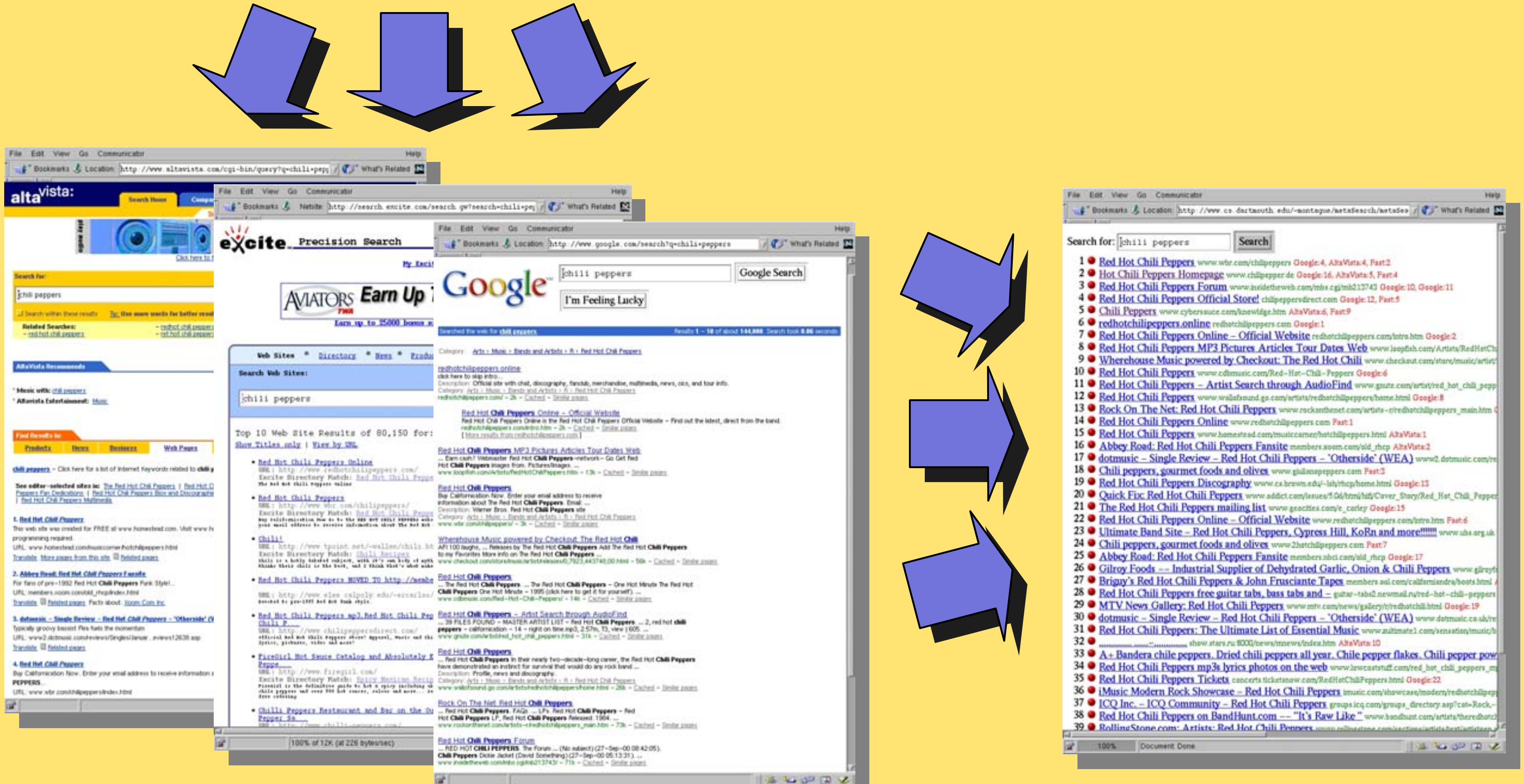
Models for Metasearch

Javed Aslam

The Metasearch Problem

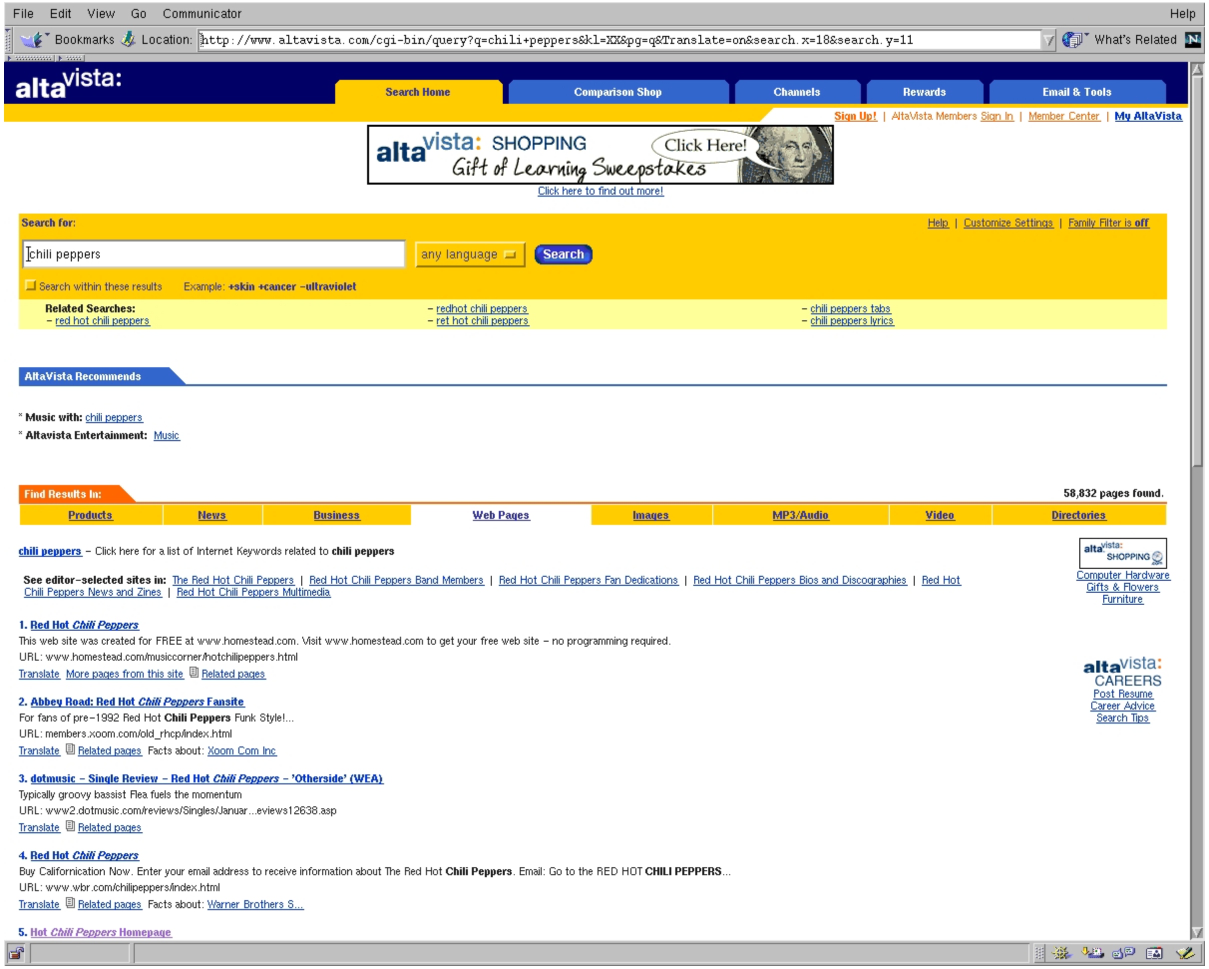
Search for: `chili peppers`



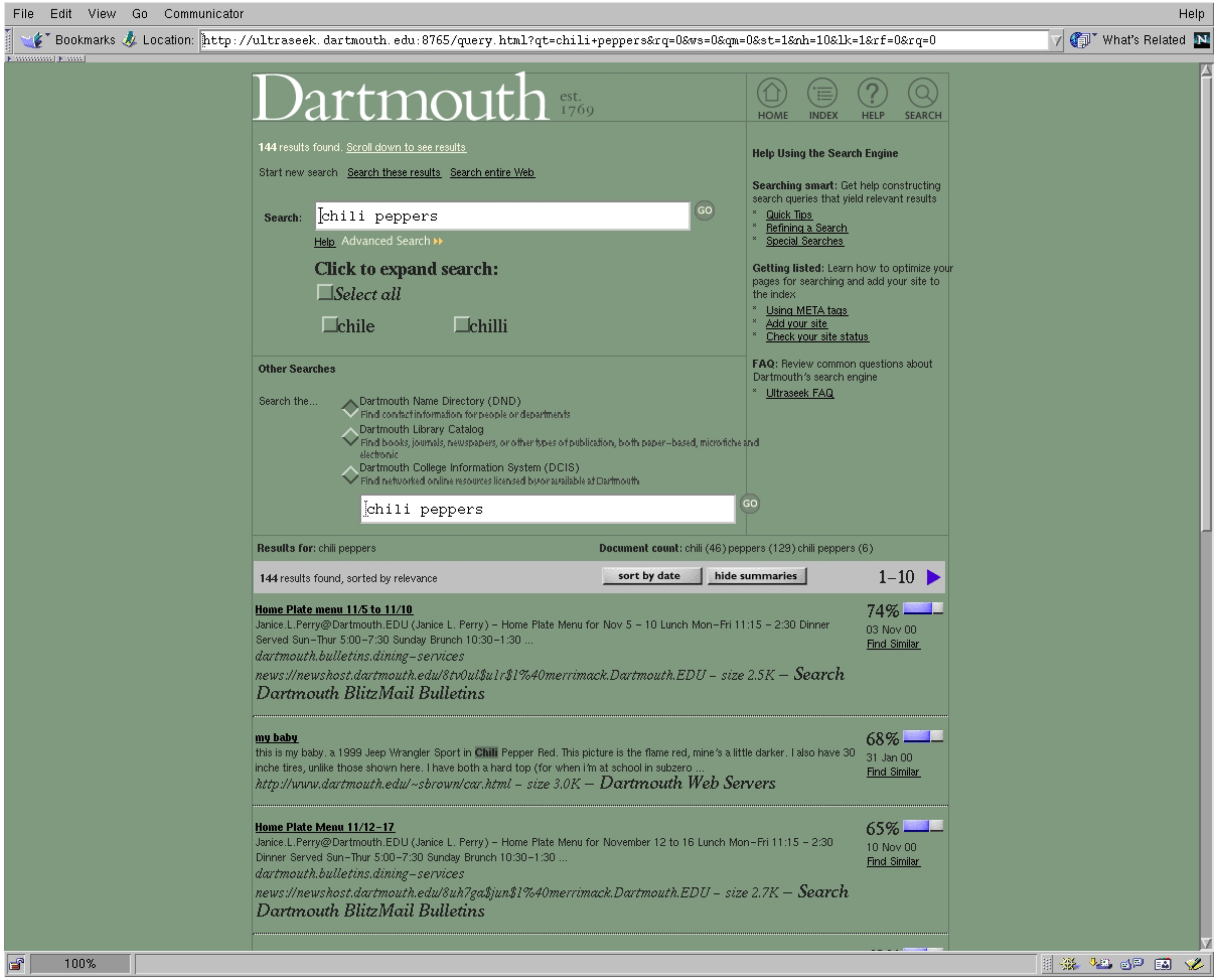
Search Engines

- Provide a ranked list of documents.
- May provide relevance scores.
- May have performance information.

Search Engine: Alta Vista



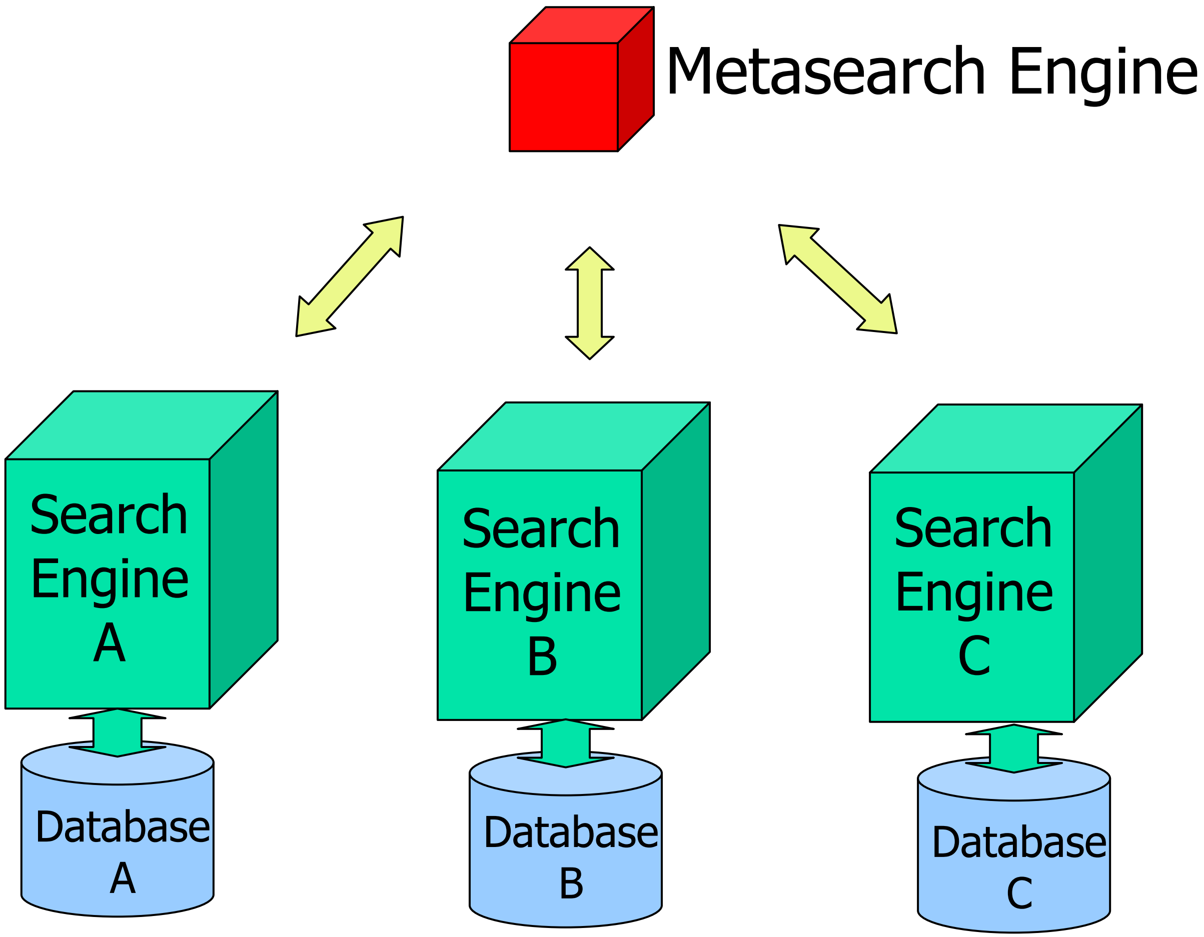
Search Engine: Ultraseek



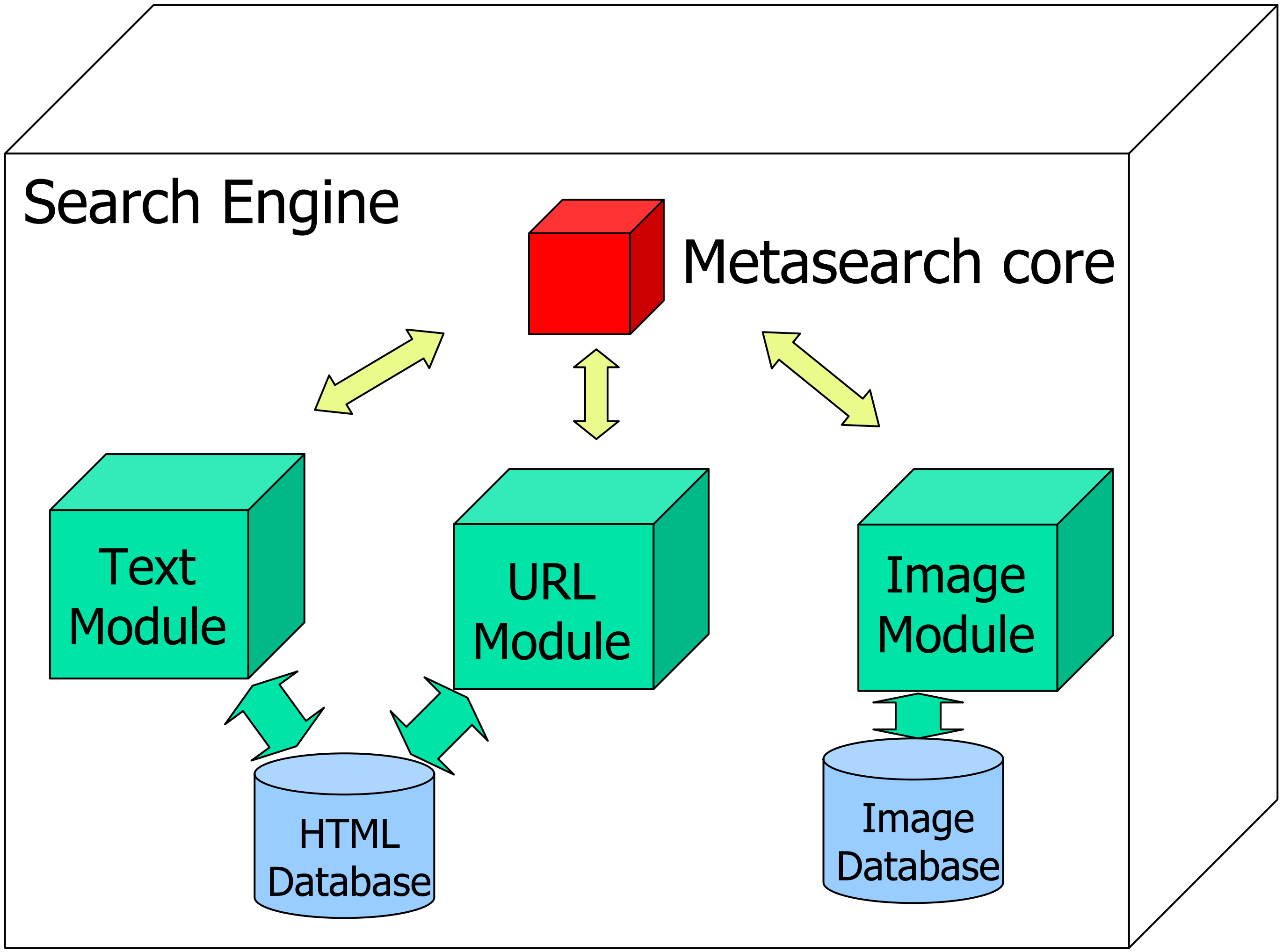
Search Engine: inq102 TREC3

```
Queryid (Num):          50
Total number of documents over all queries
    Retrieved:       50000
    Relevant:         9805
    Rel_ret:          7305
Interpolated Recall - Precision Averages:
    at 0.00         0.8992
    at 0.10         0.7514
    at 0.20         0.6584
    at 0.30         0.5724
    at 0.40         0.4982
    at 0.50         0.4272
    at 0.60         0.3521
    at 0.70         0.2915
    at 0.80         0.2173
    at 0.90         0.1336
    at 1.00         0.0115
Average precision (non-interpolated)
for all rel docs (averaged over queries)
                    0.4226
Precision:
  At    5 docs:     0.7440
  At   10 docs:     0.7220
  At   15 docs:     0.6867
  At   20 docs:     0.6740
  At   30 docs:     0.6267
  At  100 docs:     0.4902
  At  200 docs:     0.3848
  At  500 docs:     0.2401
  At 1000 docs:     0.1461
R-Precision (precision after R
(= num_rel for a query) docs retrieved):
    Exact:          0.4524
```

External Metasearch



Internal Metasearch



Outline

- ✓ Introduce problem
- Characterize problem
- Survey current techniques
- Describe new approaches
  - decision theory, social choice theory
  - experiments with TREC data
- Upper bounds for metasearch
- Future work

Classes of Metasearch Problems

	no training data	training data
ranks only	Borda, Condorcet, rCombMNZ	Bayes
relevance scores	CombMNZ	LC model

Outline

- ✓ Introduce problem
- ✓ Characterize problem
- Survey current techniques
- Describe new approaches
  - decision theory, social choice theory
  - experiments with TREC data
- Upper bounds for metasearch
- Future work

Classes of Metasearch Problems

	no training data	training data
ranks only	Borda, Condorcet, rCombMNZ	Bayes
relevance scores	CombMNZ	LC model

CombSUM [Fox, Shaw, Lee, et al.]

- Normalize scores: [0,1].
- For each doc:
  - sum relevance scores given to it by each system (use 0 if unretrieved).
- Rank documents by score.
- Variants: MIN, MAX, MED, ANZ, MNZ

CombMNZ [Fox, Shaw, Lee, et al.]

- Normalize scores: [0,1].
- For each doc:
  - sum relevance scores given to it by each system (use 0 if unretrieved), and
  - multiply by number of systems that retrieved it (MNZ).
- Rank documents by score.

How well do they perform?

- Need *performance metric*.
- Need *benchmark data*.

Metric: Average Precision

R	1/1	
N		
R	2/3	
N		
R	3/5	0.6917
N		
N		
R	4/8	

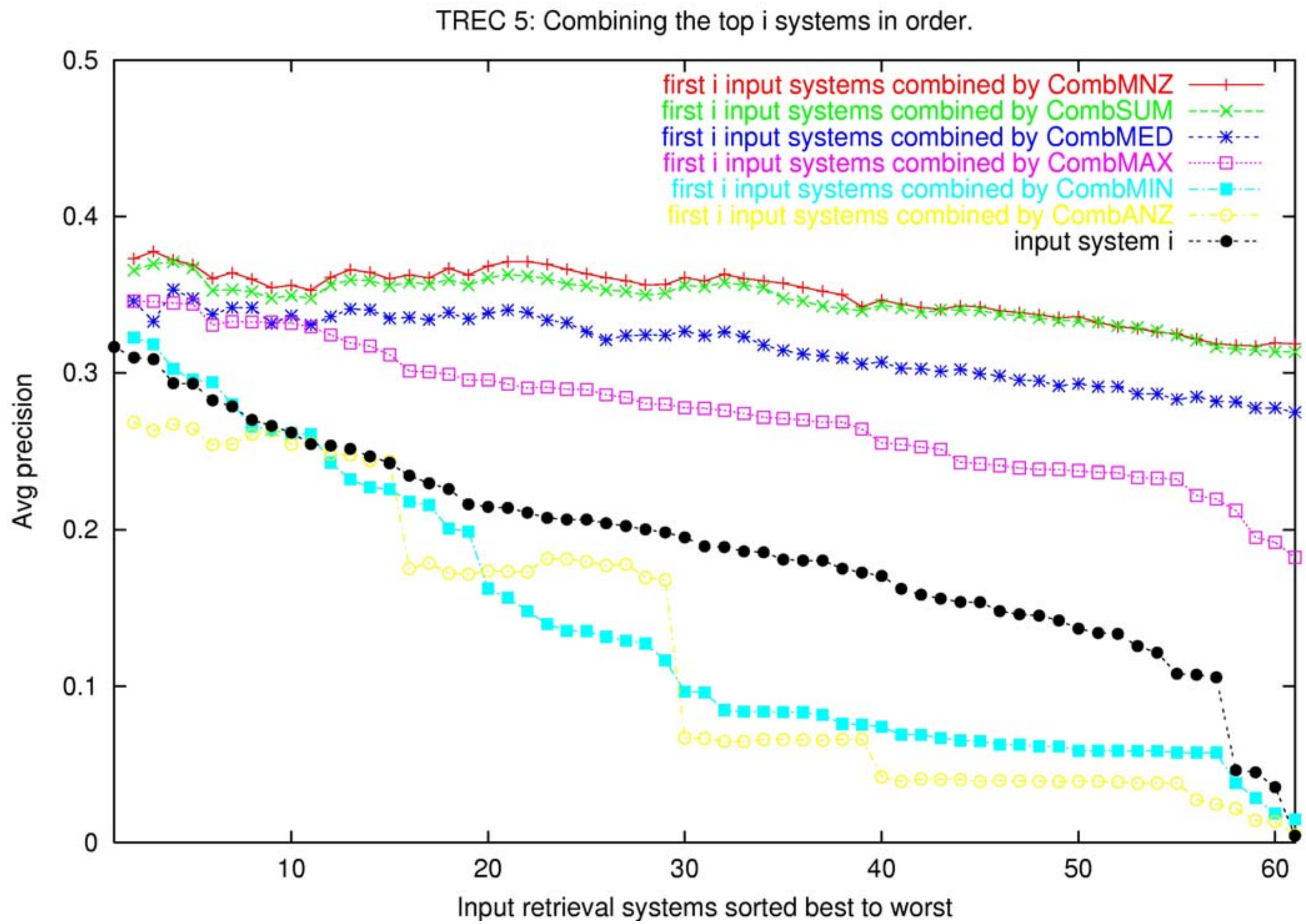
Benchmark Data: TREC

- Annual *Text Retrieval Conference.*
- Millions of documents (AP, NYT, etc.)
- 50 queries.
- Dozens of retrieval engines.
- Output lists available.
- Relevance judgments available.

Data Sets

Data set	Number systems	Number queries	Number of docs
TREC3	40	50	1000
TREC5	61	50	1000
Vogt	10	10	1000
TREC9	105	50	1000

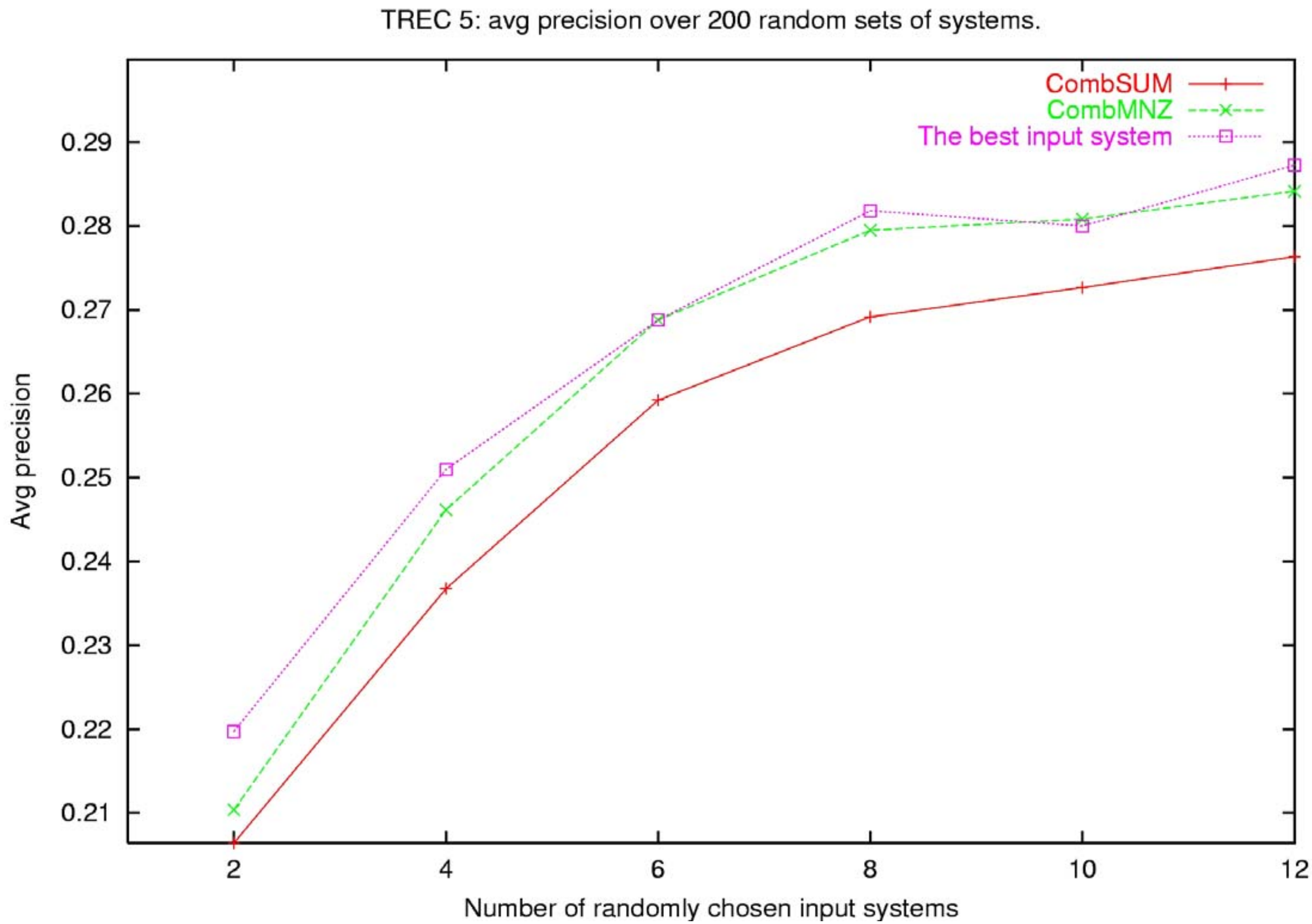
CombX on TREC5 Data



Experiments

- Randomly choose *n* input systems.
- For each query:
  - combine, trim, calculate avg precision.
- Calculate mean avg precision.
- Note best input system.
- Repeat (statistical significance).

CombMNZ on TREC5



Outline

- ✓ Introduce problem
- ✓ Characterize problem
- ✓ Survey current techniques
- Describe new approaches
  - decision theory, social choice theory
  - experiments with TREC data
- Upper bounds for metasearch
- Future work

New Approaches [Aslam, Montague]

- Analog to *decision theory*.
  - Requires only rank information.
  - Training required.
- Analog to *election strategies*.
  - Requires only rank information.
  - No training required.

Classes of Metasearch Problems

	no training data	training data
ranks only	Borda, Condorcet, rCombMNZ	Bayes
relevance scores	CombMNZ	LC model

Decision Theory

- Consider two alternative explanations for some observed data.
  - Medical example:
    - Perform a set of blood tests.
    - Does patient have disease or not?
- Optimal method for choosing among the explanations: *likelihood ratio test*.
  [Neyman-Pearson Lemma]

Metasearch via Decision Theory

- Metasearch analogy:
  - *Observed data* – document rank info over all systems.
  - *Hypotheses* – document is relevant or not.
- Ratio test: $$O_{rel} = \frac{\Pr[rel \mid r_1, r_2, \ldots, r_n]}{\Pr[irr \mid r_1, r_2, \ldots, r_n]}$$

Bayesian Analysis

$$P_{rel} = \Pr[rel \mid r_1, r_2, \ldots, r_n]$$

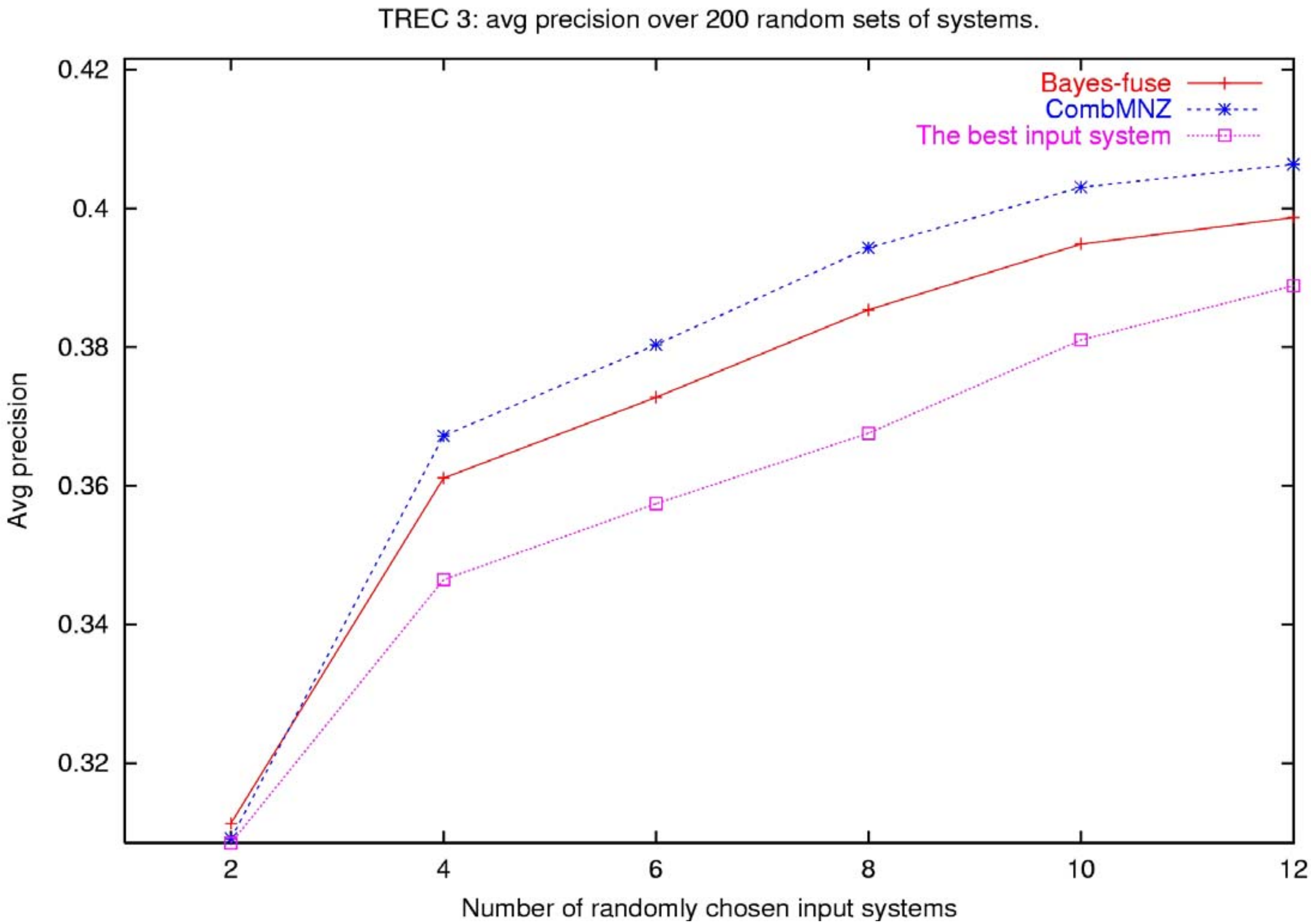
$$P_{rel} = \frac{\Pr[r_1, r_2, \ldots, r_n \mid rel] \cdot \Pr[rel]}{\Pr[r_1, r_2, \ldots, r_n]}$$

$$O_{rel} = \frac{\Pr[r_1, r_2, \ldots, r_n \mid rel] \cdot \Pr[rel]}{\Pr[r_1, r_2, \ldots, r_n \mid irr] \cdot \Pr[irr]}$$

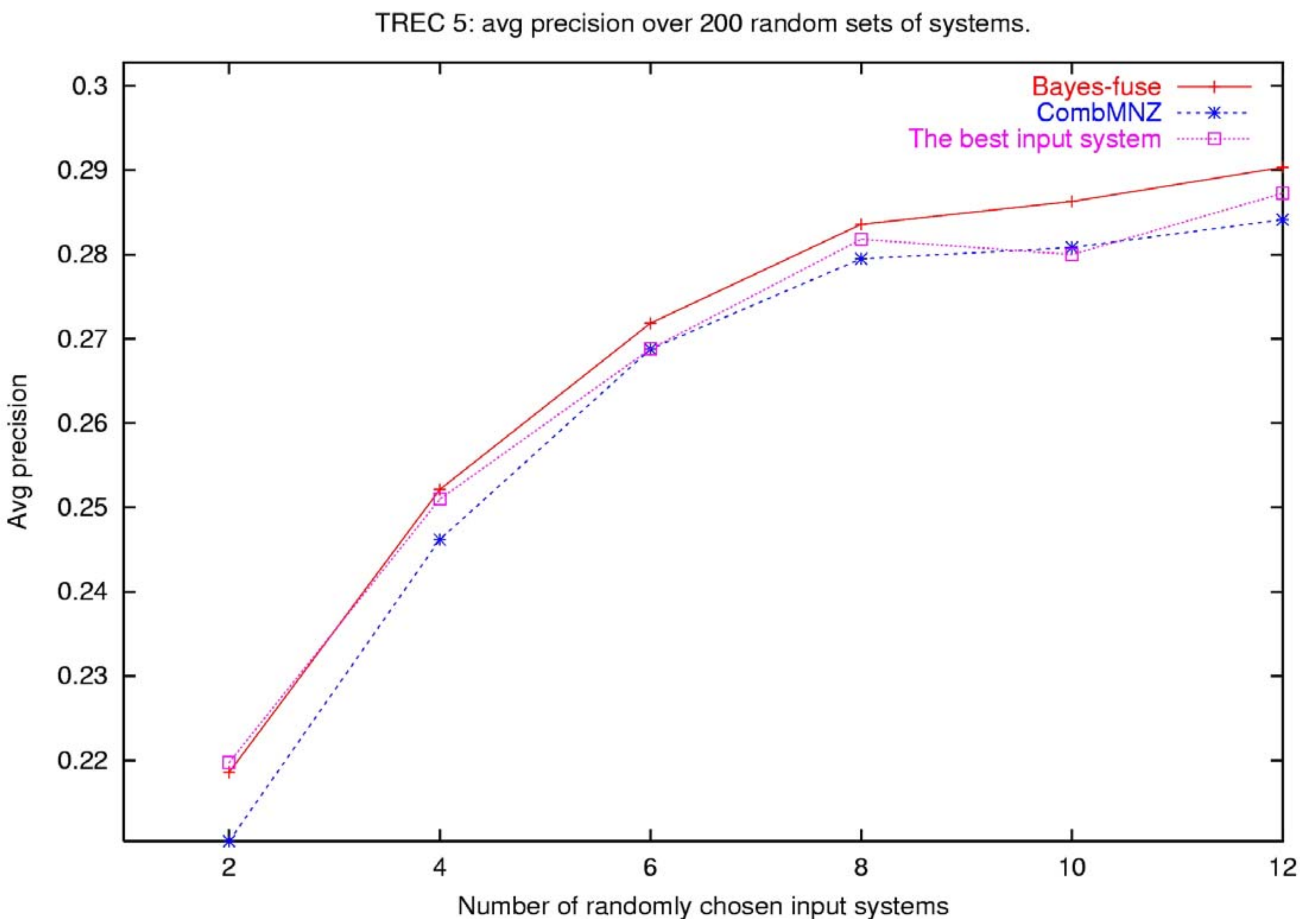
$$O_{rel} \cong \frac{\Pr[rel] \cdot \prod_i \Pr[r_i \mid rel]}{\Pr[irr] \cdot \prod_i \Pr[r_i \mid irr]}$$

$$LO_{rel} \sim \sum_i \log \frac{\Pr[r_i \mid rel]}{\Pr[r_i \mid irr]}$$

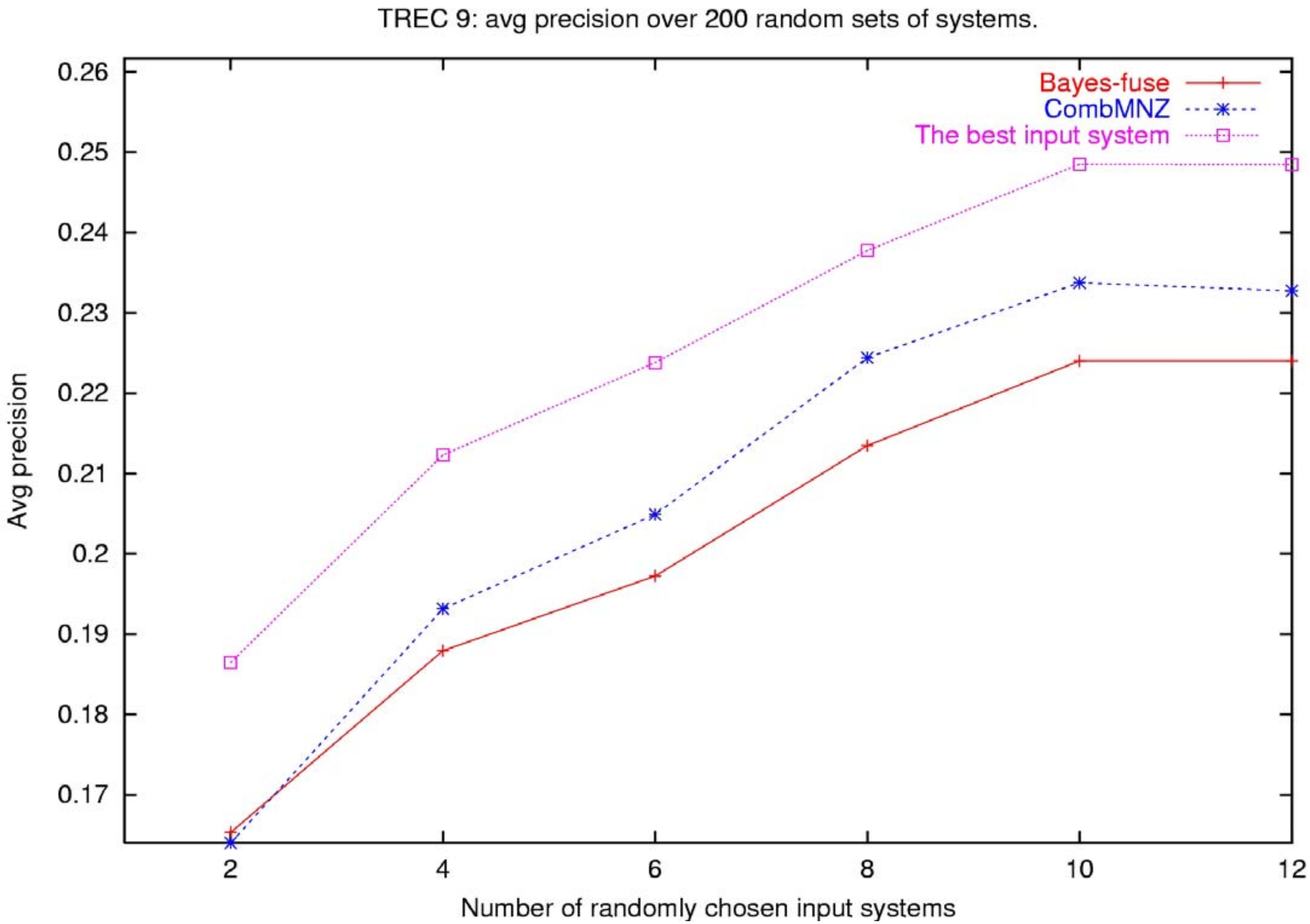
Bayes on TREC3



Bayes on TREC5



Bayes on TREC9



Beautiful theory, but…

In theory, there is no difference between theory and practice; in practice, there is.

–variously: Chuck Reid, Yogi Berra

Issue: *independence assumption…*

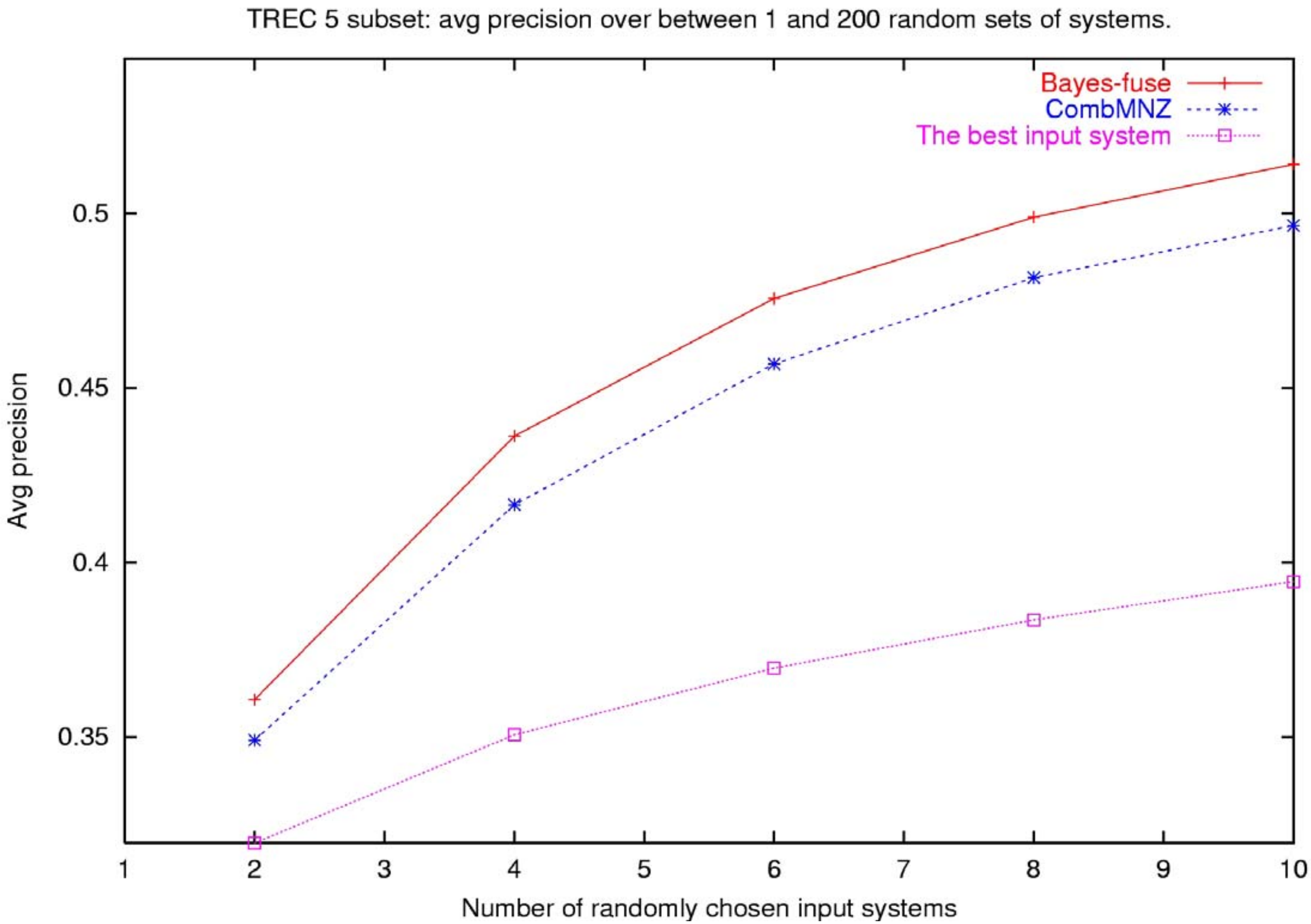
Naïve-Bayes Assumption

$$O_{rel} = \frac{\Pr[r_1, r_2, \ldots, r_n \mid rel] \cdot \Pr[rel]}{\Pr[r_1, r_2, \ldots, r_n \mid irr] \cdot \Pr[irr]}$$

$$\Downarrow$$

$$O_{rel} \cong \frac{\Pr[rel] \cdot \prod_i \Pr[r_i \mid rel]}{\Pr[irr] \cdot \prod_i \Pr[r_i \mid irr]}$$

Bayes on Vogt Data



New Approaches [Aslam, Montague]

- Analog to *decision theory*.
  - Requires only rank information.
  - Training required.
- Analog to *election strategies*.
  - Requires only rank information.
  - No training required.

Classes of Metasearch Problems

	no training data	training data
ranks only	Borda, Condorcet, rCombMNZ	Bayes
relevance scores	CombMNZ	LC model

Election Strategies

- Plurality vote.
- Approval vote.
- Run-off.
- Preferential rankings:
  - instant run-off,
  - Borda count (positional),
  - Condorcet method (head-to-head).

Metasearch Analogy

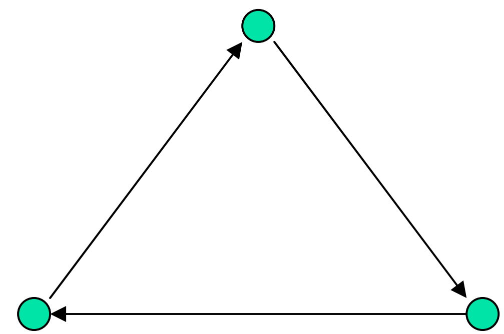
- Documents are *candidates*.
- Systems are *voters* expressing preferential rankings among candidates.

Condorcet Voting

- Each ballot ranks all candidates.
- Simulate head-to-head run-off between each pair of candidates.
- Condorcet winner: candidate that beats all other candidates, head-to-head.

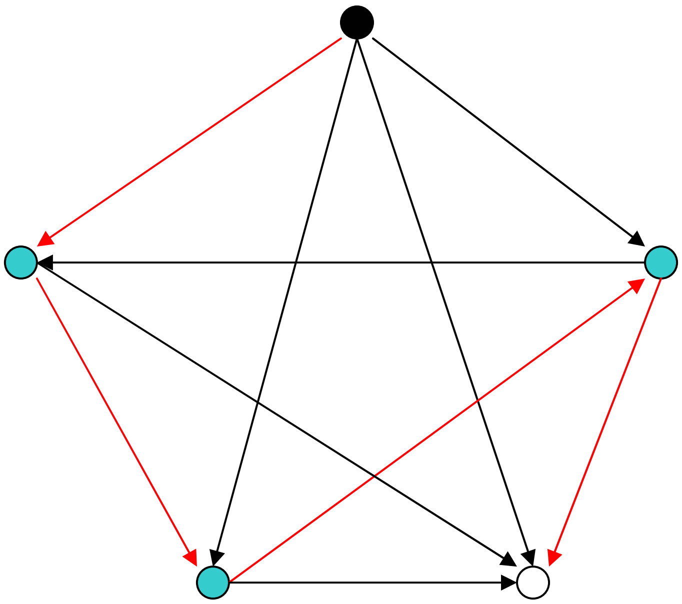
Condorcet Paradox

- Voter 1: A, B, C
- Voter 2: B, C, A
- Voter 3: C, A, B



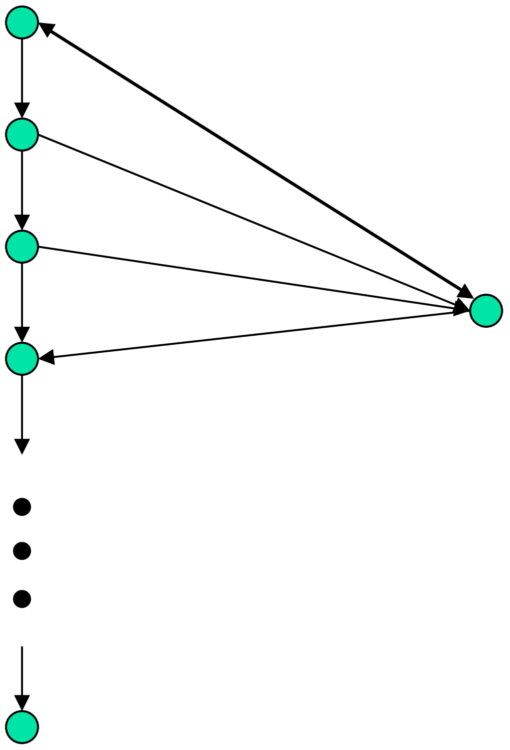
- Cyclic preferences: cycle in Condorcet graph.
- Condorcet consistent path: Hamiltonian.
- For metasearch: any CC path will do.

Condorcet Consistent Path



Hamiltonian Path Proof

Inductive Step:

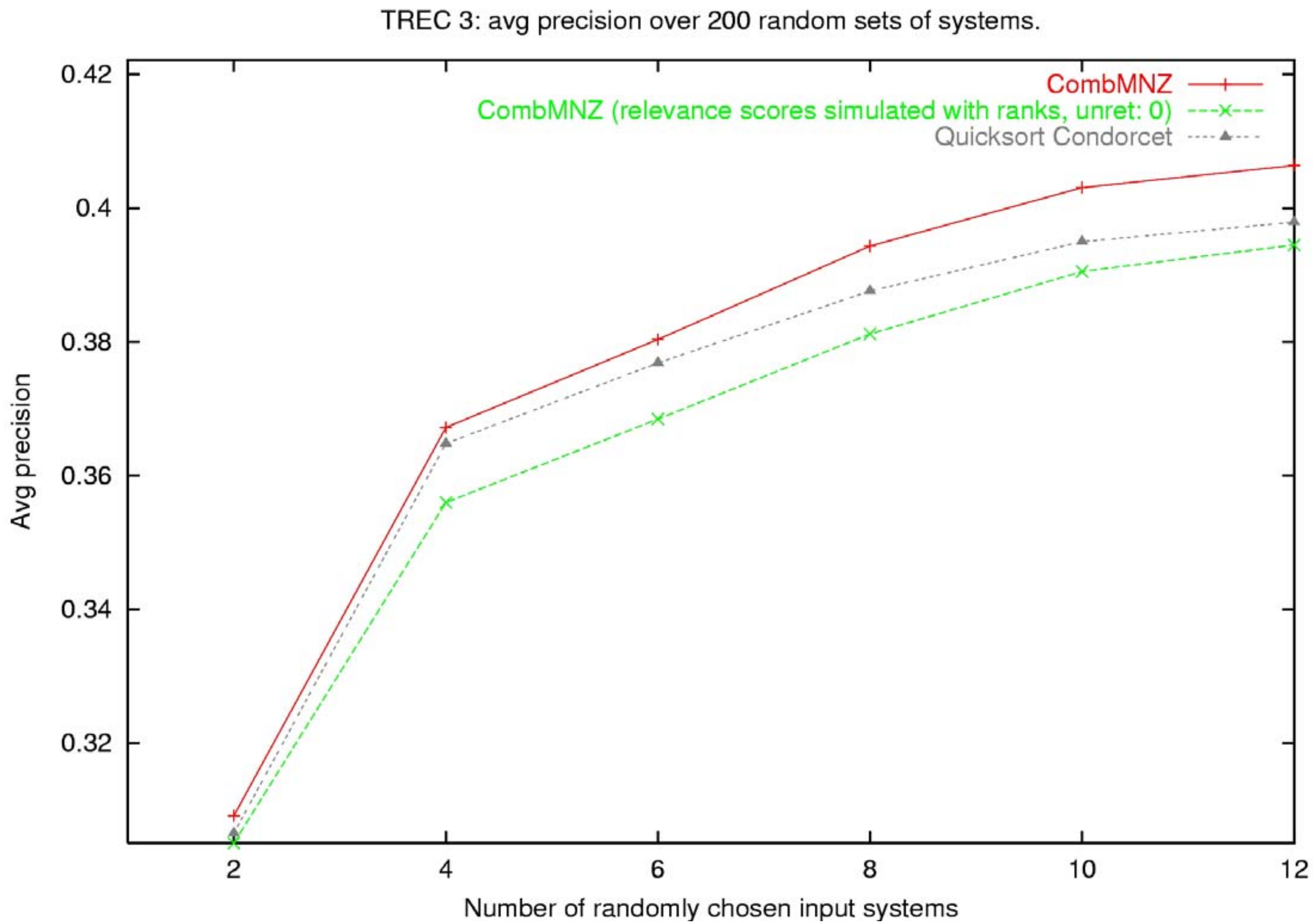


Base Case:

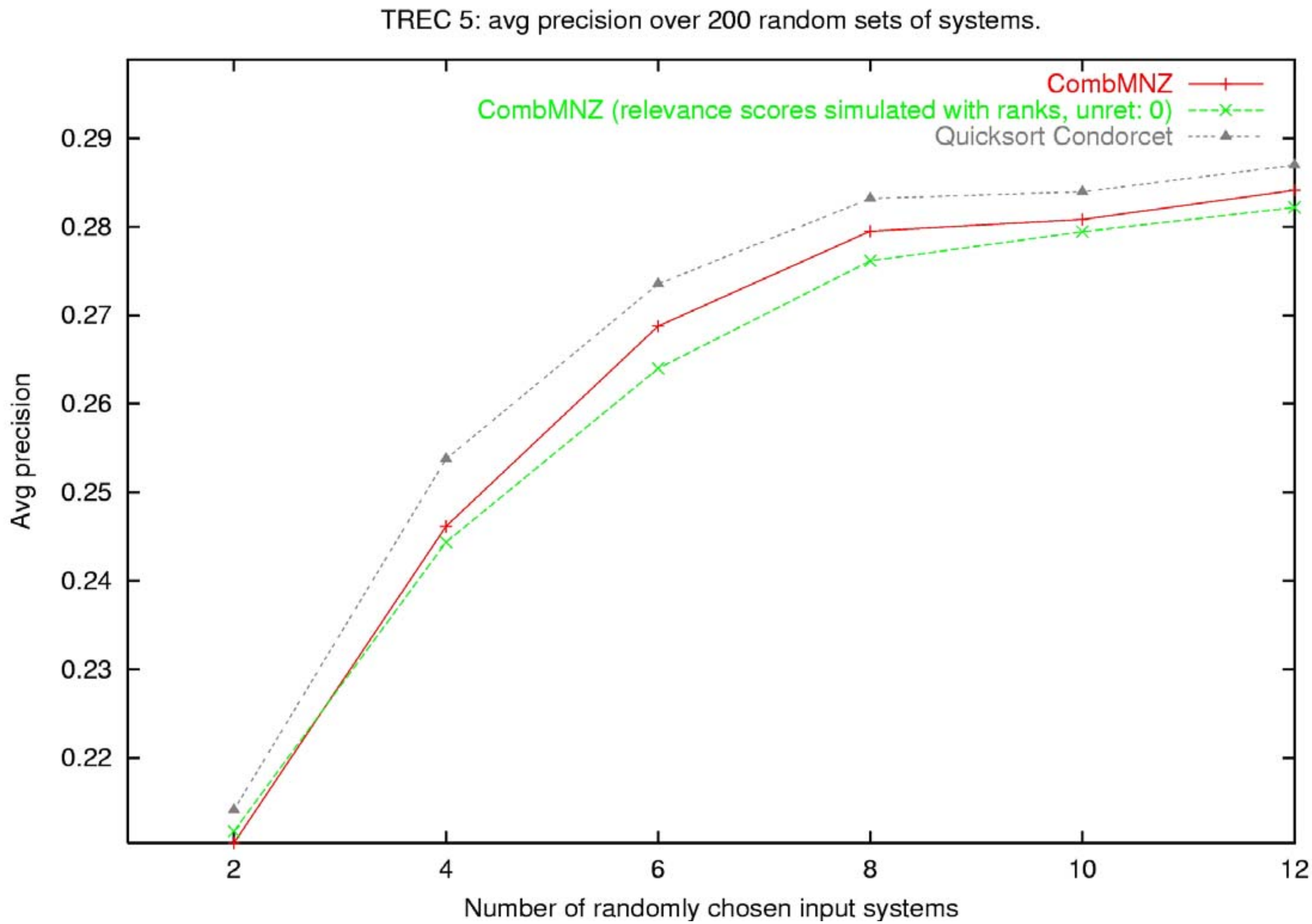
Condorcet-fuse: Sorting

- Insertion-sort suggested by proof.
- Quicksort too; $O(n \log n)$ comparisons.
  - n documents.
- Each comparison: $O(m)$.
  - m input systems.
- Total: $O(m\ n \log n)$.
- *Need not compute entire graph.*

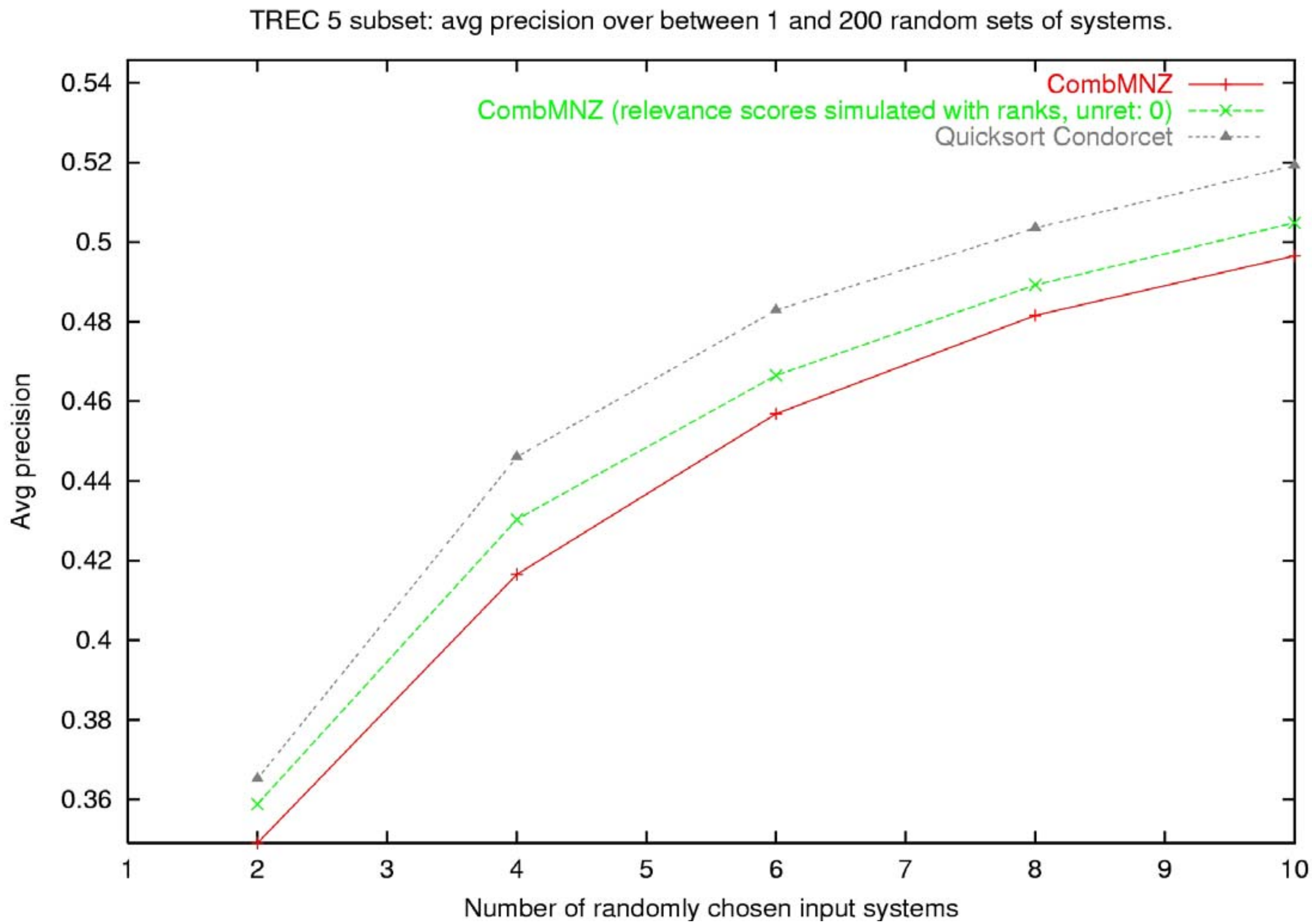
Condorcet-fuse on TREC3



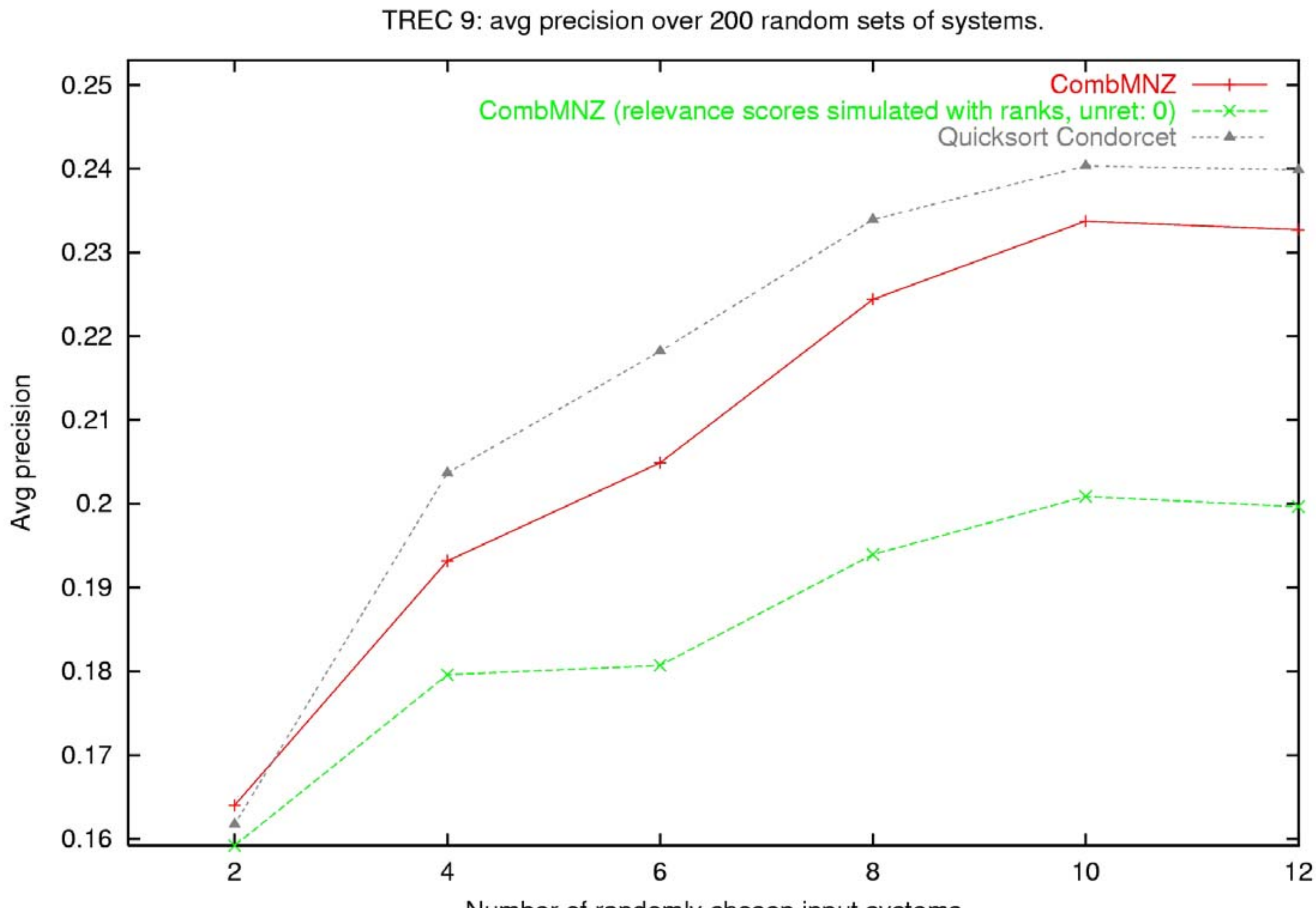
Condorcet-fuse on TREC5



Condorcet-fuse on Vogt



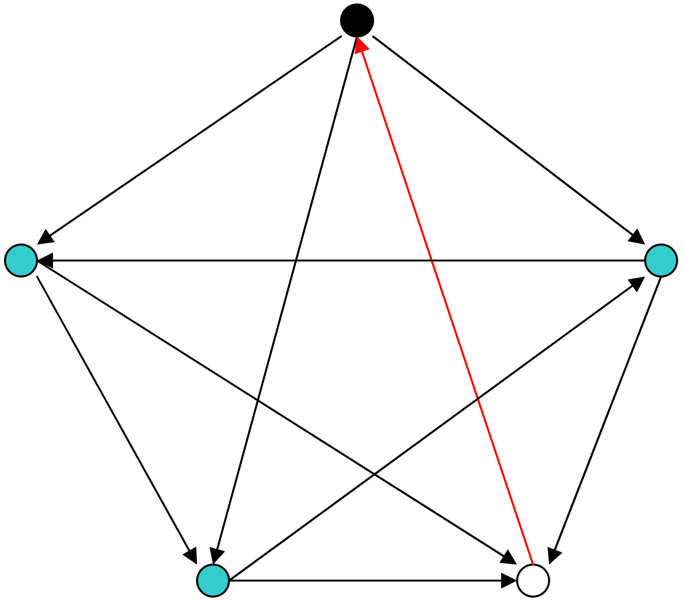
Condorcet-fuse on TREC9



Breaking Cycles

SCCs are properly ordered.

How are ties within an SCC broken? (Quicksort)



Outline

- ✓ Introduce problem
- ✓ Characterize problem
- ✓ Survey current techniques
- ✓ Describe new approaches
  - ✓ decision theory, social choice theory
  - ✓ experiments with TREC data
- Upper bounds for metasearch
- Future work

Upper Bounds on Metasearch

- How good can metasearch be?
- Are there fundamental limits that methods are approaching?
- Need an analog to running time lower bounds…

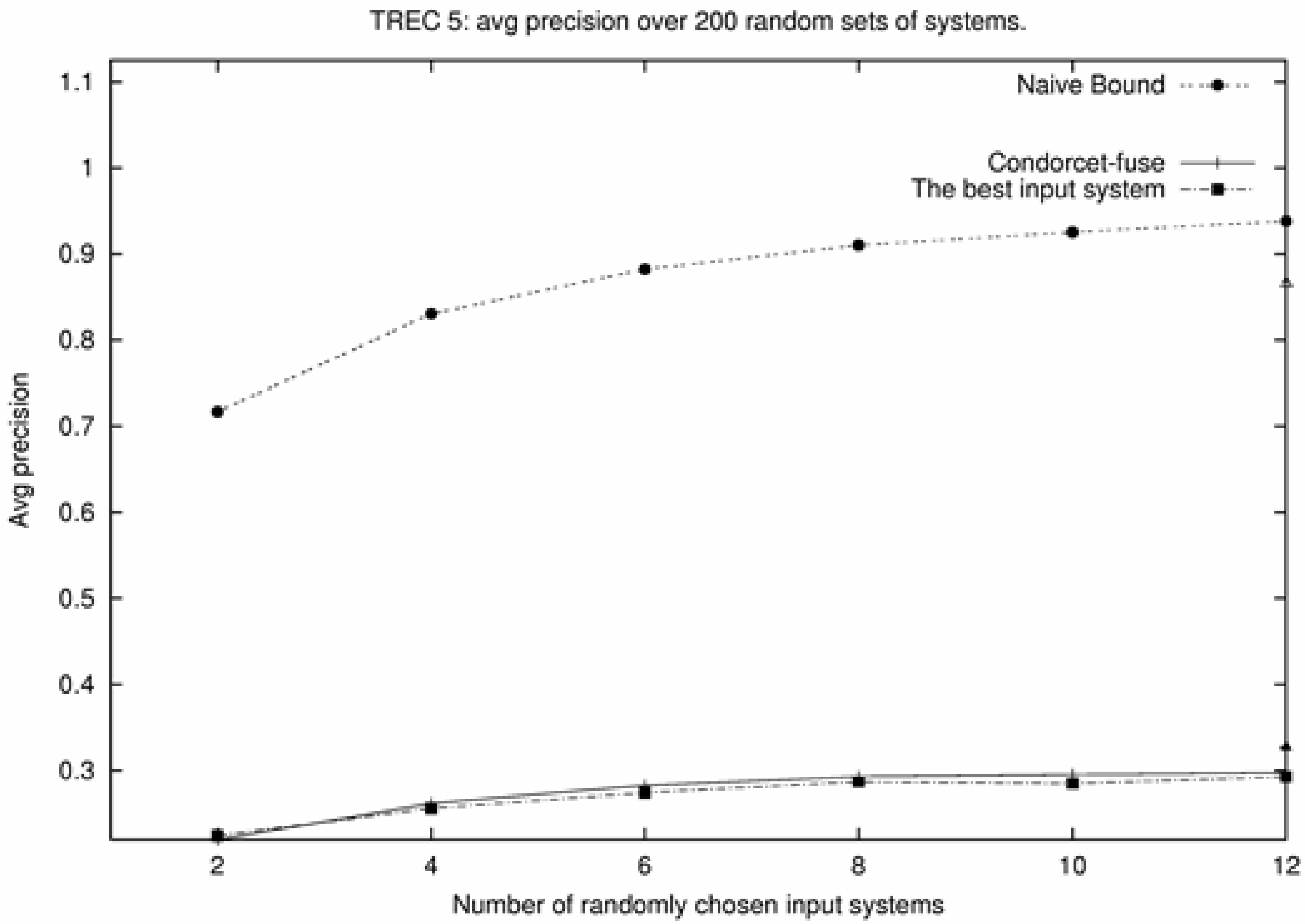
Upper Bounds on Metasearch

- Constrained oracle model:
  - omniscient metasearch oracle,
  - constraints placed on oracle that any reasonable metasearch technique must obey.
- What are "reasonable" constraints?

Naïve Constraint

- *Naïve* constraint:
  - Oracle may only return docs from underlying lists.
  - Oracle may return these docs in any order.
  - Omniscient oracle will return relevants docs above irrelevant docs.

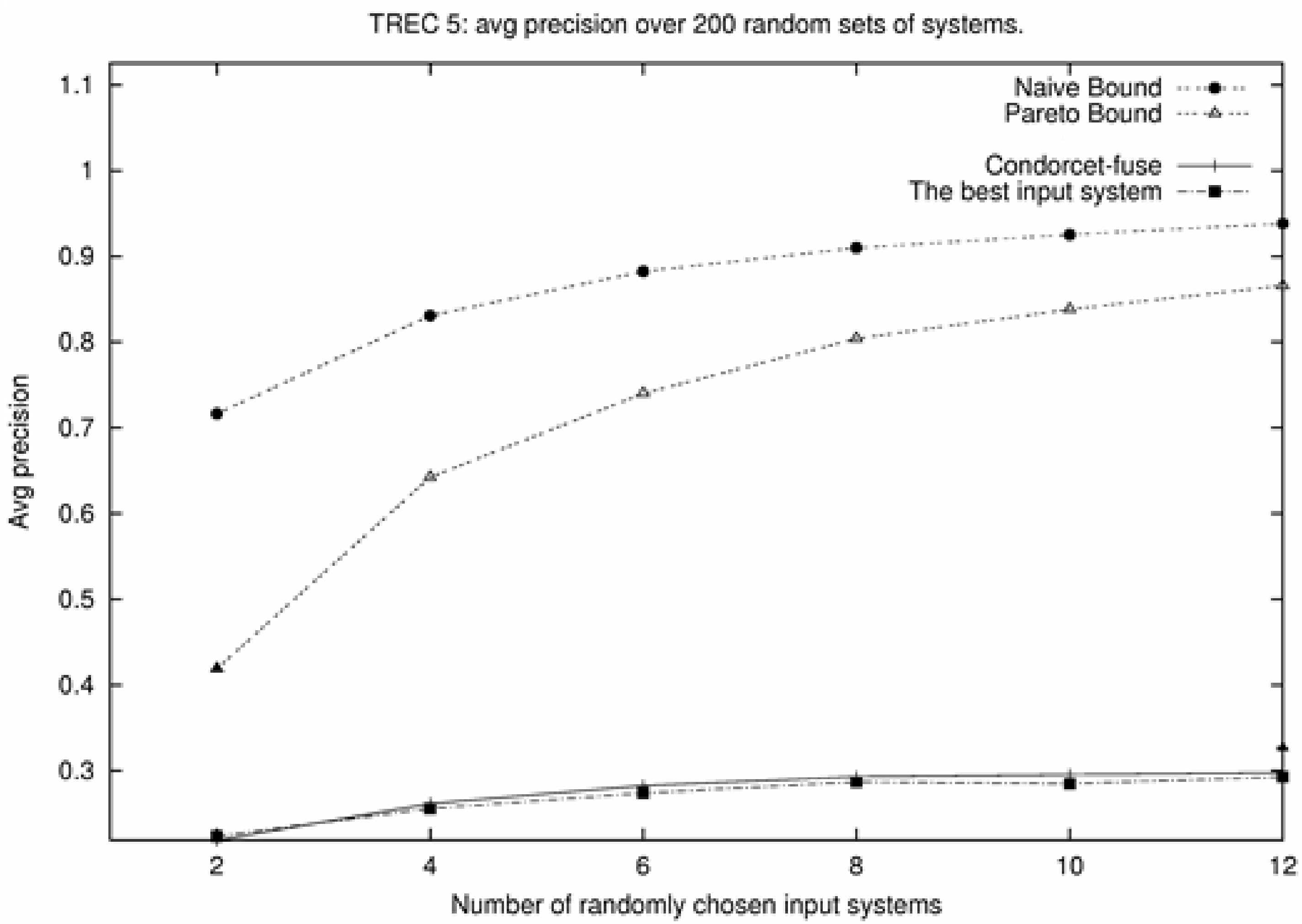
TREC5: Naïve Bound



Pareto Constraint

- *Pareto* constraint:
  - Oracle may only return docs from underlying lists.
  - Oracle must respect *unanimous* will of underlying systems.
  - Omniscient oracle will return relevants docs above irrelevant docs, subject to the above constraint.

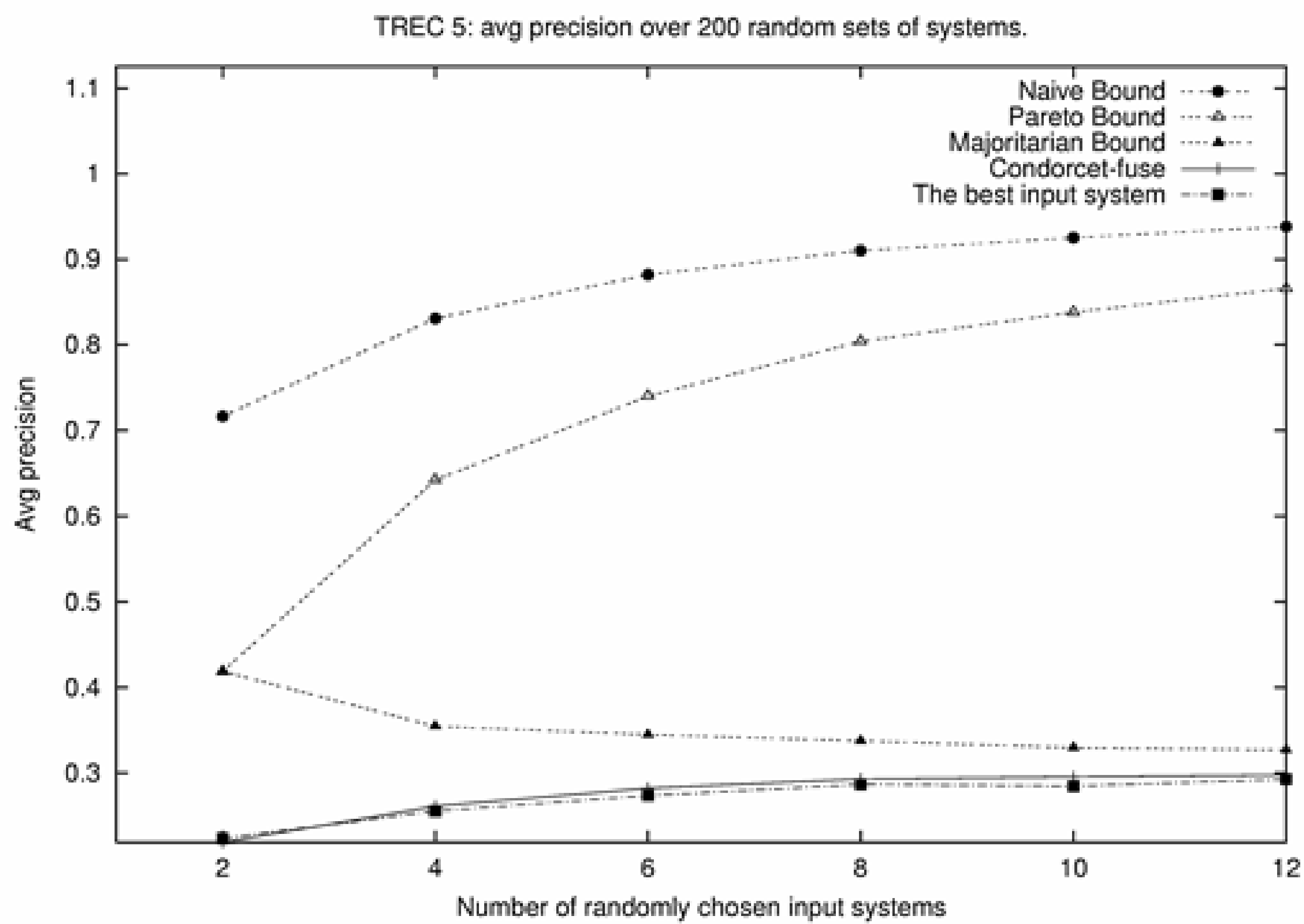
TREC5: Pareto Bound



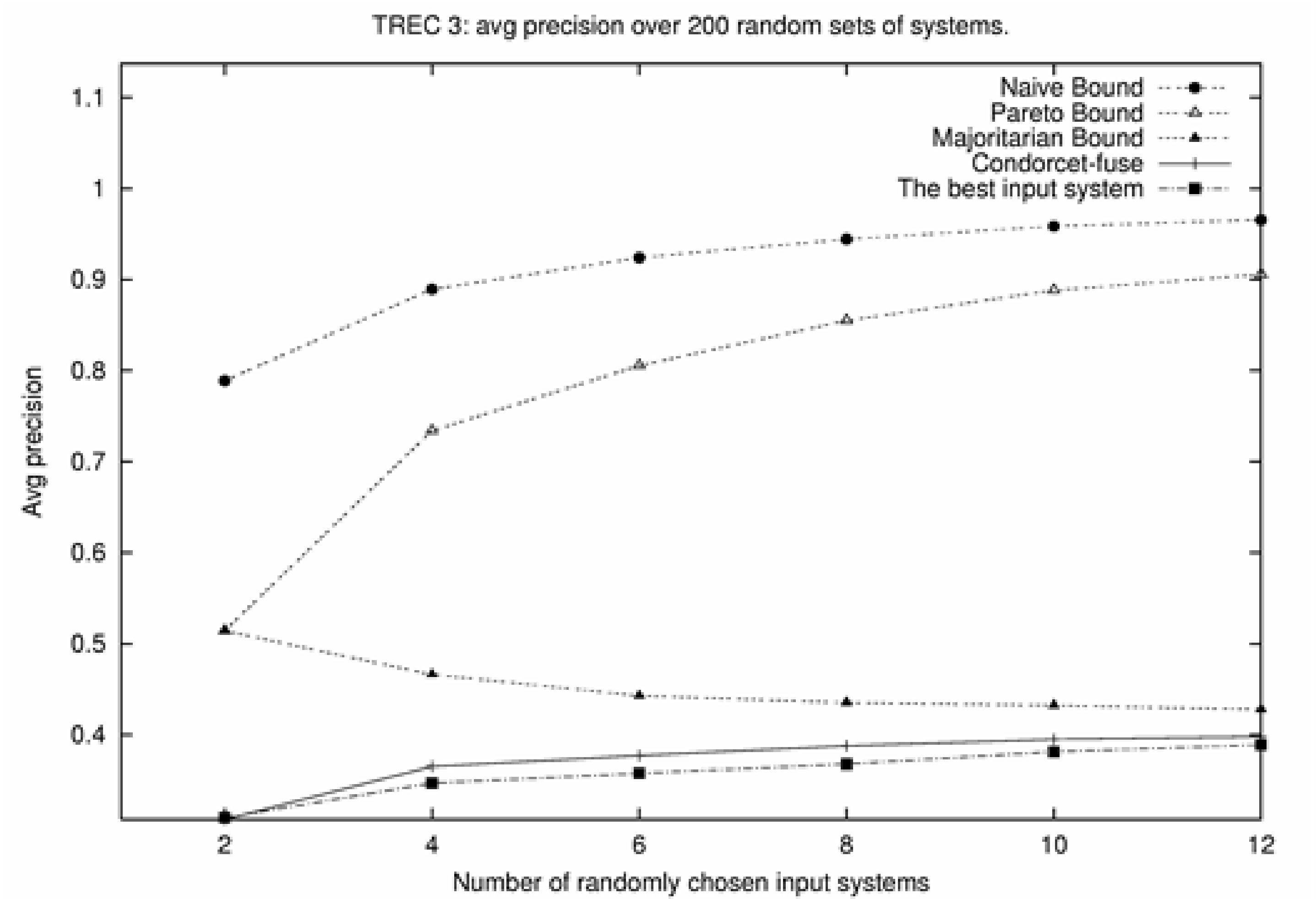
Majoritarian Constraint

- *Majoritarian* constraint:
  - Oracle may only return docs from underlying lists.
  - Oracle must respect *majority* will of underlying systems.
  - Omniscient oracle will return relevant docs above irrelevant docs and break cycles optimally, subject to the above constraint.

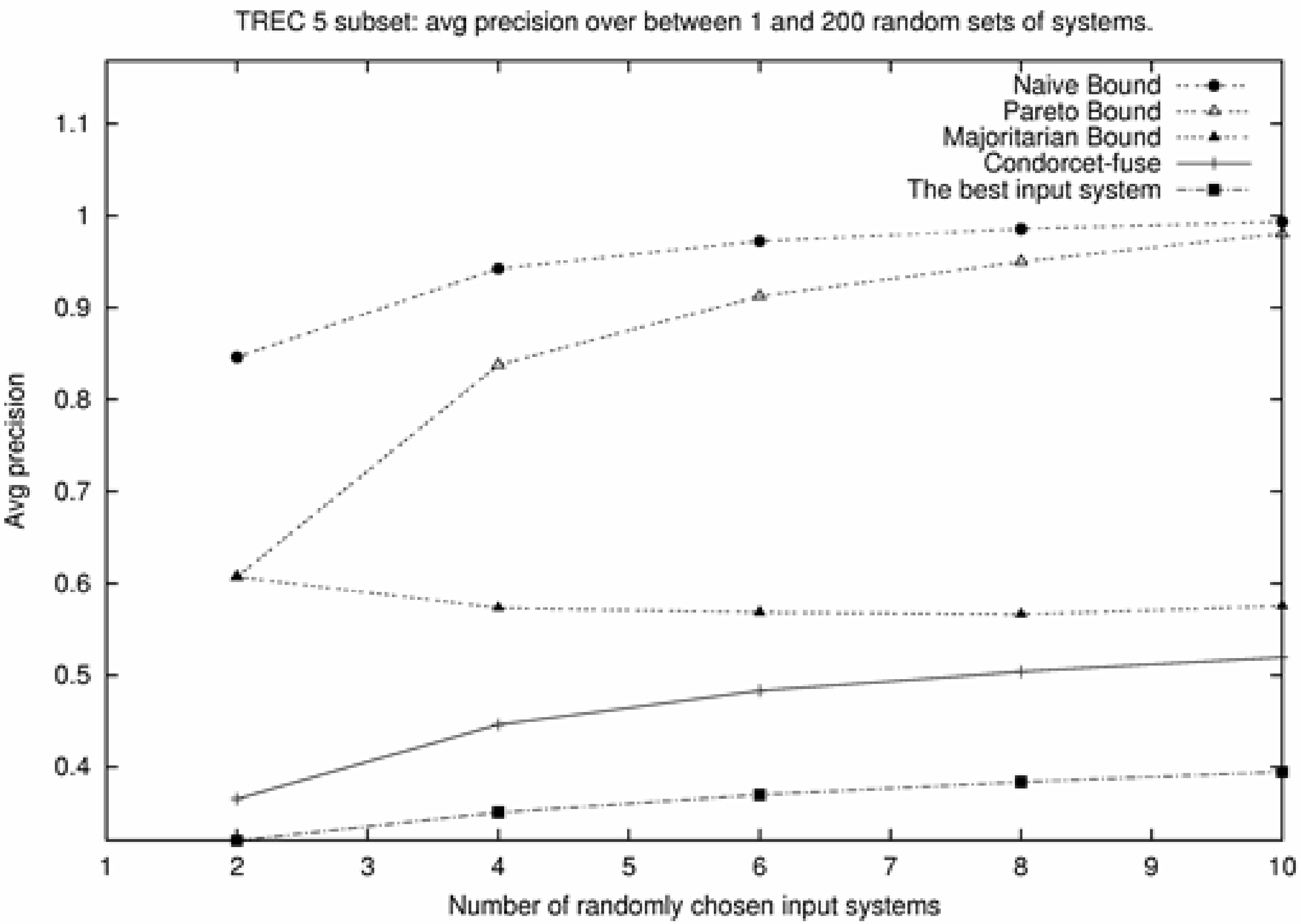
TREC5: Majoritarian Bound



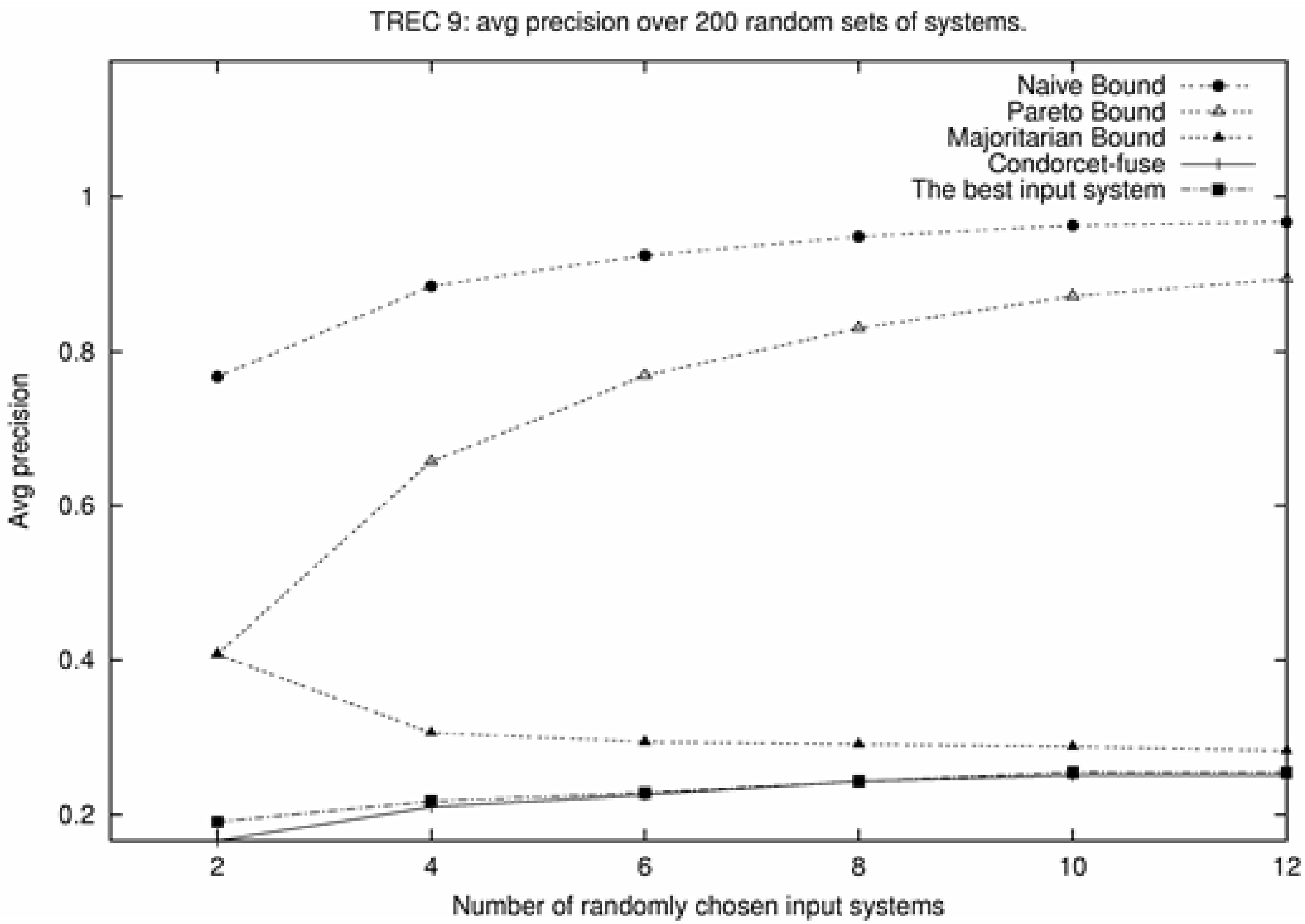
Upper Bounds: TREC3



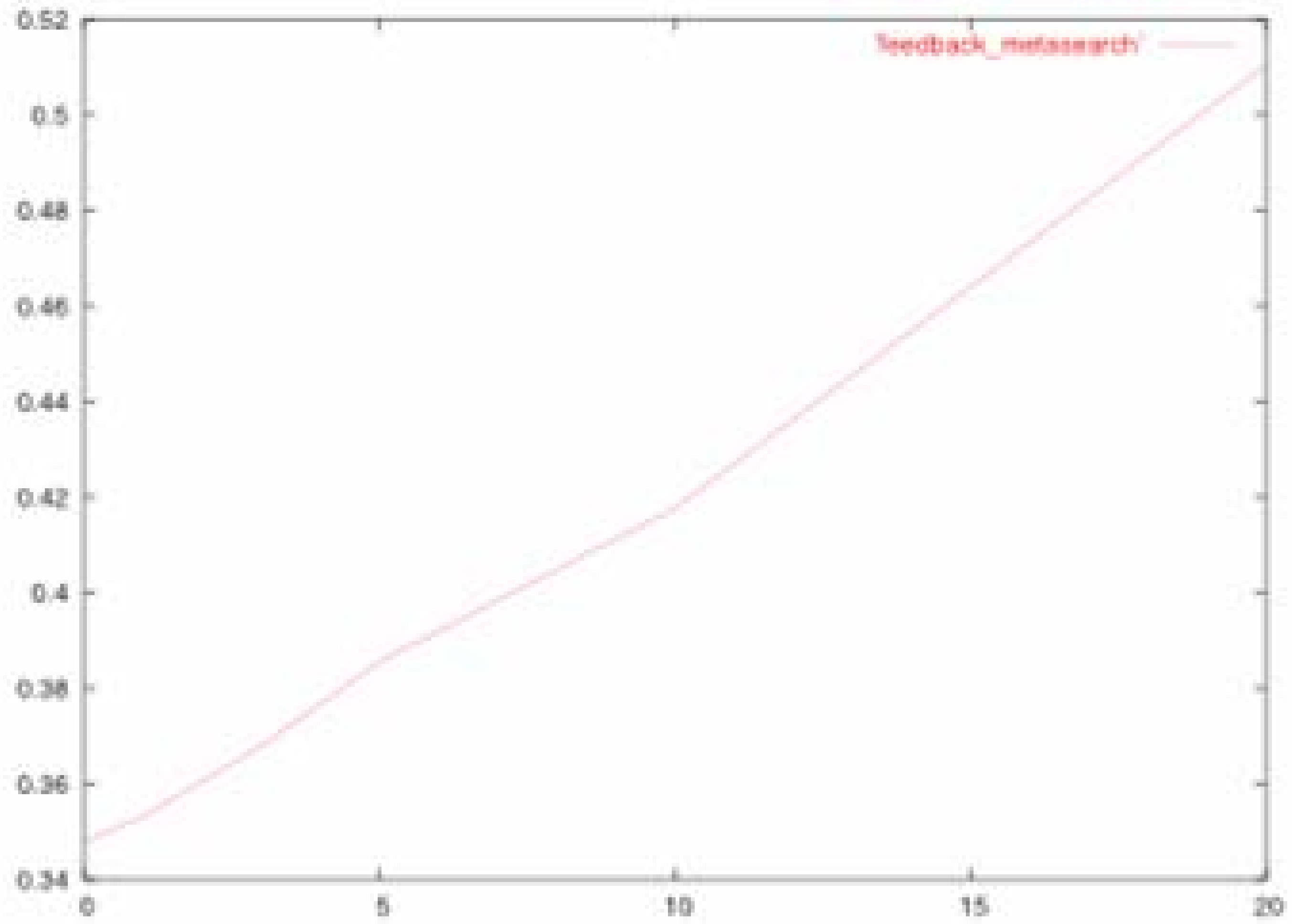
Upper Bounds: Vogt



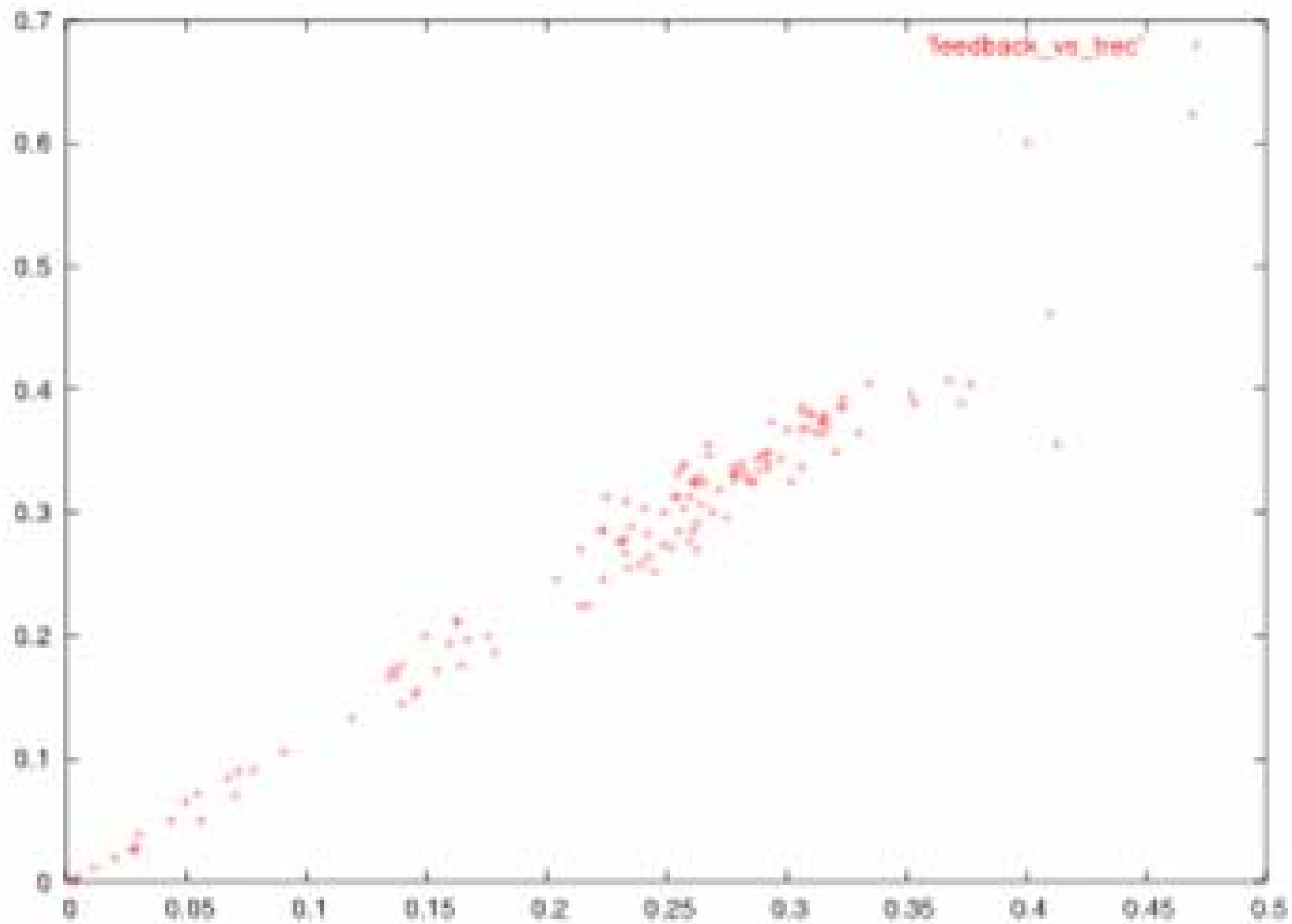
Upper Bounds: TREC9



TREC8: Avg Prec vs Feedback



TREC8: System Assessments vs TREC

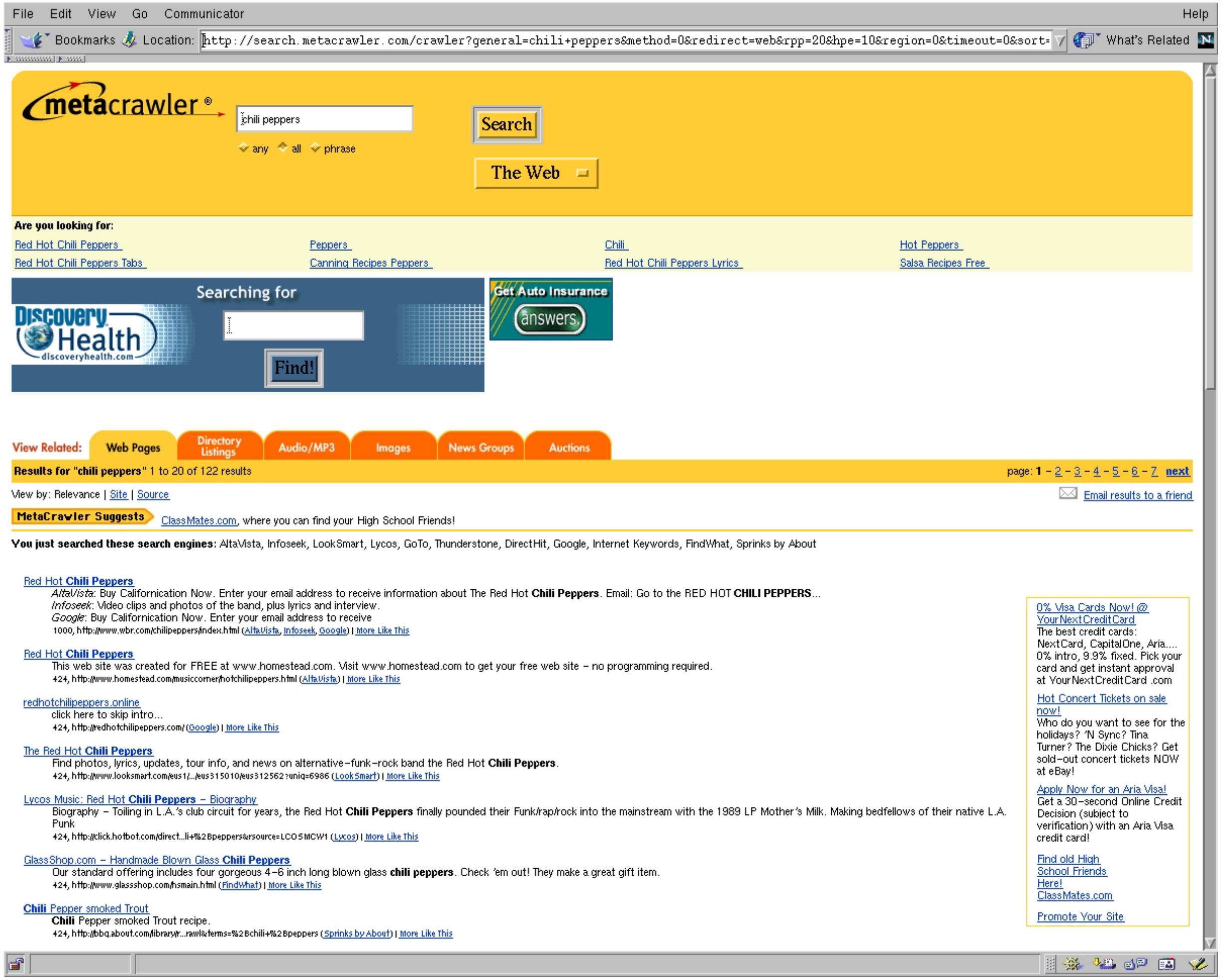


Metasearch Engines

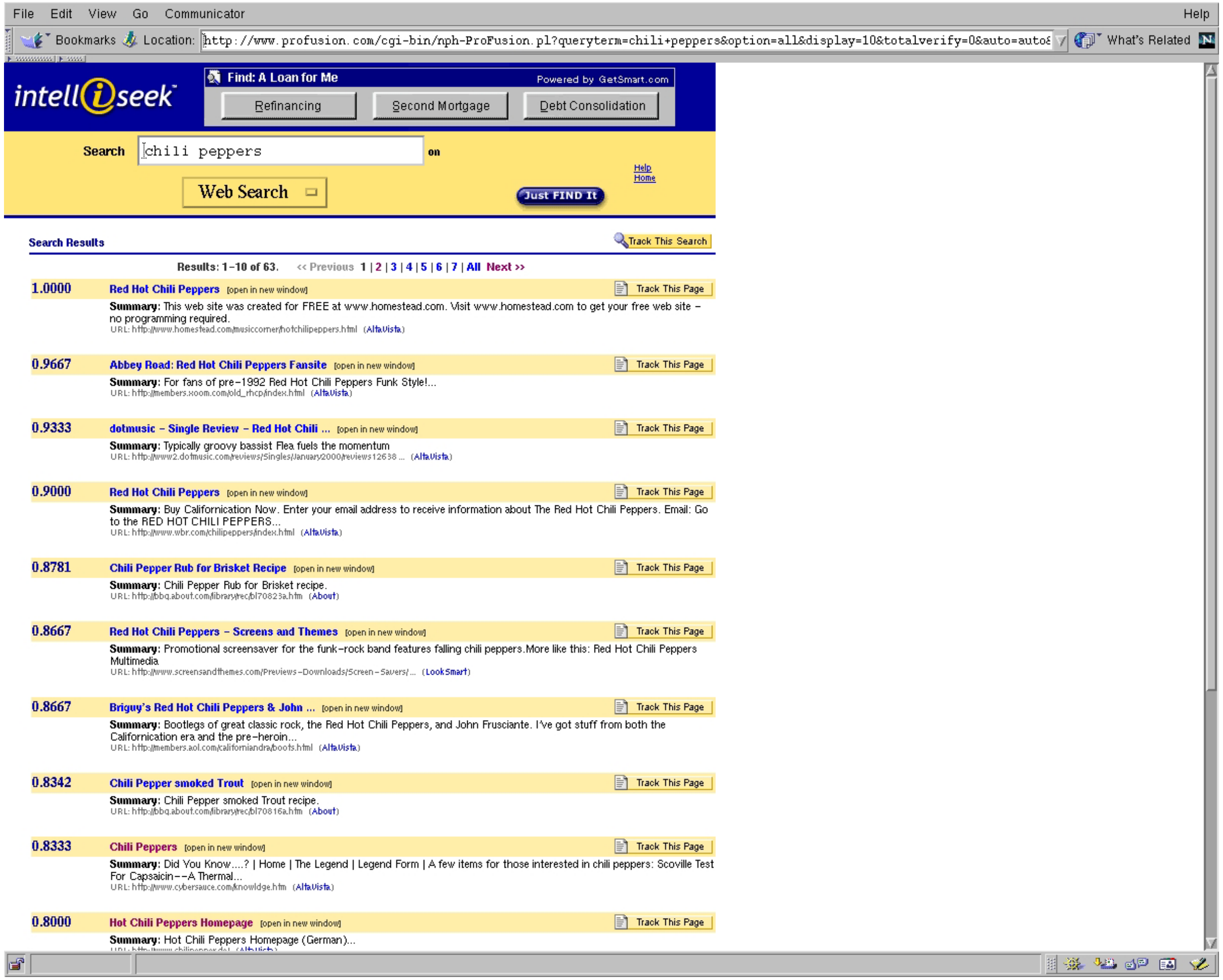
- Query multiple search engines.
- May or may not combine results.

Metasearch: Dogpile

Metasearch: Metacrawler



Metasearch: Profusion



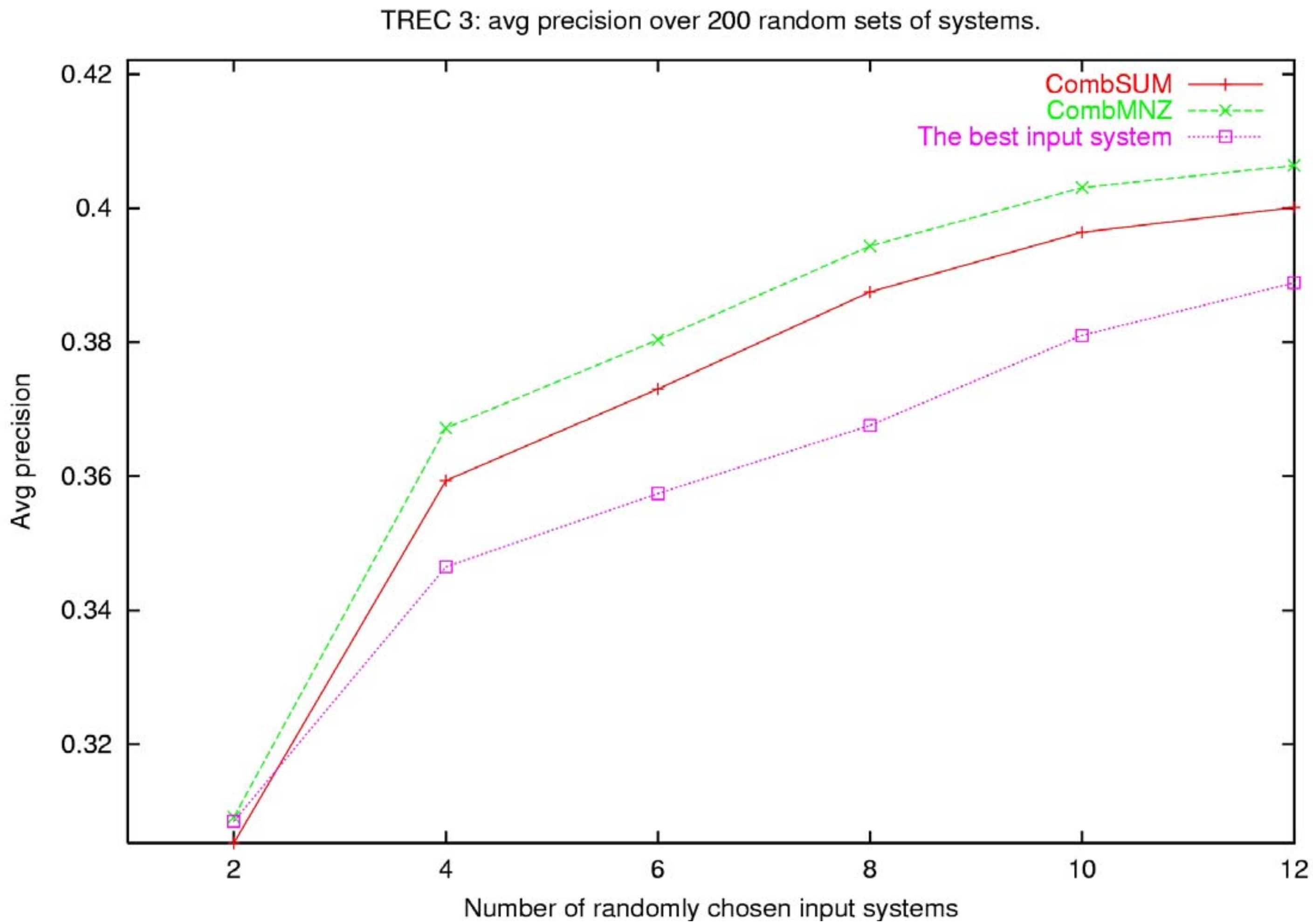
Characterizing Metasearch

- Three axes:
  - common *vs*. disjoint database,
  - relevance scores *vs*. ranks,
  - training data *vs*. no training data.

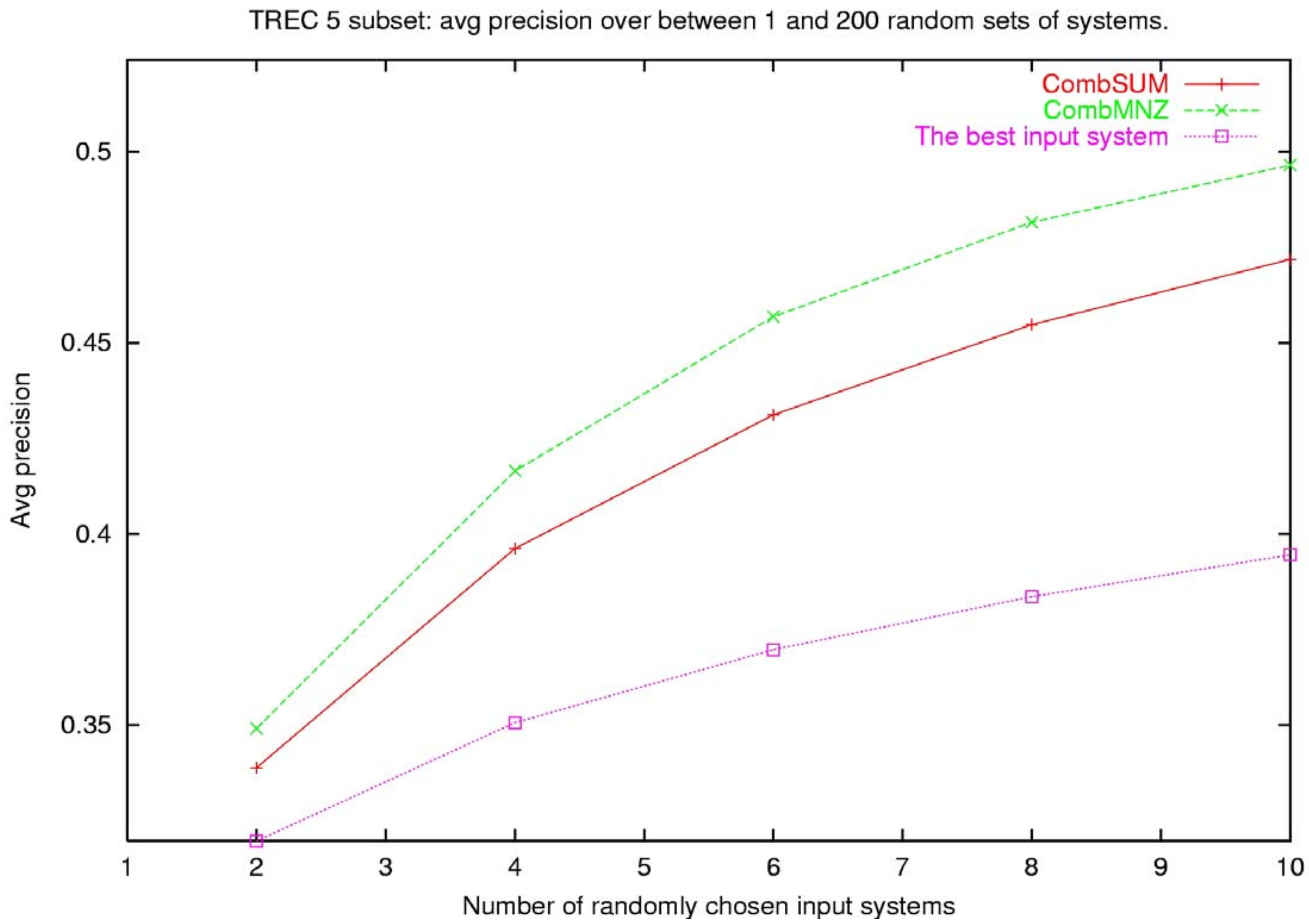
Axis 1: DB Overlap

- High overlap
  - data fusion.
- Low overlap
  - collection fusion (distributed retrieval).
- *Very different techniques for each…*
- This work: data fusion.

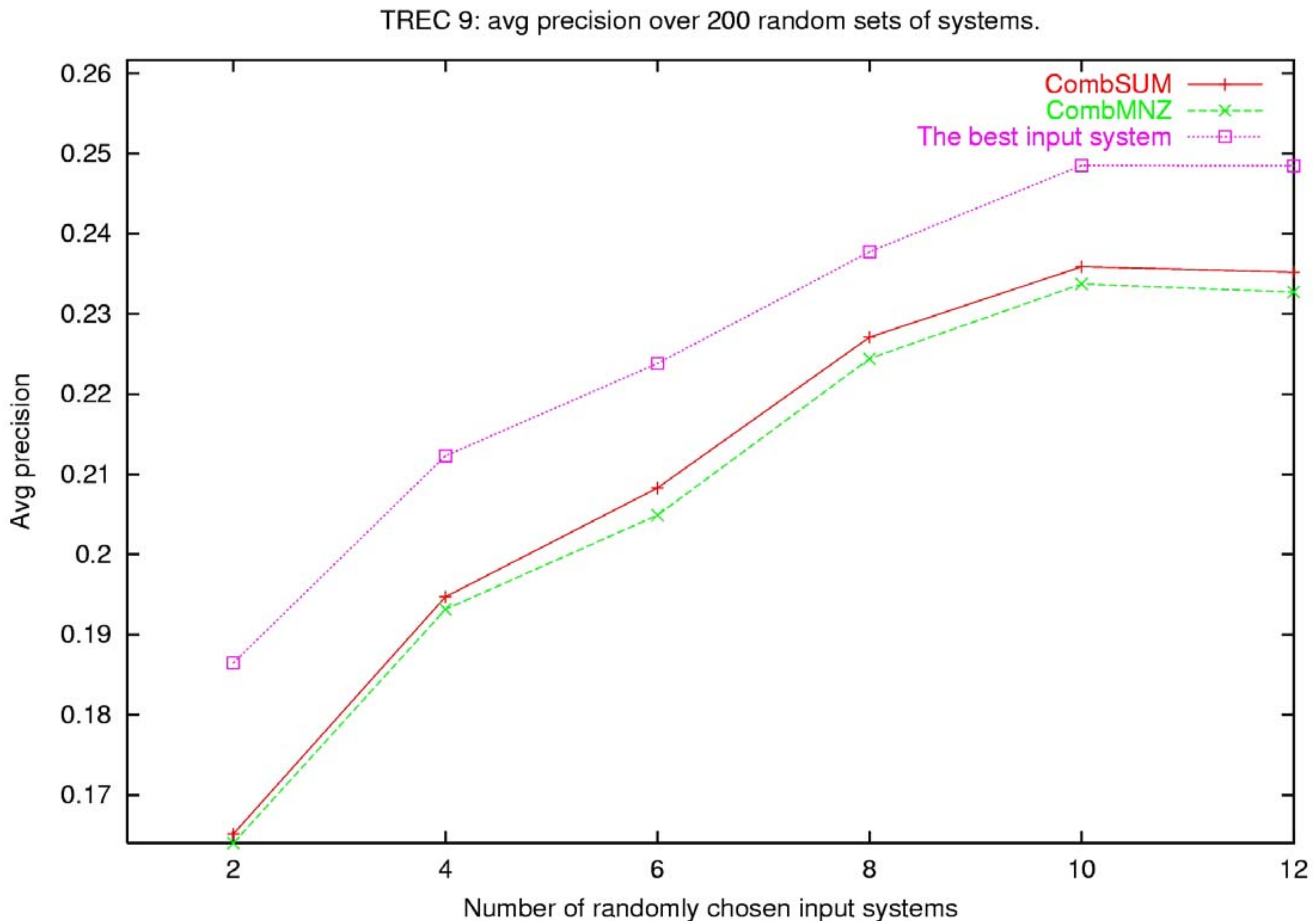
CombMNZ on TREC3



CombMNZ on Vogt



CombMNZ on TREC9



Borda Count

- Consider an n candidate election.
- For each ballot:
  - assign n points to top candidate,
  - assign n-1 points to next candidate,
  - …
- Rank candidates by point sum.

Borda Count: Election 2000

- Ideological order: Nader, Gore, Bush.
- Ideological voting:
  - Bush voter: Bush, Gore, Nader.
  - Nader voter: Nader, Gore, Bush.
  - Gore voter:
    - Gore, Bush, Nader.
    - Gore, Nader, Bush.

50/50, 100/0

Election 2000: Ideological Florida Voting

	Gore	Bush	Nader
50/50	14,734,379	13,185,542	7,560,864
100/0	14,734,379	14,639,267	6,107,138

Gore Wins

Borda Count: Election 2000

- Ideological order: Nader, Gore, Bush.
- Manipulative voting:
  - Bush voter: Bush, Nader, Gore.
  - Gore voter: Gore, Nader, Bush.
  - Nader voter: Nader, Gore, Bush.

Election 2000: Manipulative Florida Voting

Gore	Bush	Nader
11,825,203	11,731,816	11,923,765

Nader Wins

Future Work

- Bayes
  - approximate dependence.
- Condorcet
  - weighting, dependence.
- Upper bounds
  - other constraints.
- Meta-retrieval
  - Metasearch is approaching fundamental limits.
  - Need to incorporate user feedback: learning…