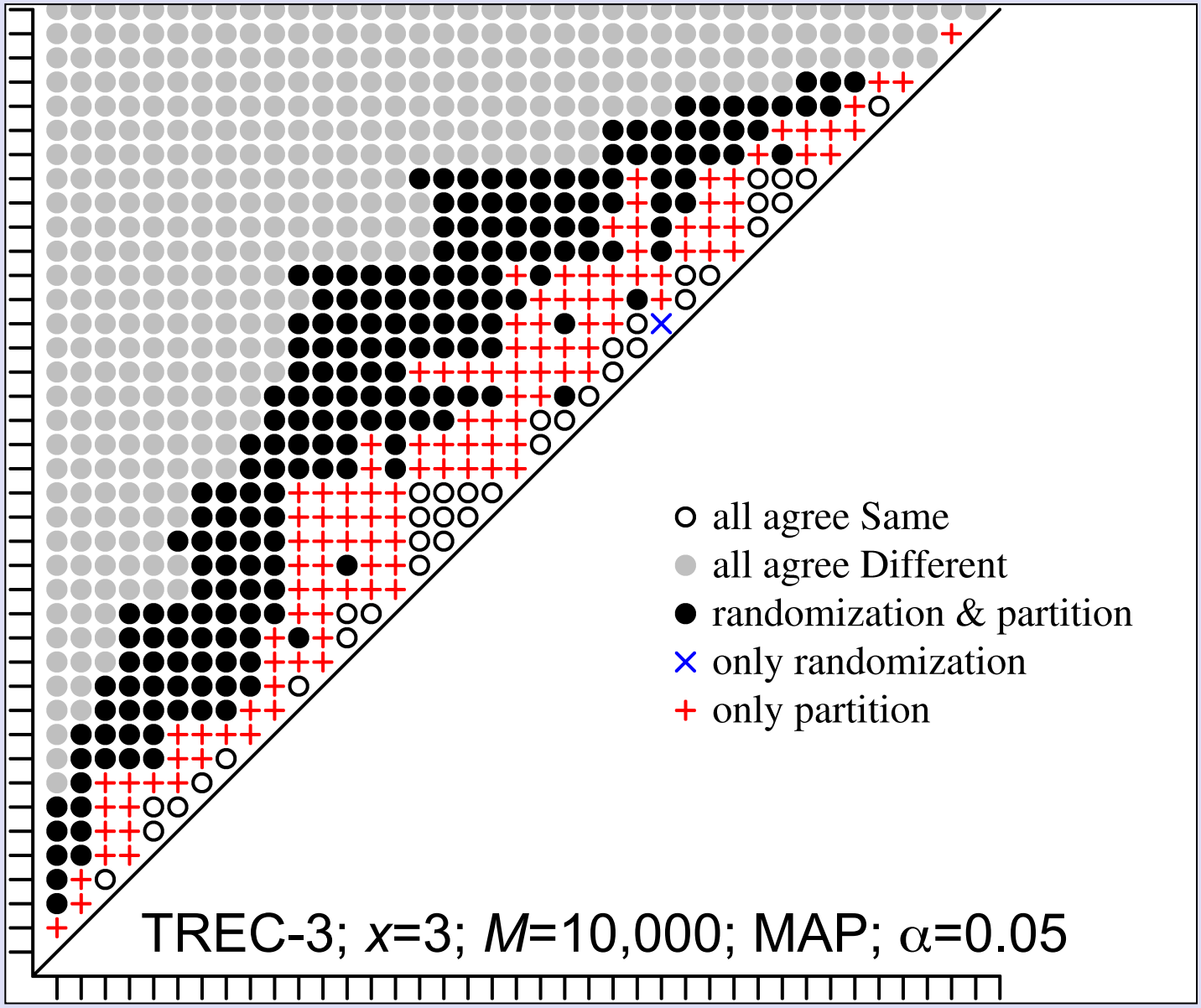
Coll.	Meas.	*x*=2		*x*=3		*x*=5	
TREC-3	MAP	0.075	[0.071, 0.082]	0.064	[0.060, 0.069]	0.055	[0.052, 0.058]
		0.029	[0.026, 0.031]	0.032	[0.030, 0.034]	0.033	[0.031, 0.034]
	P(10)	0.130	[0.122, 0.140]	0.106	[0.099, 0.112]	0.081	[0.076, 0.086]
		0.065	[0.061, 0.069]	0.065	[0.061, 0.071]	0.055	[0.052, 0.060]
TREC-8	MAP	0.088	[0.082, 0.094]	0.078	[0.070, 0.084]	0.069	[0.065, 0.074]
		0.039	[0.035, 0.042]	0.044	[0.040, 0.047]	0.049	[0.046, 0.053]
	P(10)	0.122	[0.115, 0.134]	0.098	[0.093, 0.109]	0.071	[0.066, 0.076]
		0.061	[0.055, 0.065]	0.061	[0.057, 0.067]	0.048	[0.045, 0.053]
Terabyte	infAP	0.064	[0.064, 0.071]	0.058	[0.055, 0.064]	---	
		0.032	[0.032, 0.035]	0.037	[0.035, 0.040]		

Mean [min, max] length of confidence interval on system effect for different collections and different number of partitions. Intervals for models with no interactions are on top and models with interactions are on bottom.

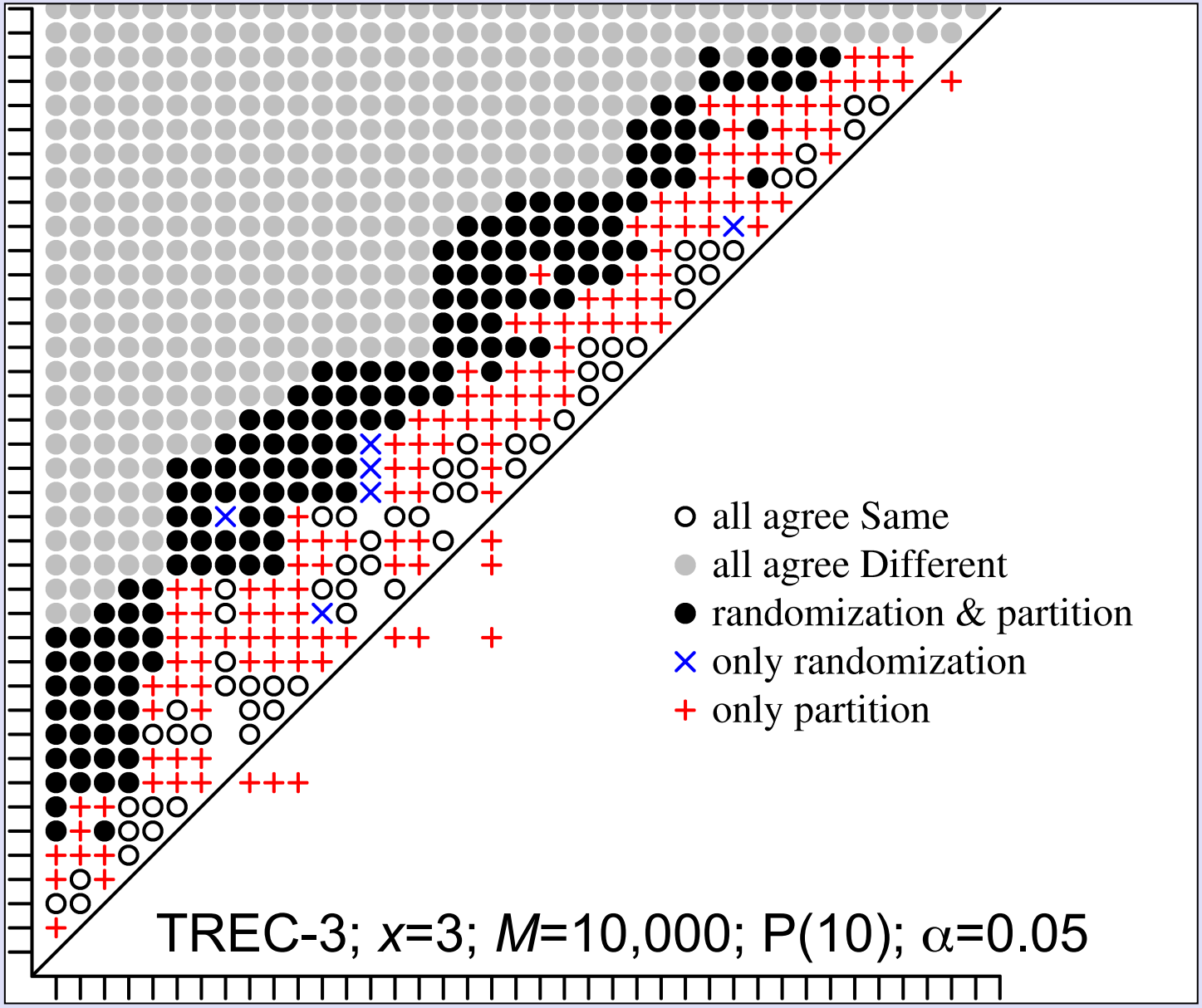
Distinguishing Among Systems

- Use confidence intervals to compute p-values which allow us to infer the likelihood that a given pair of systems are different
  - Use Benjamini-Hochberg correction for multiple comparisons
  - member of the family of methods that control false discovery rate
- Compare to sets of different systems as determined through (uncorrected) t-test and randomization test

Decision for system i > system j



Decision for system i > system j



Distinguishing Among Systems

		TREC-3 (780 pairs)		TREC-8 (8256 pairs)		Terabyte (3160 pairs)
		MAP	P(10)	MAP	P(10)	infAP
t-test	$\alpha = 0.05$	409 (52.4%)	411 (52.7%)	4164 (50.4%)	4317 (52.3%)	1261 (39.9%)
	$\alpha = 0.01$	310 (39.7%)	324 (41.5%)	3437 (41.6%)	3695 (44.8%)	976 (30.9%)
Random-ization	$\alpha = 0.05$	619 (79.4%)	579 (74.2%)	6325 (76.6%)	5663 (68.6%)	2114 (66.9%)
	$\alpha = 0.01$	544 (69.7%)	491 (62.9%)	5571 (67.5%)	4903 (59.4%)	1700 (53.8%)
3 Parts	$\alpha = 0.05$	741 (95.0%)	712 (91.3%)	7413 (89.8%)	7112 (86.1%)	2662 (84.2%)
	$\alpha = 0.01$	728 (93.3%)	693 (88.8%)	7150 (86.6%)	6786 (82.2%)	2540 (80.4%)
2 Parts	$\alpha = 0.05$	743 (95.3%)	712 (91.3%)	7510 (91.0%)	7254 (87.9%)	2742 (86.8%)
	$\alpha = 0.01$	730 (93.6%)	700 (89.7%)	7269 (88.0%)	6930 (83.9%)	2637 (83.4%)

Distinguishing Among Systems

- All tests agree on the majority of pairs
- Randomization and partition methods find (many) more significant differences than t-test
- Partition method cannot distinguish among systems whose difference is small relative to size of the residuals across the entire run set
- Never observed a conflict (two different tests both distinguish system pair, but prefer different runs)

Number of Partitions

- Fewer partitions give slightly better results
  - smaller confidence intervals smaller
  - greater number of significant differences found
  - confidence interval size reflects total variability in system; fewer partitions produce larger partitions and somewhat more stable scores
- Fewer partitions easier to use in practice

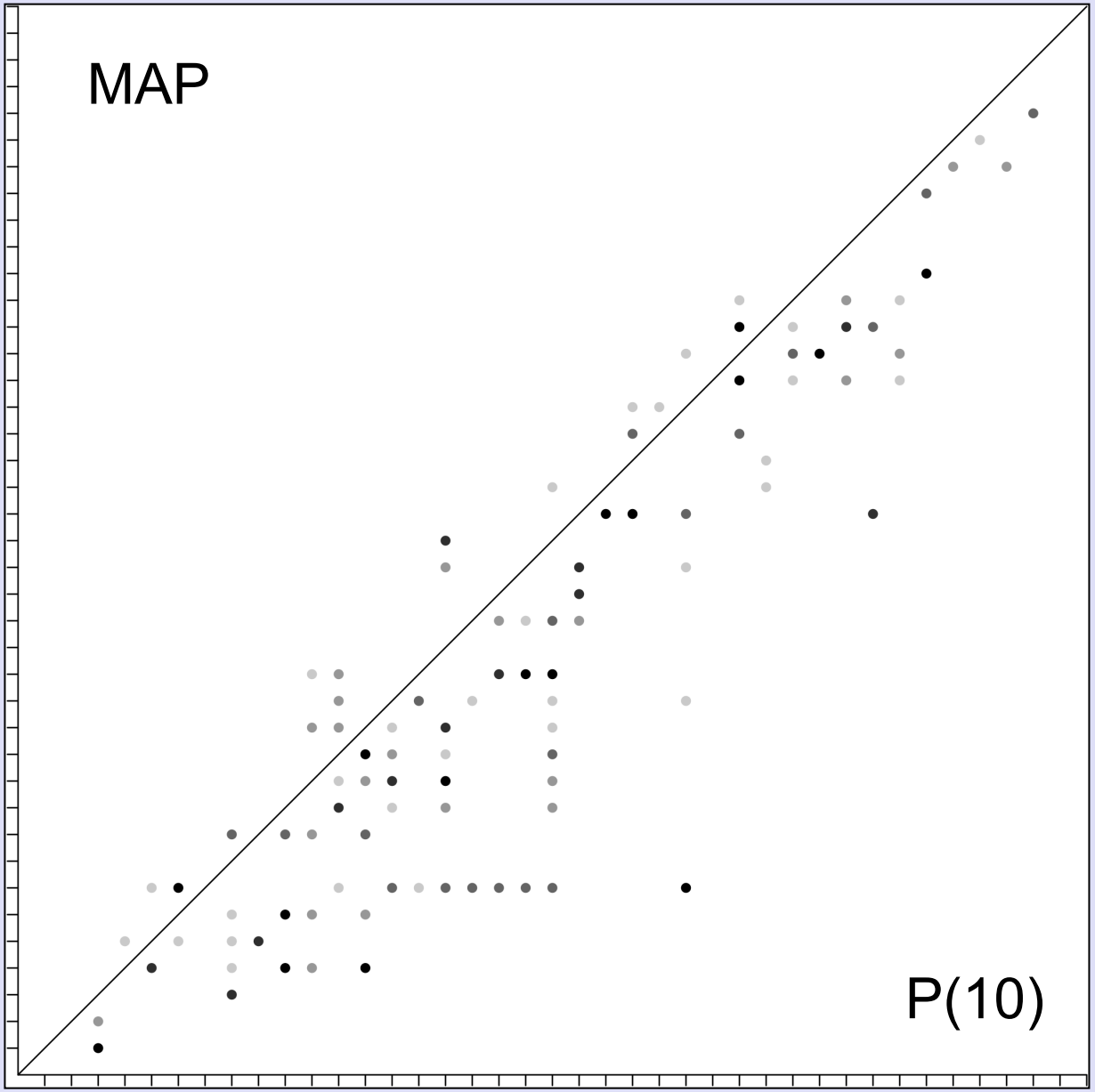
Additional Experiments

- How is partition method impacted by original split of documents?
- How many runs are necessary?

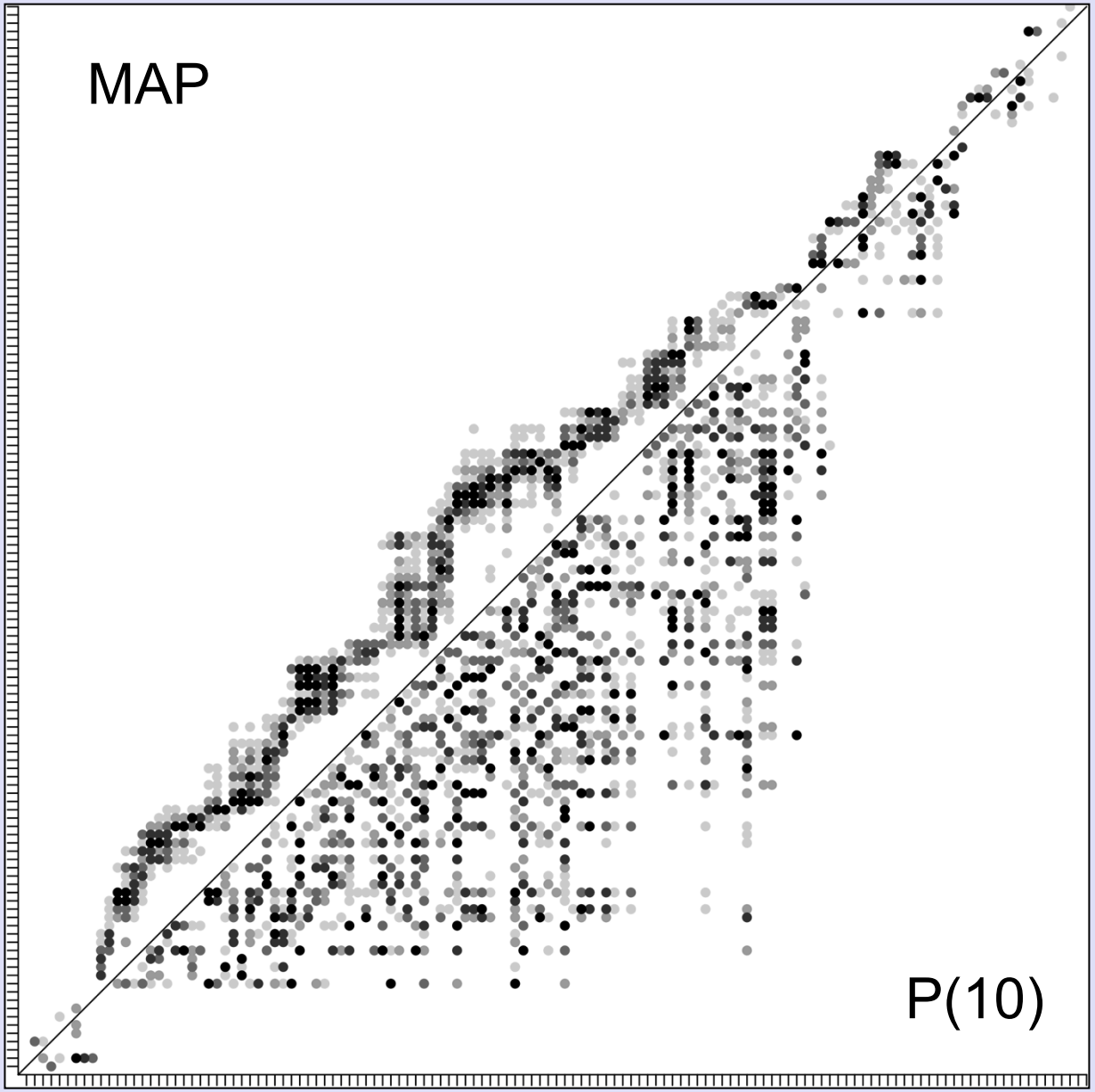
Stability Over Different Partitions

- Generate 10 new 2-partition splits of document collection (so 11 total different splits)
- Calculate p-values as above for each split; for each run pair, count number of times significance decision is the same
  - six possible agreement outcomes per pair
    11-0 10-1 9-2 8-3 7-4 6-5
    White… Gray… Darker Gray… Black
  - used α=0.05, *M*=10,000

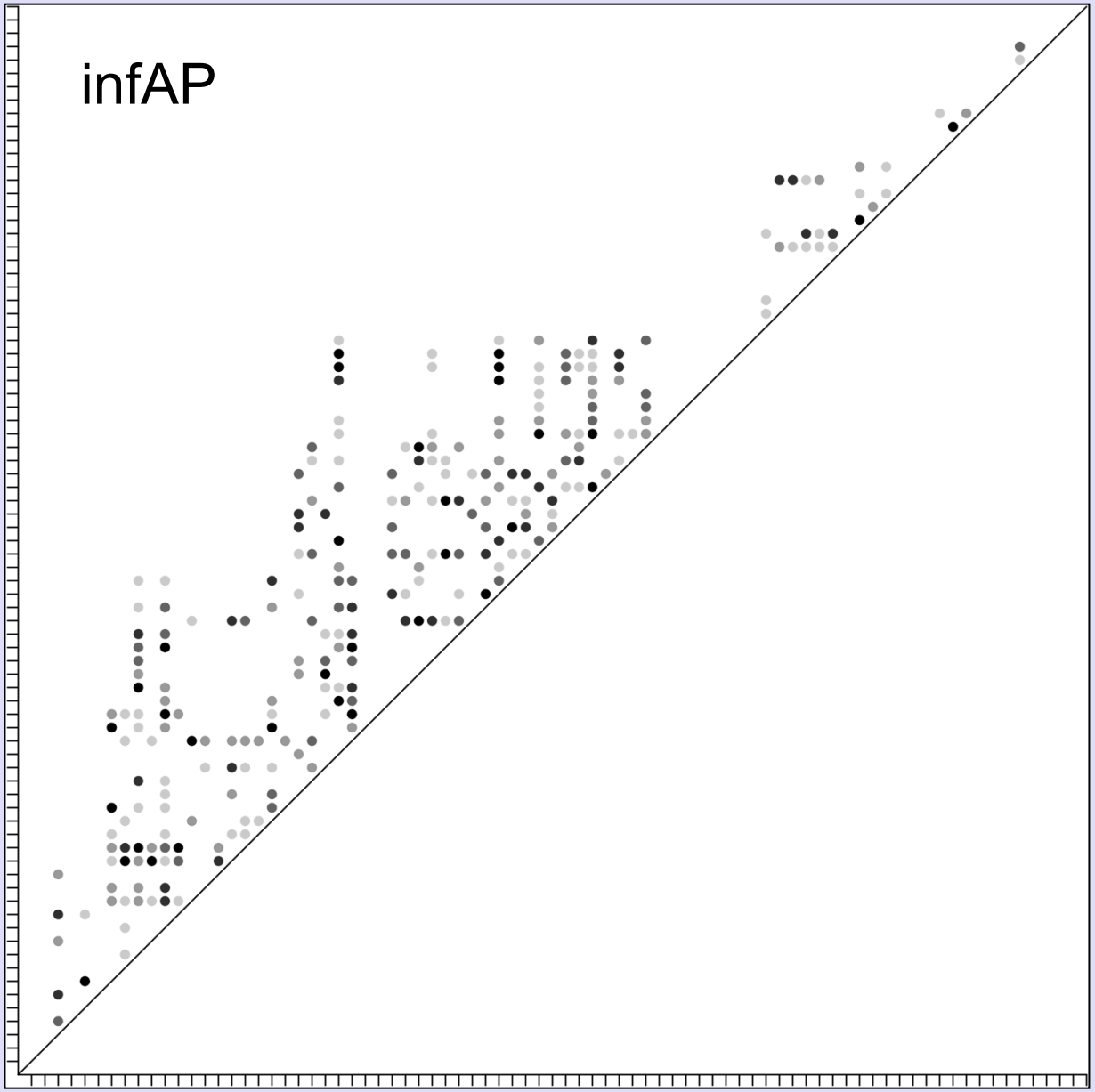
Agreement Over Different Splits: TREC-3



Agreement Over Different Splits: TREC-8



Agreement Over Different Splits: Terabyte



Agreement Over Different Splits

(x=2; α=0.05; 11 splits total)

	TREC-3 (780 pairs)		**TREC-8 (8256 pairs)**		**Terabyte (3160 pairs)**
	MAP	P(10)	MAP	P(10)	infAP
10:1	8 (1.0%)	23 (2.9%)	235 (2.8%)	357 (4.3%)	95 (3.0%)
9:2	5 (0.6%)	17 (2.2%)	133 (1.6%)	244 (3.0%)	60 (1.9%)
8:3	2 (0.3%)	17 (2.2%)	92 (1.1%)	182 (2.2%)	43 (1.4%)
7:4	1 (0.1%)	11 (1.4%)	95 (1.2%)	155 (1.9%)	38 (1.2%)
6:5	2 (0.3%)	14 (1.8%)	86 (1.0%)	165 (2.0%)	33 (1.0%)
Total	18 (2.3%)	82 (10.5%)	641 (7.8%)	1103 (13.4%)	269 (8.5%)

Agreement Over Different Splits

- Fewer than 15% of run pairs ever have any disagreement
  - but those are likely the pairs we care about!
- Basing significance decisions on a single split probably a bad idea, so incorporate multiple independent splits into protocol
  - unclear how best to combine decisions
  - most conservative method of forcing unanimous agreement still finds more differences than randomization test or t-test

Distinguishing Among Systems

(x=2; α=0.05)

	TREC-3 (780 pairs)		TREC-8 (8256 pairs)		Terabyte (3160 pairs)
	MAP	P(10)	MAP	P(10)	infAP
t-test	409 (52.4%)	411 (52.7%)	4164 (50.4%)	4317 (52.3%)	1261 (39.9%)
Randomization	619 (79.4%)	579 (74.2%)	6325 (76.6%)	5663 (68.6%)	2114 (66.9%)
3 Partitions	741 (95.0%)	712 (91.3%)	7413 (89.8%)	7112 (86.1%)	2662 (84.2%)
2 Partitions	743 (95.3%)	712 (91.3%)	7510 (91.0%)	7254 (87.9%)	2742 (86.8%)
Unanimous decision over 11 2-partition splits	733 (94.0%)	677 (86.8%)	7152 (86.6%)	6616 (80.1%)	2595 (82.1%)

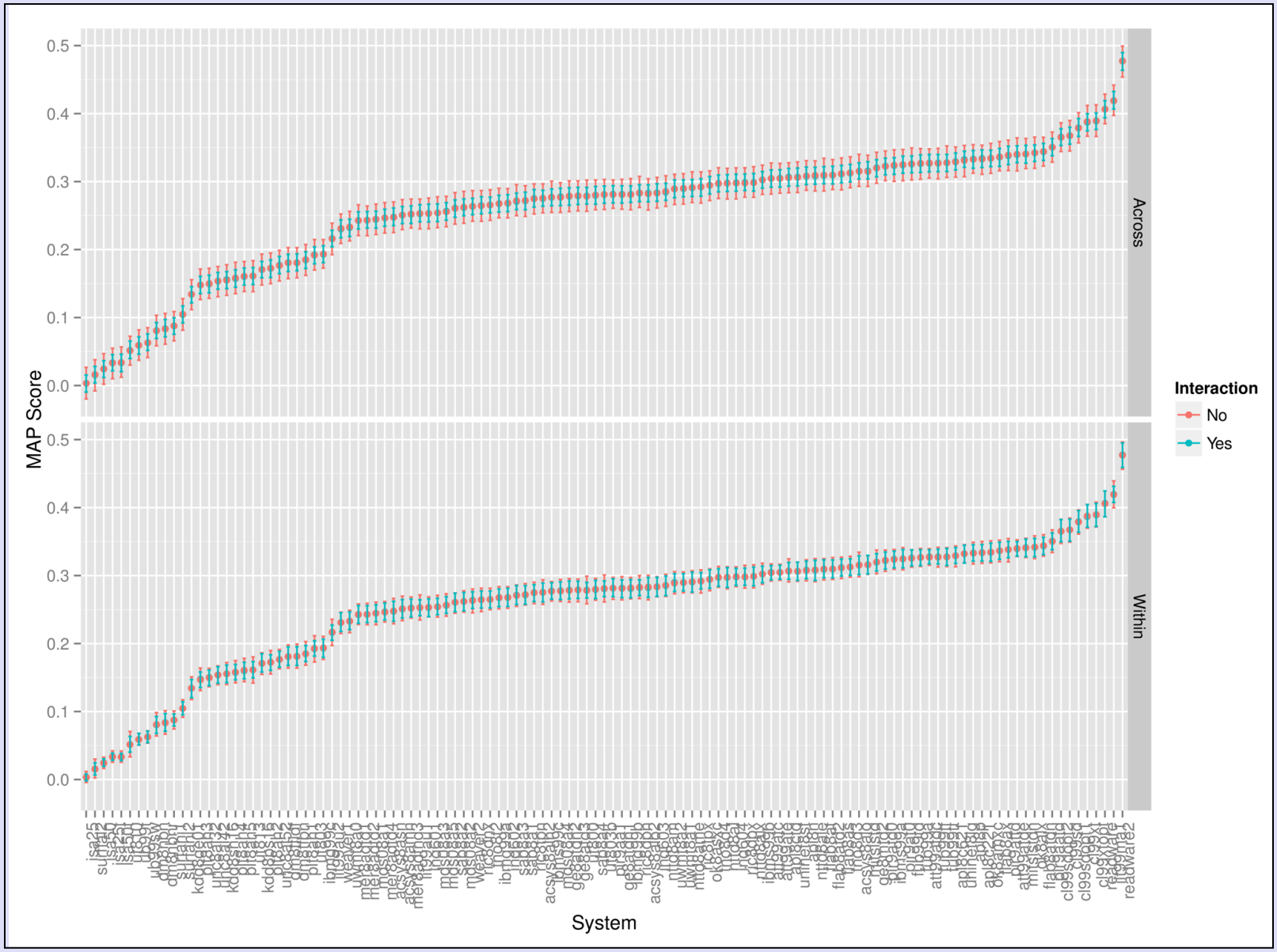
How Many Runs?

- All experiments used entire set of runs contributed to a TREC track.
- Can the method be used with a small set of runs such as those produced by a single research group's experiment?

Within- vs. Across-Teams Models

- Perform replicates process using just the set of runs contributed by a single TREC participant; compare to results when using whole track's set of runs as run set.
  - used single split with x=3
  - M=1000 for individual groups; M=10,000 for all
  - MAP as measure (so just TREC-3 and TREC-8 colls)
  - α = 0.05
- TREC-3: 17 groups submitted a pair of runs
- TREC-8: 32 groups each submitted 2-5 runs
  - total of 119 runs involved in some group's run set

Within- vs. Across-Teams



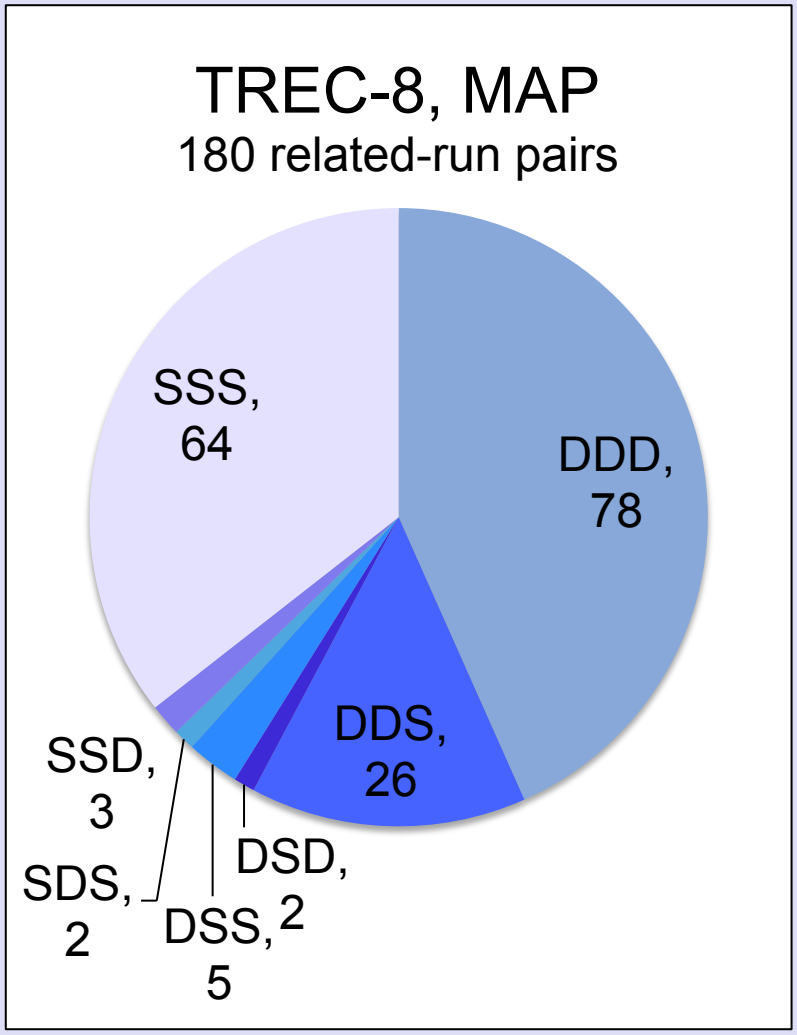
Within- vs. Across-Teams: CIs

	TREC-3		TREC-8	
	Interactions		Interactions	
	Without	With	Without	With
Across	0.064	0.032	0.078	0.044
Within	0.039	0.030	0.052	0.040

Mean confidence interval size on the system effect for runs: with-interactions in the model vs. without and across-all-track-systems vs. within-same-team's-systems

- Smaller confidence intervals for Within- vs. Across confirms that related runs are less variable than whole set. But further reduction for With-interactions vs. Without shows there is still a significant system-topic interactions effect.

Within- vs. Across-Teams Significance Decisions



- Triples formed by decision ('Same', 'Different') for Across-Teams, Within-Teams, & Randomization tests
- Vast majority (142/180) unanimous decision
- When Across- vs. Within- differ, generally Within- can't detect difference
  - exception: very ineffective runs whose total score is same size as residuals for Across-case
  - Within- case produces more extreme values (~ 0 or 1)

Future Work

- How best to incorporate multiple splits into single significance decision
- Relax requirement for completely balanced design in ANOVA
  - balanced design is computationally efficient
  - but restricts the partitions that can be used because of the requirement to have relevant documents in all partitions
  - handling number relevant issue in some way might cause larger number of partitions to be more favorable
- Release procedure implementing process a la trec_eval