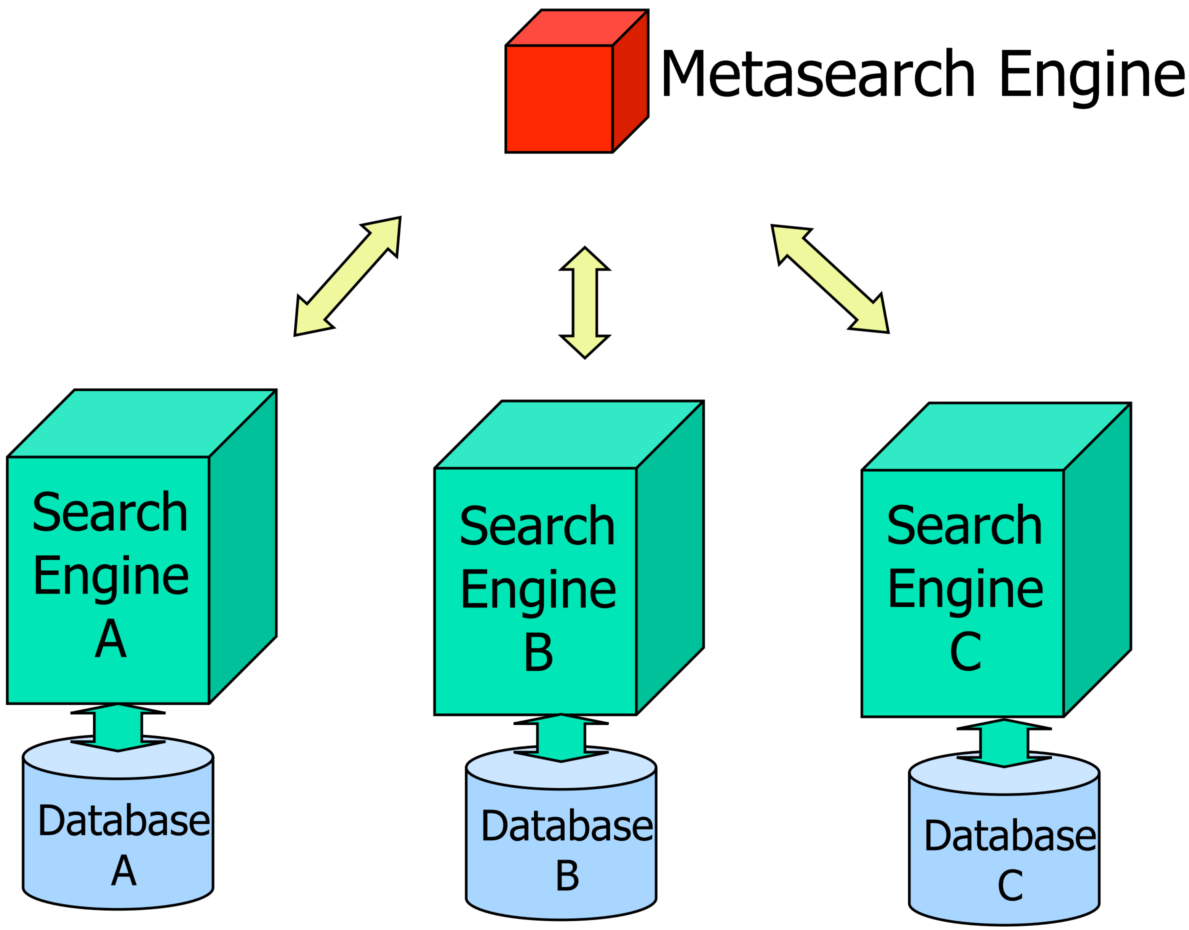
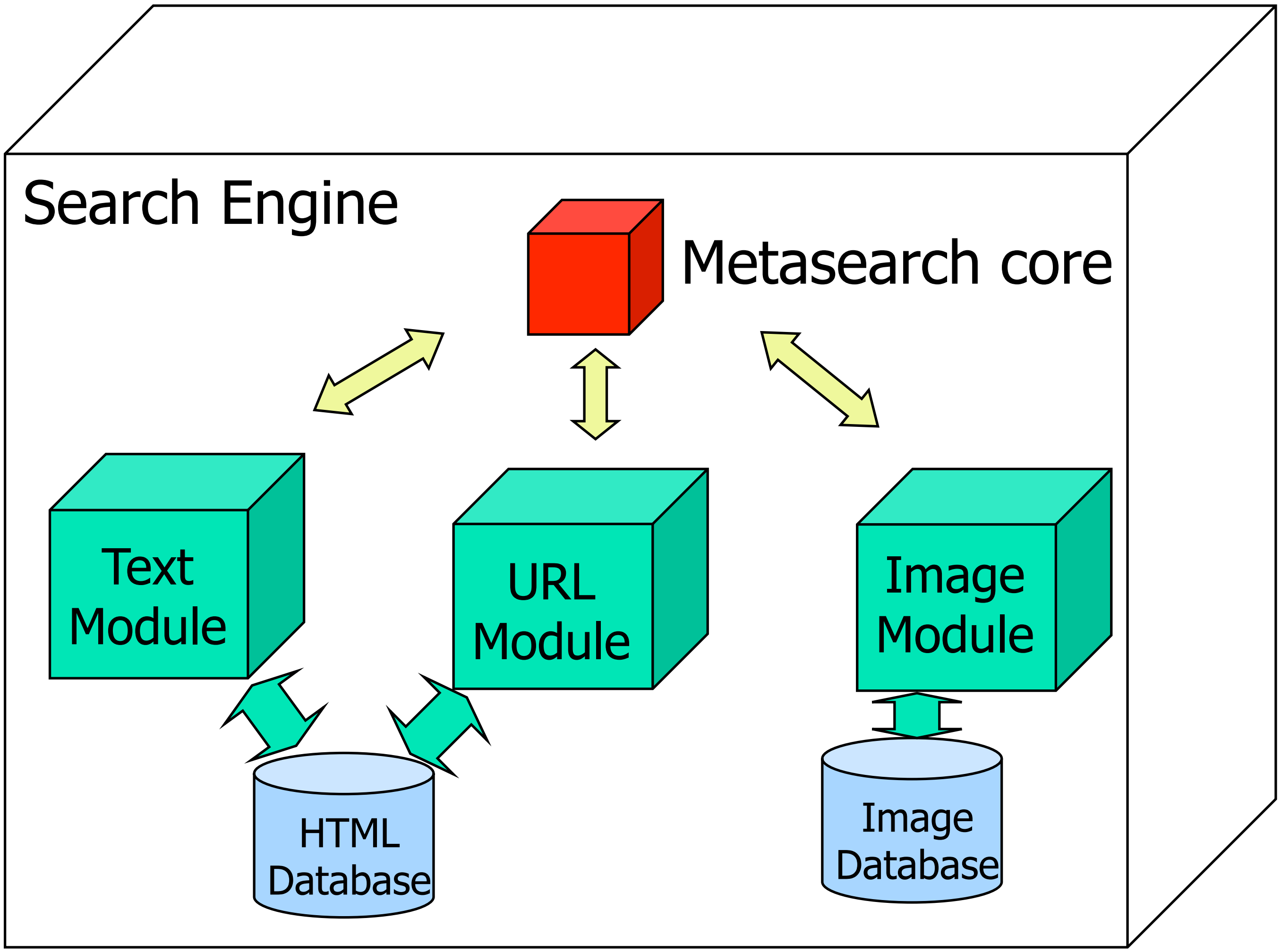
Search Engines

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

External Metasearch



Internal Metasearch



Metasearch Engines

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

Outline

- ✓ Introduce problem
- Characterize problem
- Survey techniques
- Upper bounds for metasearch

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…*
- Today: data fusion.

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 techniques
- Upper bounds for metasearch

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

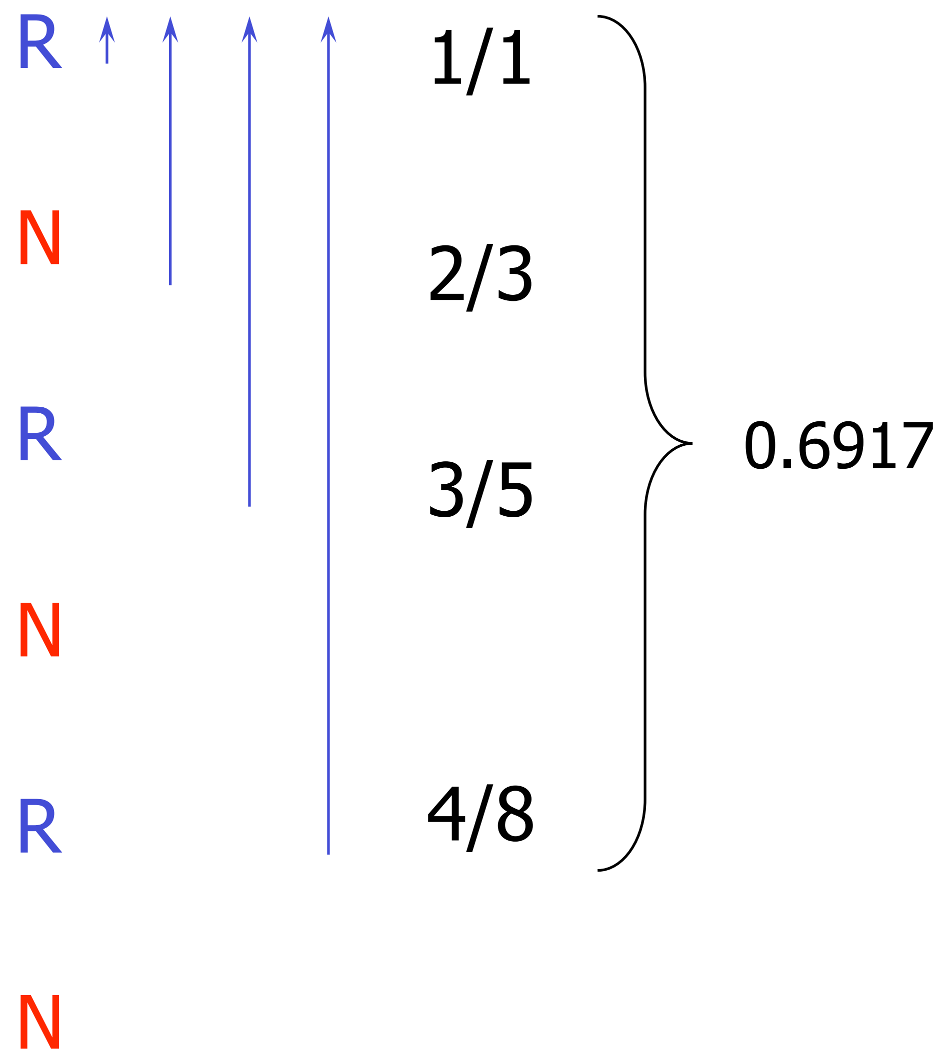
CombMNZ [Fox, Shaw, Lee, et al.]

- 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



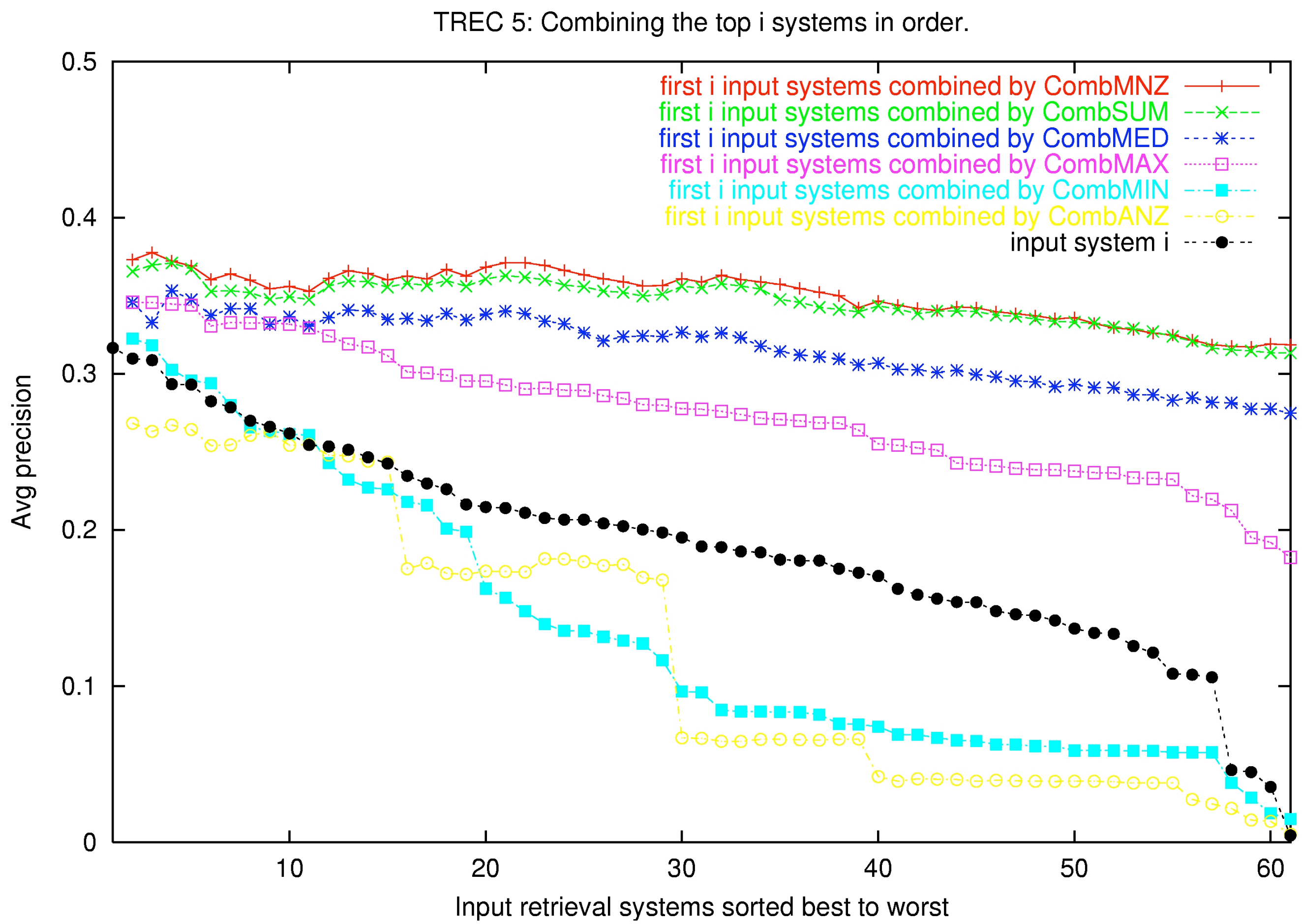
Benchmark Data: TREC

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

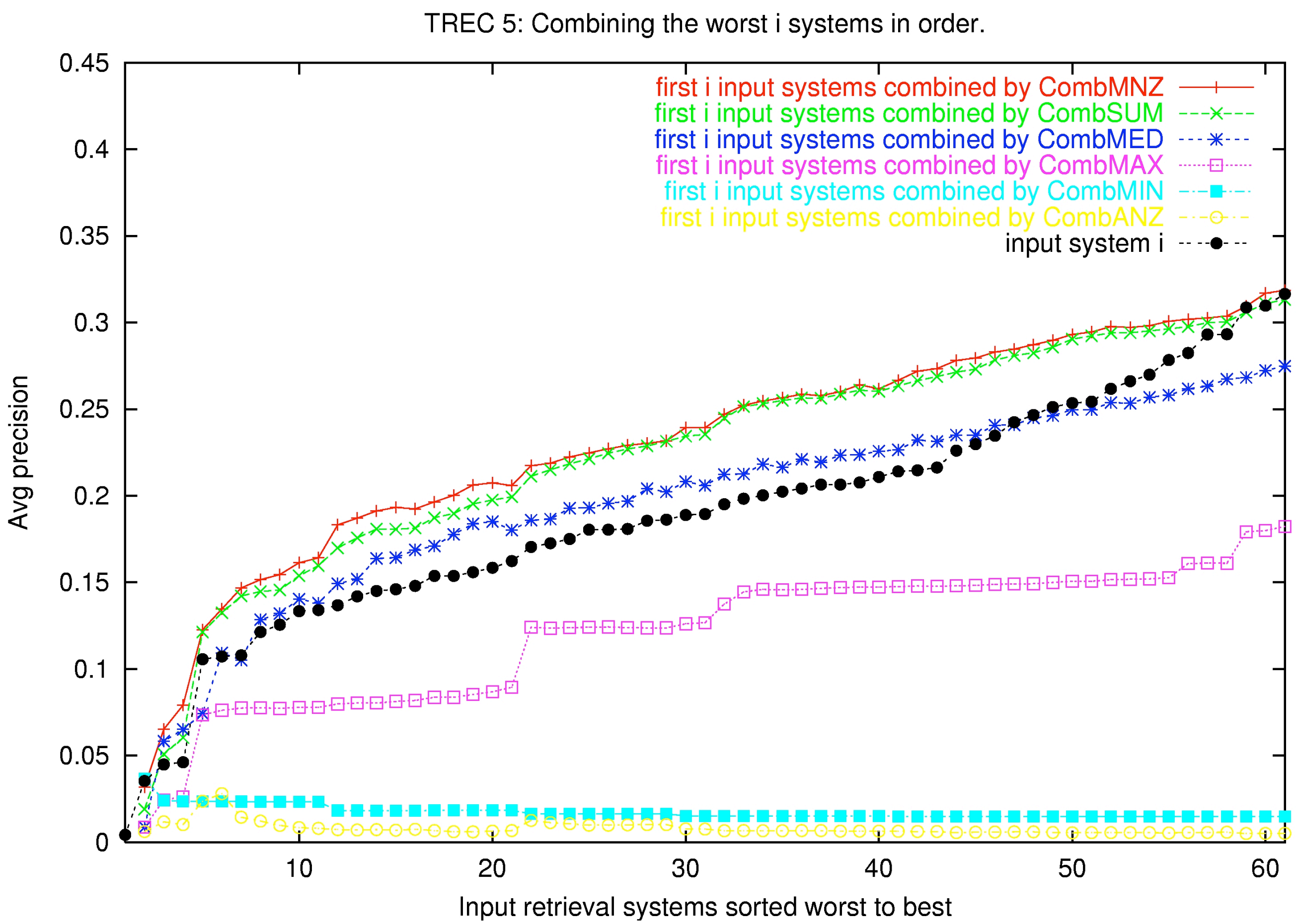
Data Sets

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



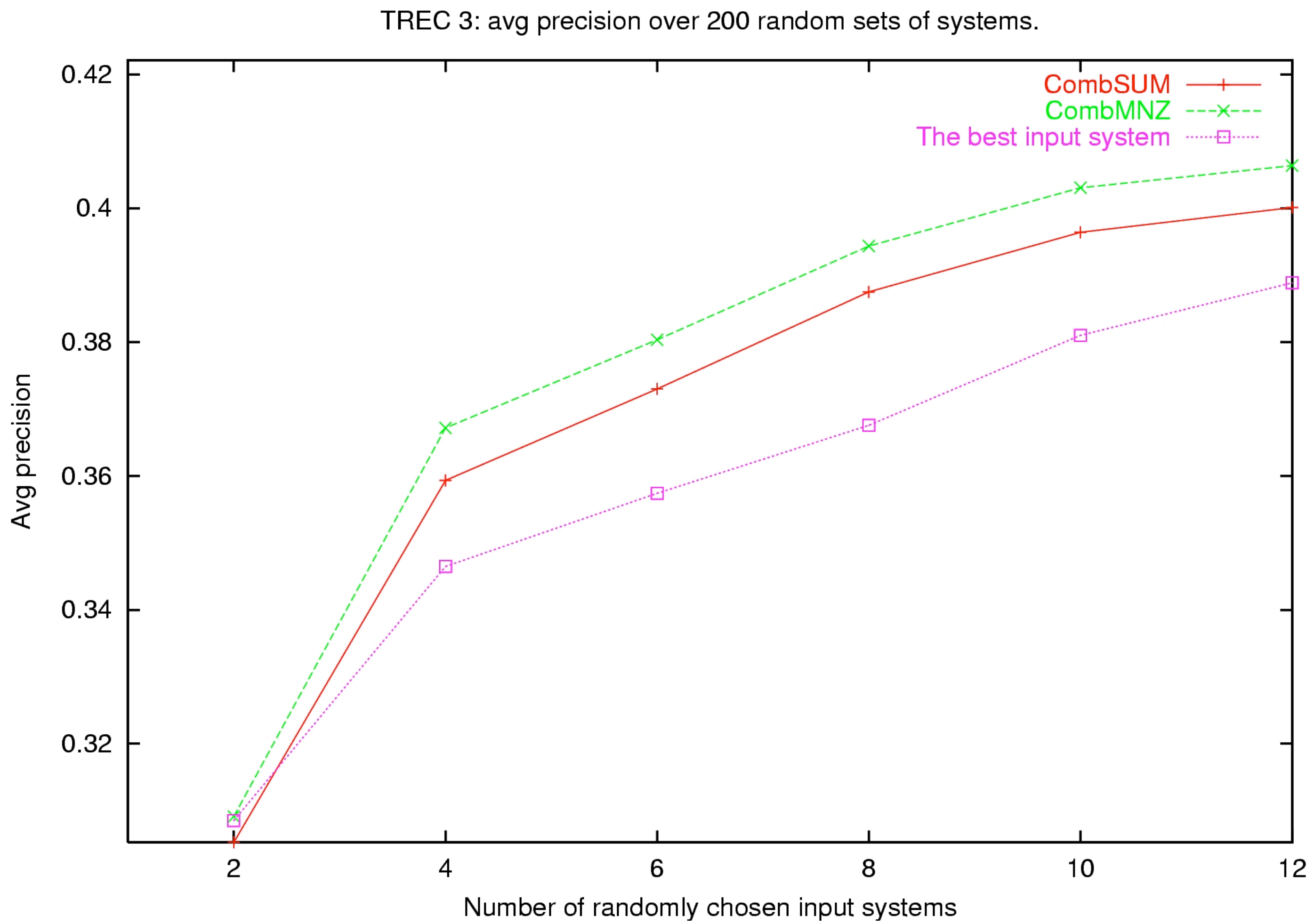
CombX on TREC5 Data, II



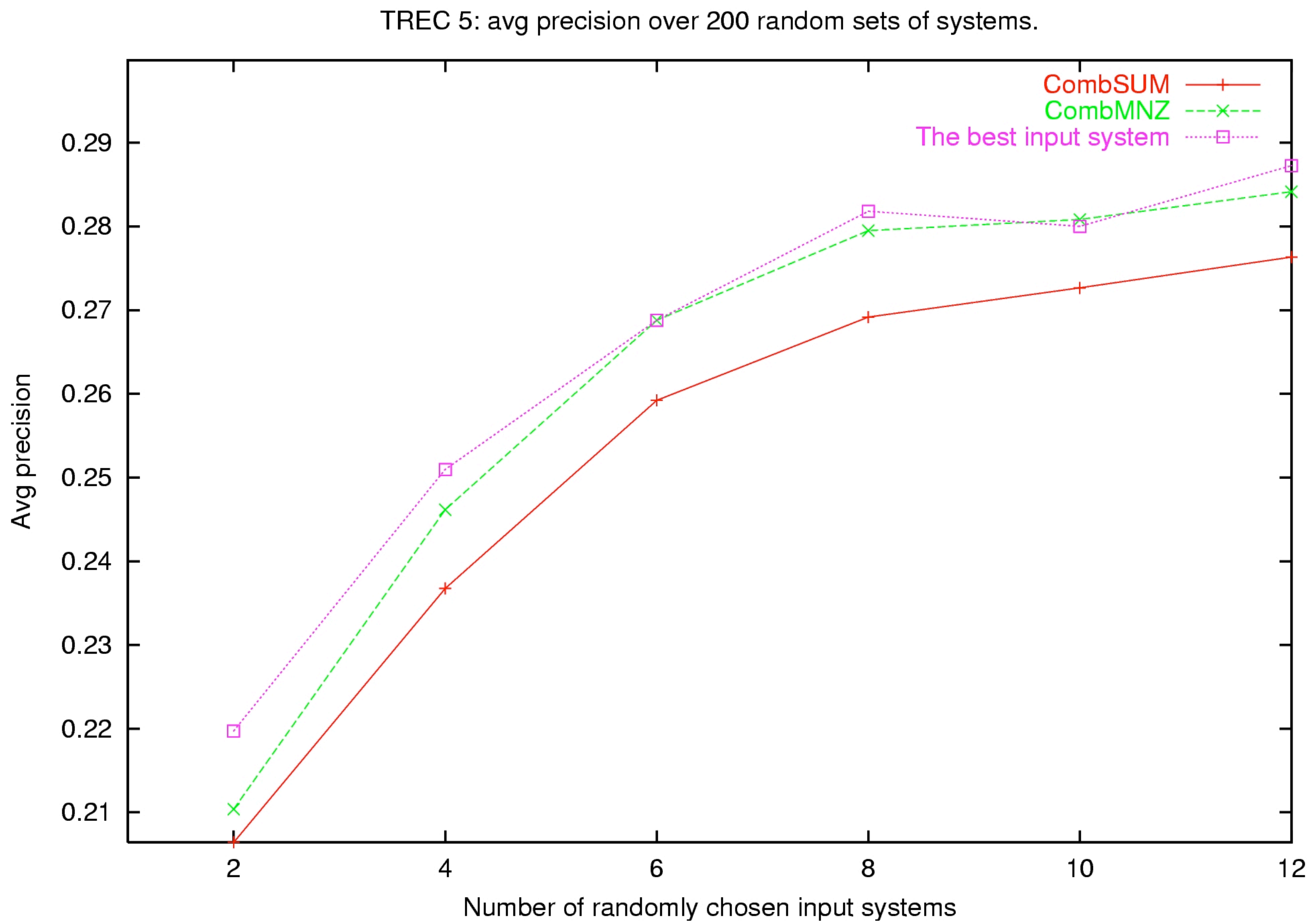
Experiments

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

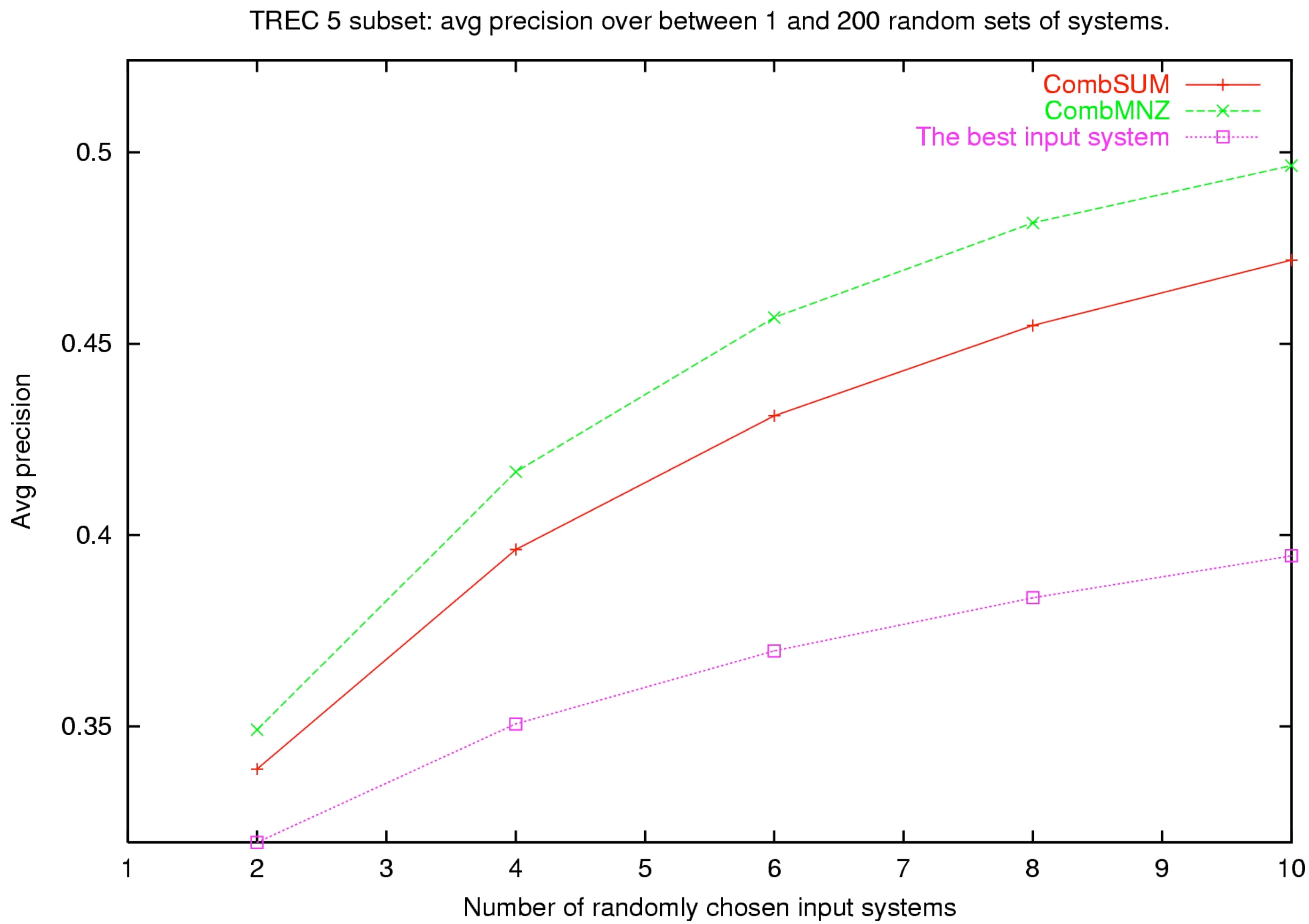
CombMNZ on TREC3



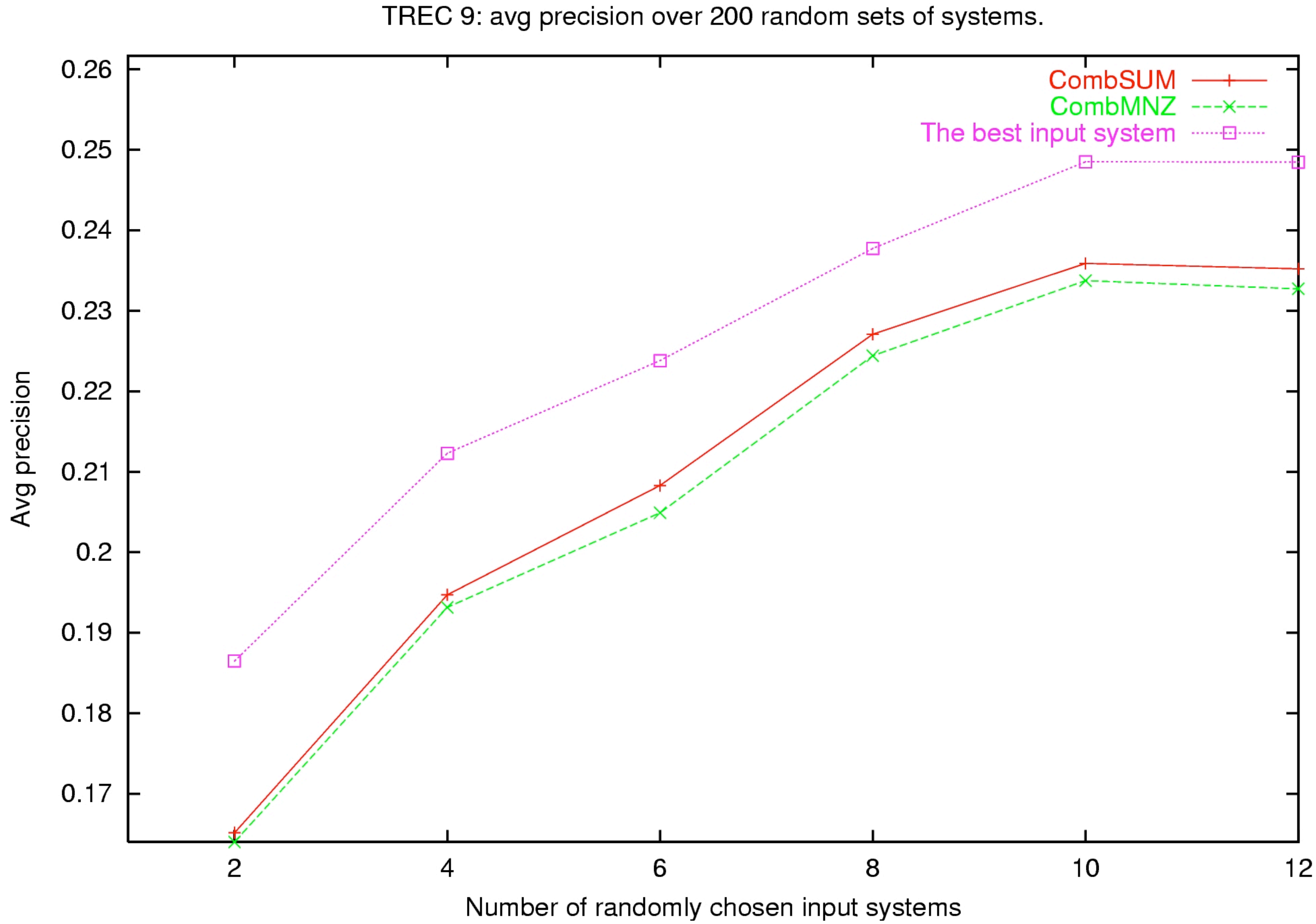
CombMNZ on TREC5



CombMNZ on Vogt



CombMNZ on TREC9



Metasearch via Voting

[Aslam, Montague]

- 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.

Borda Count

- Consider an n candidate election.
- One method for choosing winner is the Borda count. [Borda, Saari]
  - For each voter i
    - Assign n points to top candidate.
    - Assign n-1 points to next candidate.
    - …
  - Rank candidates according to point sum.

Election 2000: Florida

NATIONAL > FLORIDA

VOTER RESULTS IN FLORIDA | EXIT POLLS | HOUSE AND SENATE

Last updated: 12:32 a.m. EST, 12/14 | refresh this page | print this page

FLORIDA VOTE COUNT

	Nov. 7	Recount	Certified	12/8 Ruling
Bush ✓	1,725	930	537	193
Source: Associated Press				

25 electoral votes at stake

PRESIDENT DEC. 13

100% of precincts

Candidates	Votes	Vote %	States Won	EV
R Bush ✓	2,909,176	49 %	29	0
D Gore	2,907,451	49 %	20	0
G Nader	96,837	2 %	0	0
I Browne	18,856	0 %	0	0
RF Buchanan	17,356	0 %	0	0
I Phillips	4,280	0 %	0	0
I Hagelin	2,287	0 %	0	0

winner declared — exit polls

results as of 5:46 p.m. EST

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

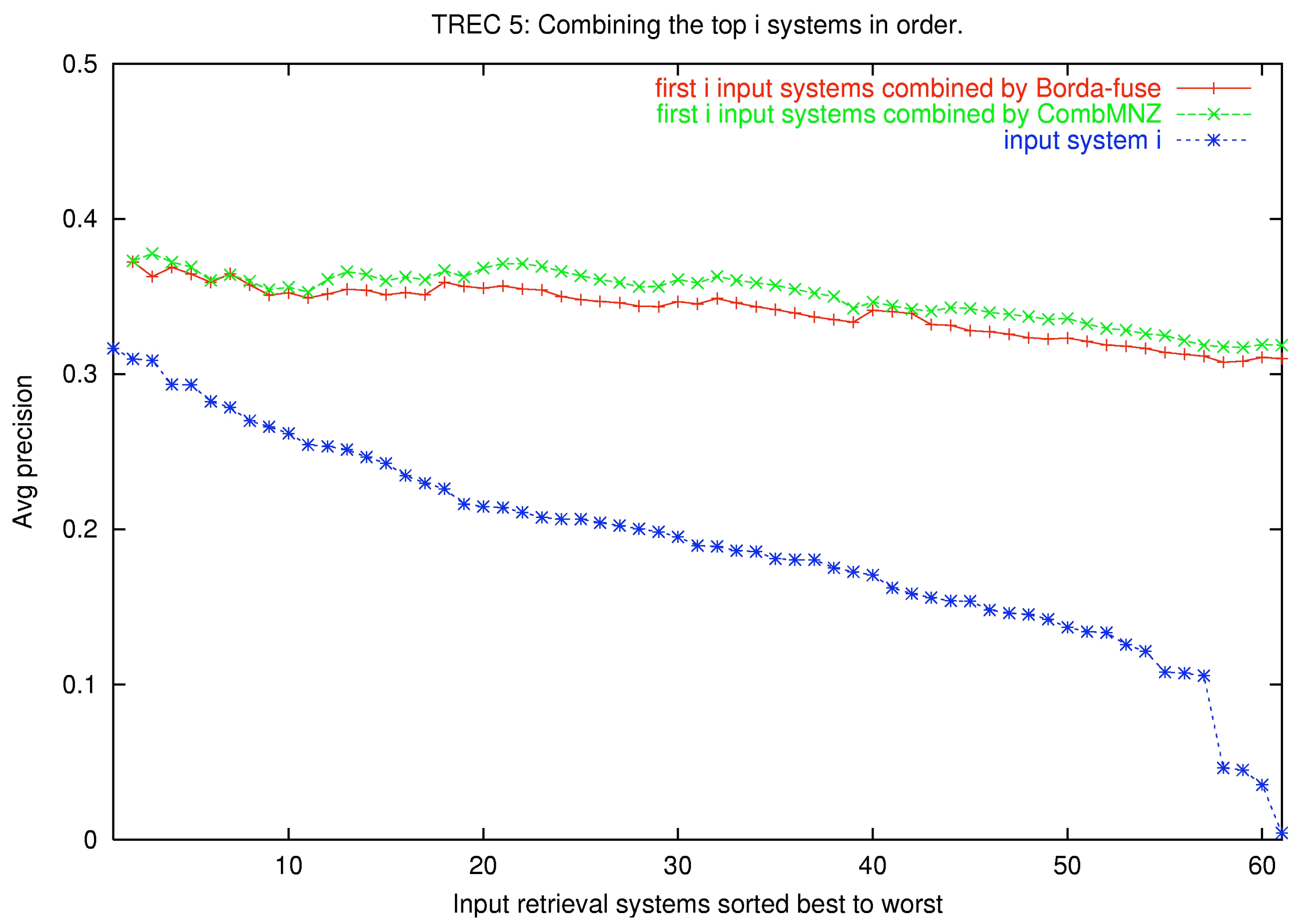
Gore	Bush	Nader
11,825,203	11,731,816	11,923,765

Nader Wins

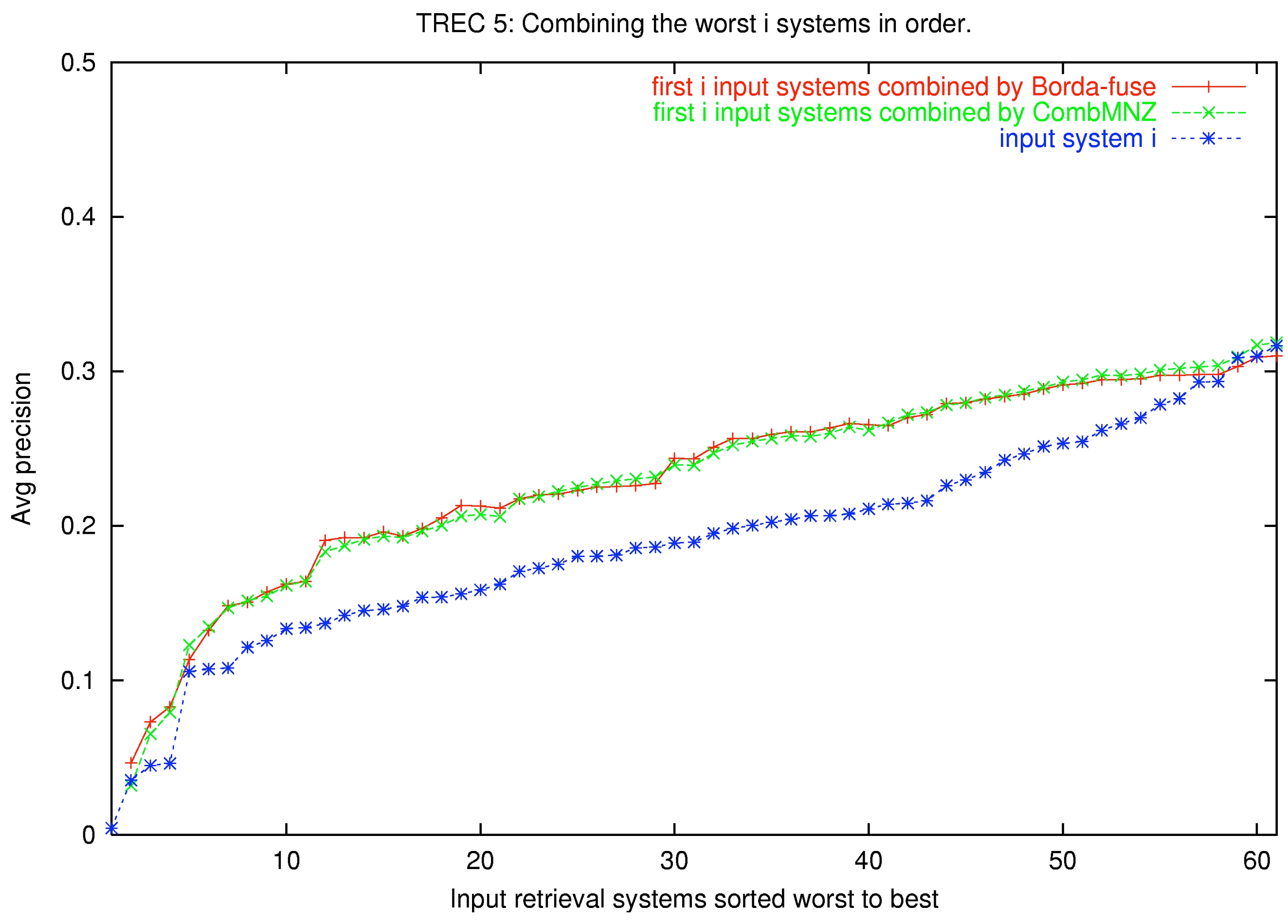
Metasearch via Borda Counts

- Metasearch analogy:
  - Documents are *candidates*.
  - Systems are *voters* providing preferential rankings.
- Issues:
  - Systems may rank different document sets.
  - How to deal with unranked documents?

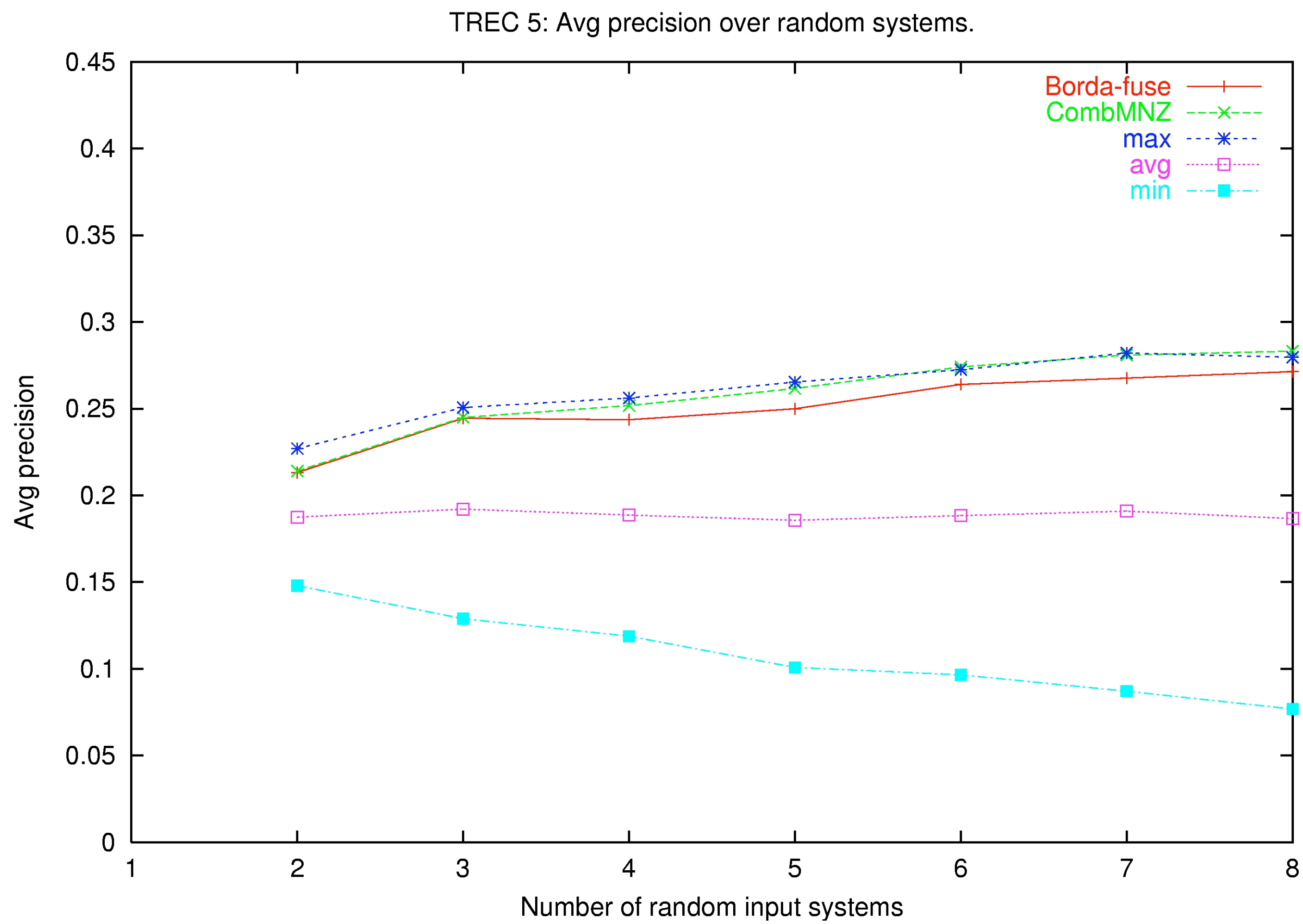
Borda on TREC5 Data, I



Borda on TREC5 Data, II



Borda on TREC5 Data, III



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.

Election 2000: Florida

NATIONAL > FLORIDA

VOTER RESULTS IN FLORIDA | EXIT POLLS | HOUSE AND SENATE

Last updated: 12:32 a.m. EST, 12/14 | refresh this page | print this page

FLORIDA VOTE COUNT

	Nov. 7	Recount	Certified	12/8 Ruling
Bush	1,725	930	537	193

Source: Associated Press

25 electoral votes at stake

PRESIDENT DEC. 13

100% of precincts

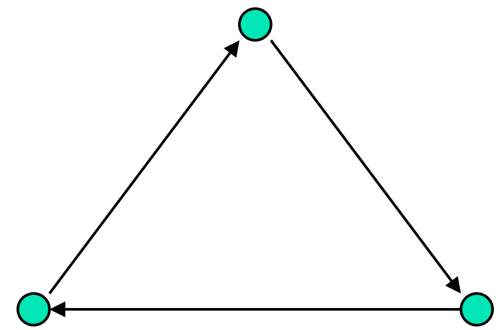
Candidates	Votes	Vote %	States Won	EV
R Bush	2,909,176	49 %	29	0
D Gore	2,907,451	49 %	20	0
G Nader	96,837	2 %	0	0
I Browne	18,856	0 %	0	0
RF Buchanan	17,356	0 %	0	0
I Phillips	4,280	0 %	0	0
I Hagelin	2,287	0 %	0	0

winner declared — exit polls

results as of 5:46 p.m. EST

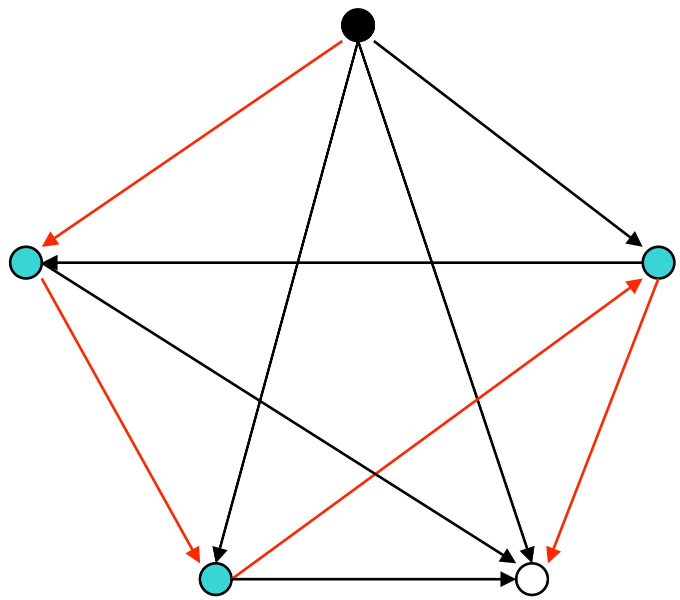
Condorcet Paradox

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



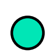
- Cyclic preferences: cycle in Condorcet graph.
- Condorcet consistent path: Hamiltonian.
- For metasearch: any CC path will do.

Condorcet Consistent Path

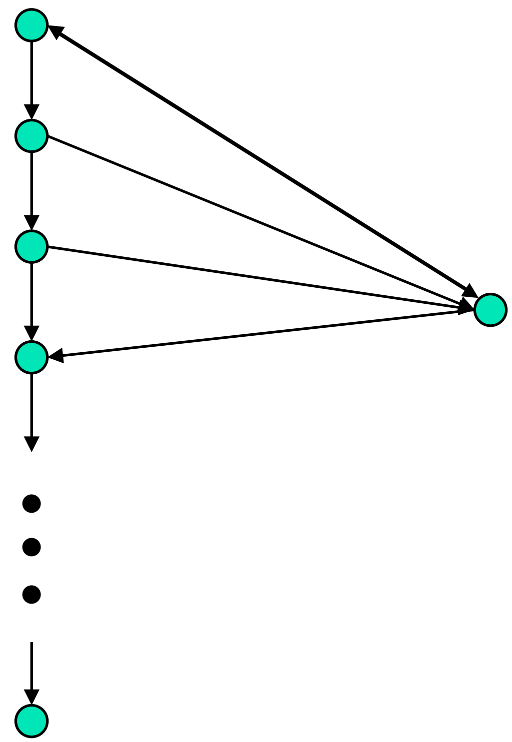


Hamiltonian Path Proof

Base Case:



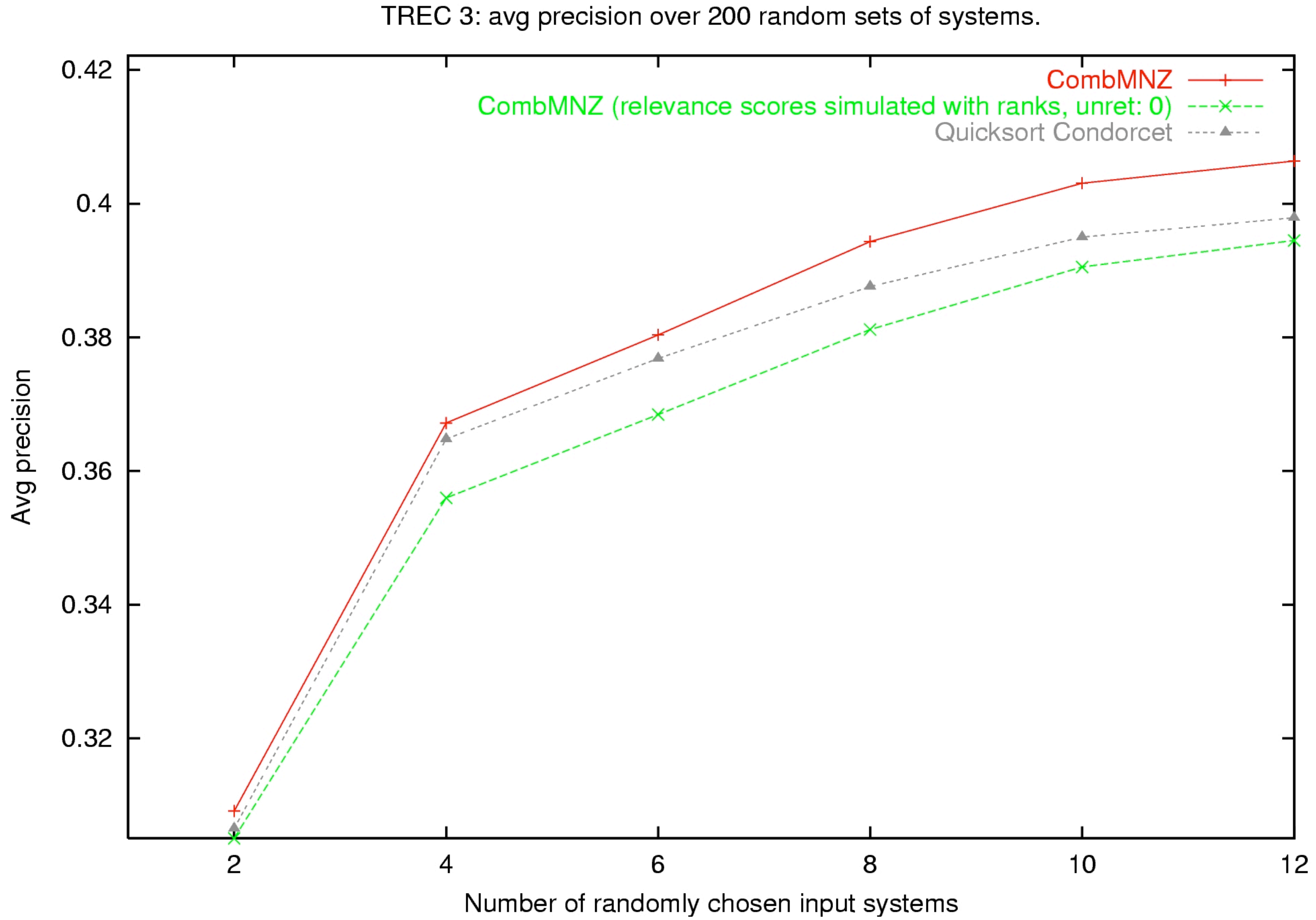
Inductive Step:



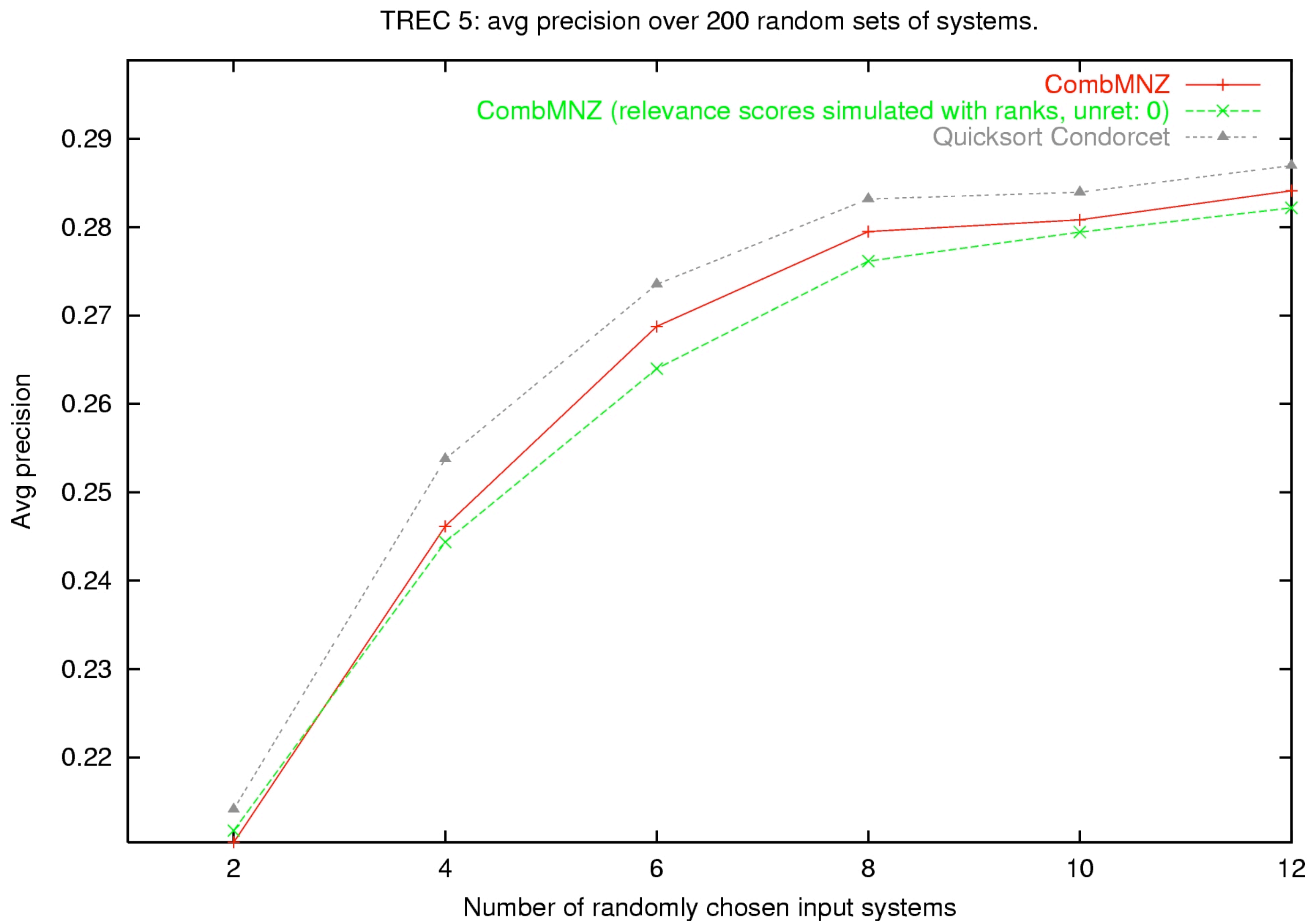
Condorcet-fuse: Sorting

- 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.*

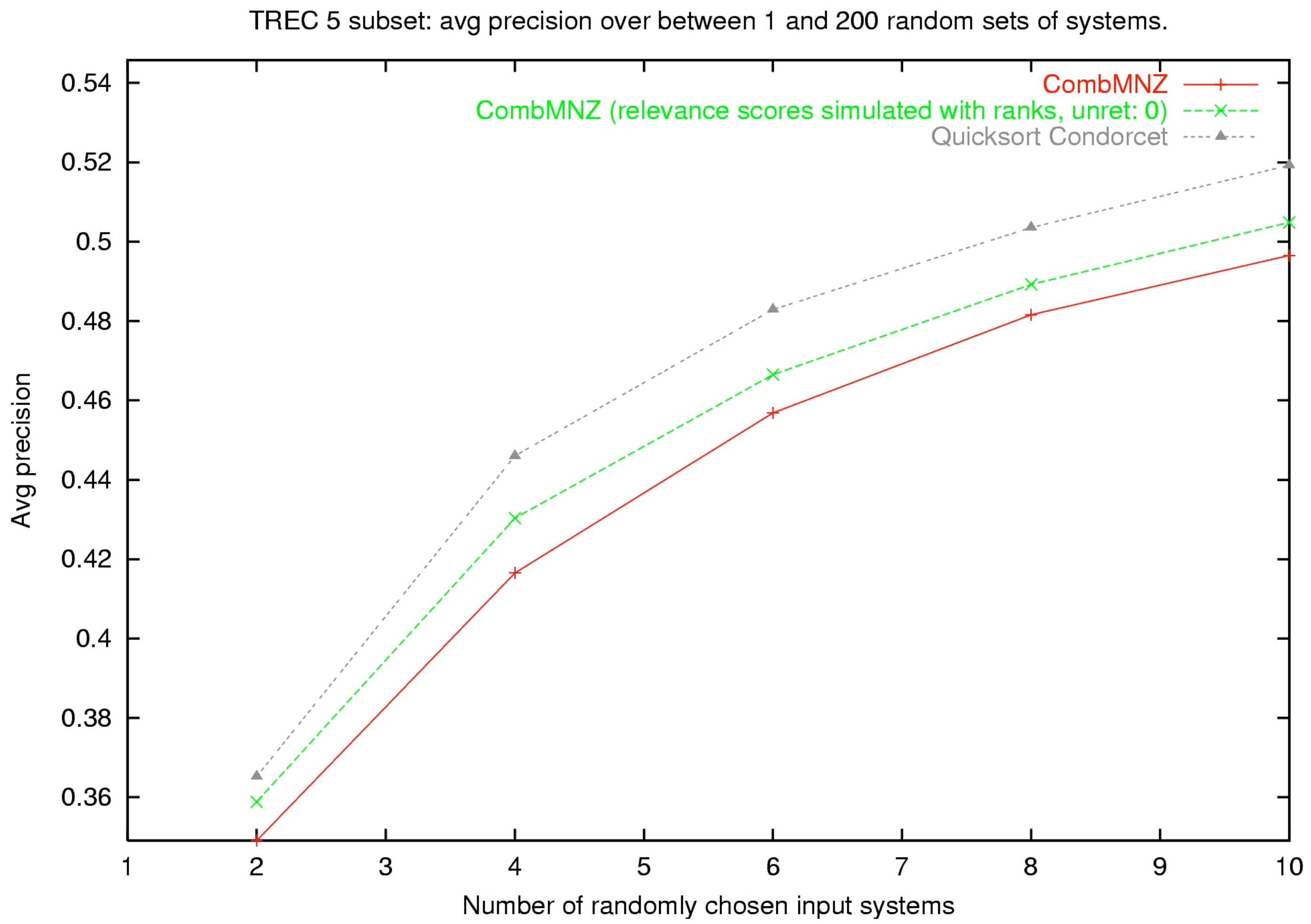
Condorcet-fuse on TREC3



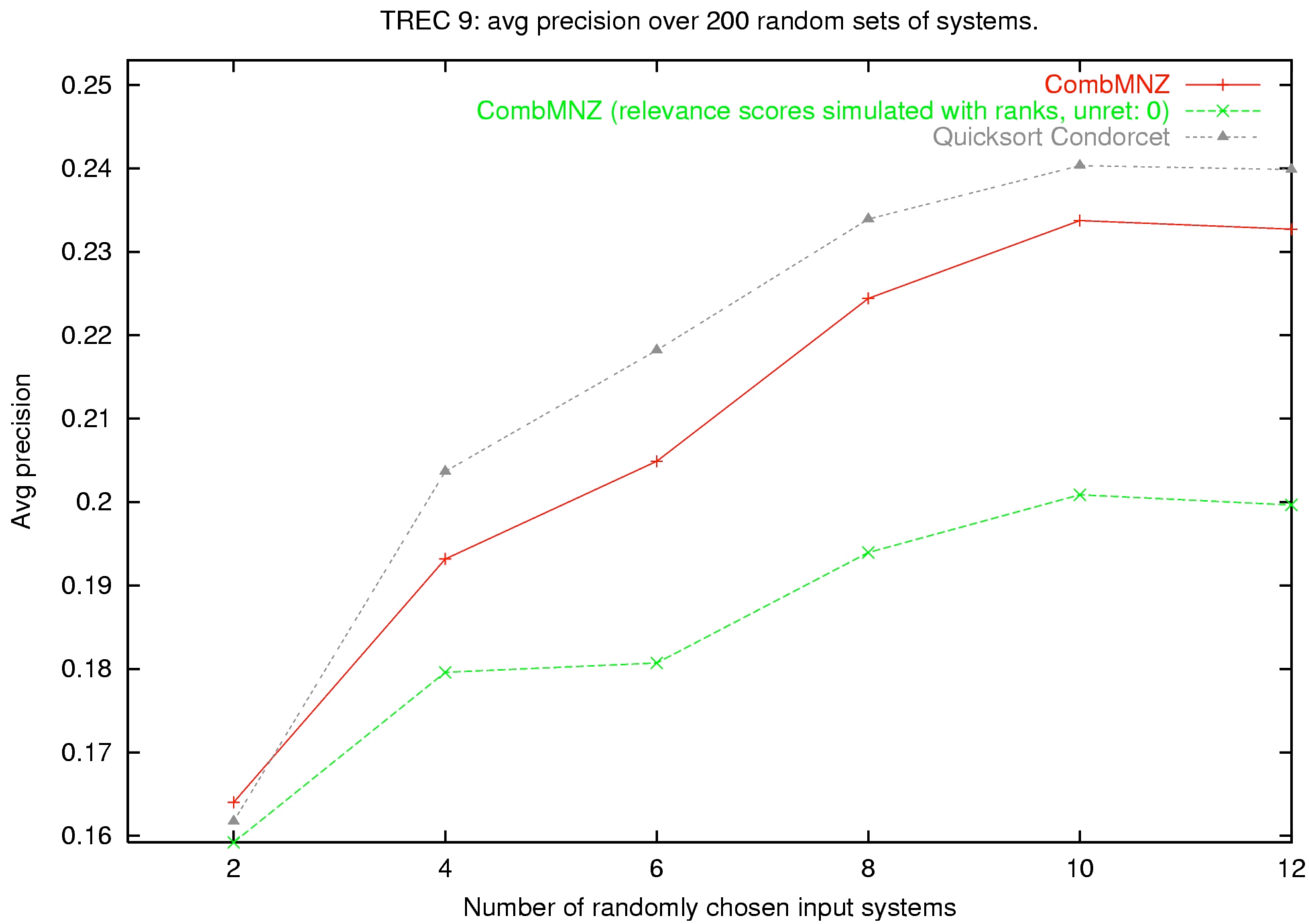
Condorcet-fuse on TREC5



Condorcet-fuse on Vogt



Condorcet-fuse on TREC9



Outline

- ✓ Introduce problem
- ✓ Characterize problem
- ✓ Survey techniques
- Upper bounds for metasearch

Upper Bounds on Metasearch

- How good can metasearch be?
- Are there fundamental limits that methods are approaching?

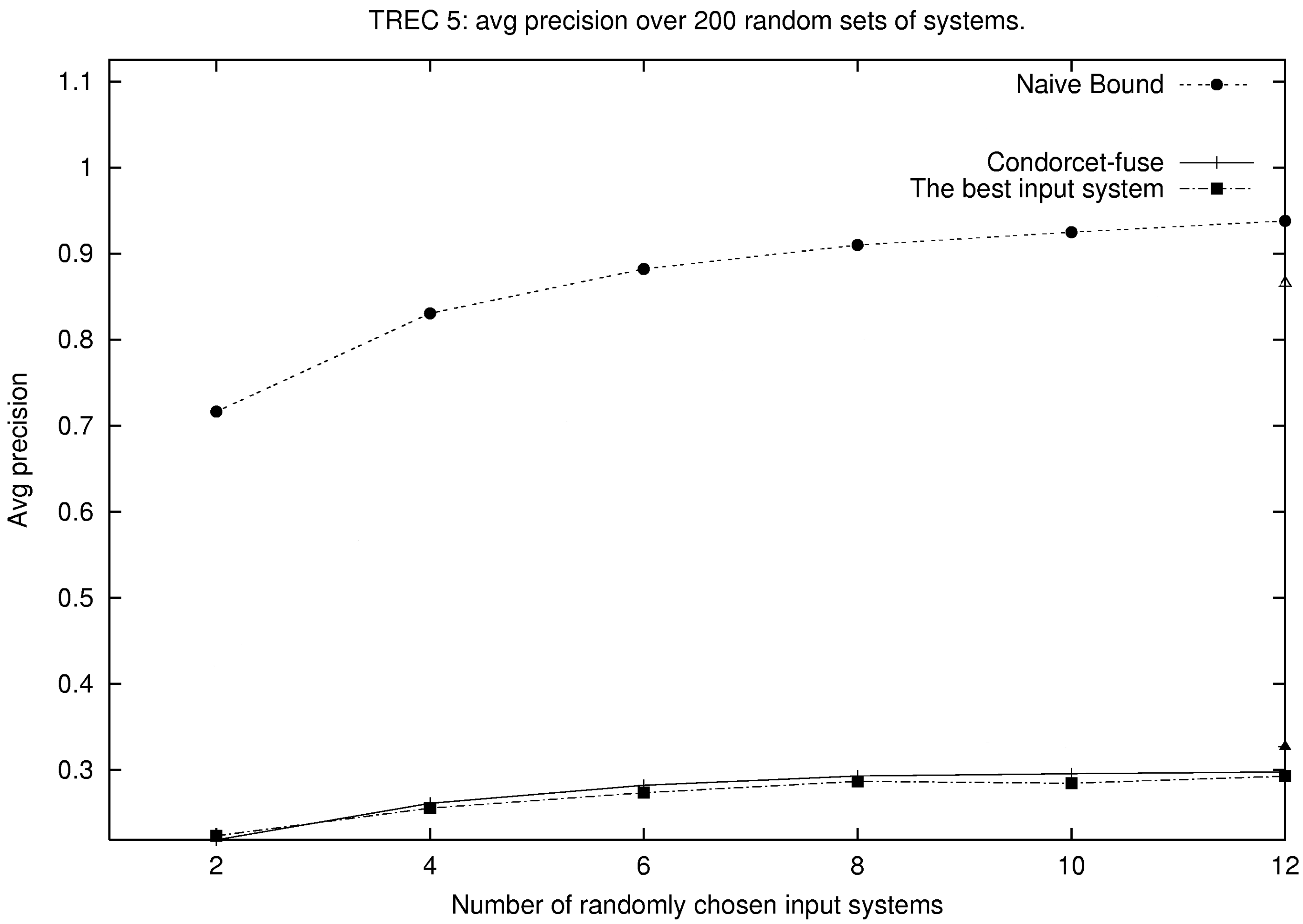
Upper Bounds on Metasearch

- 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 relevant docs above irrelevant docs.

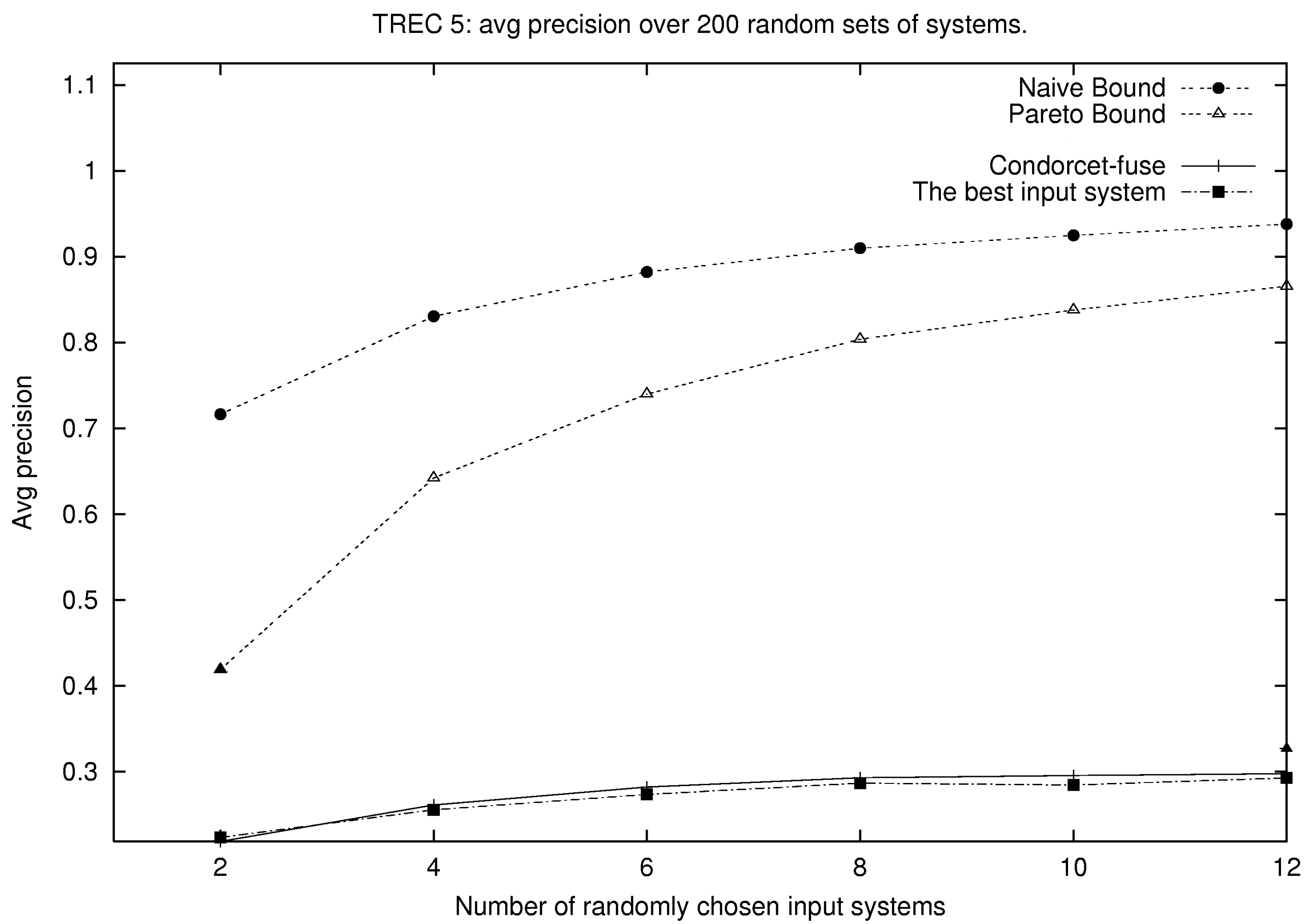
TREC5: Naïve Bound



Pareto Constraint

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

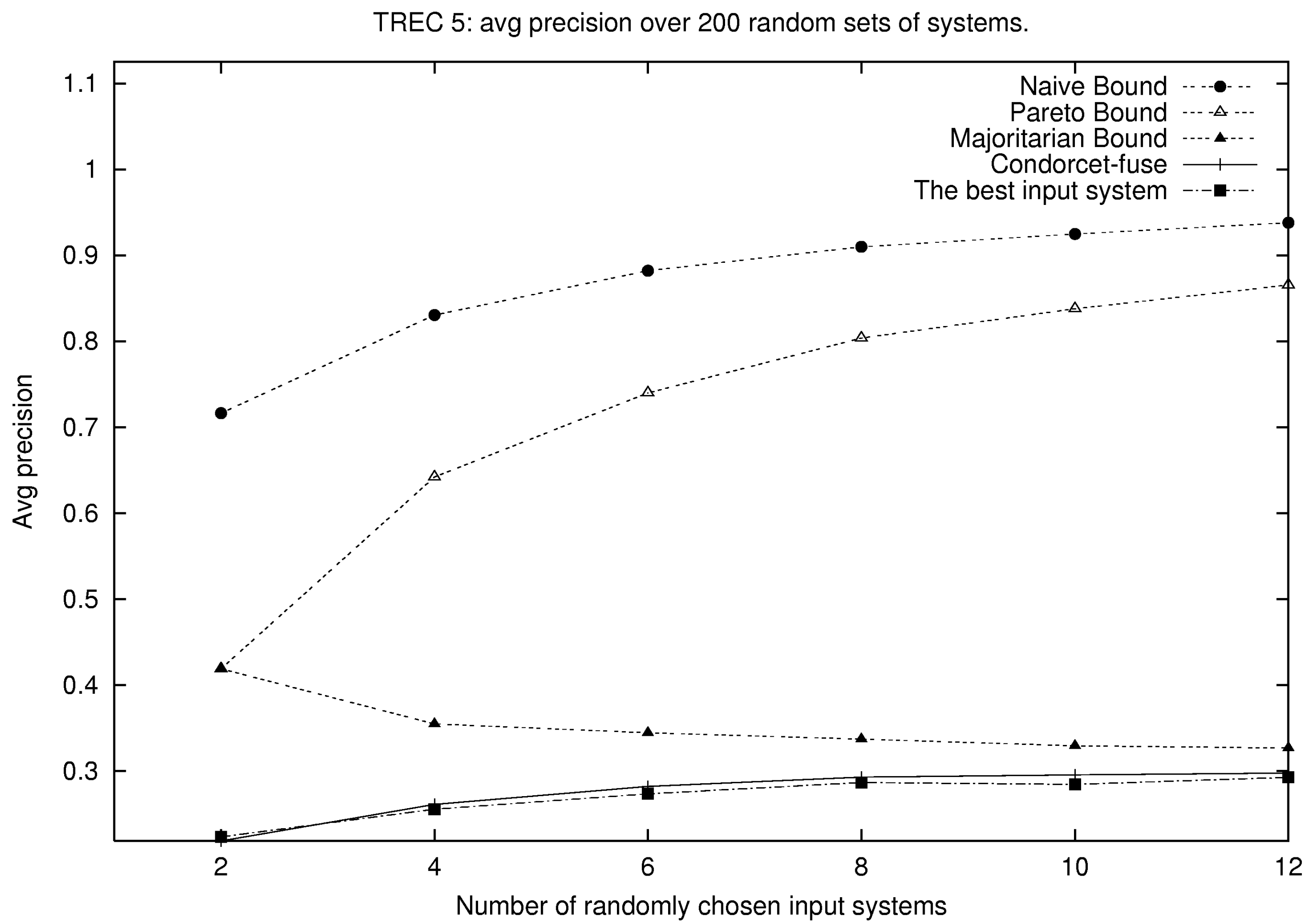
TREC5: Pareto Bound



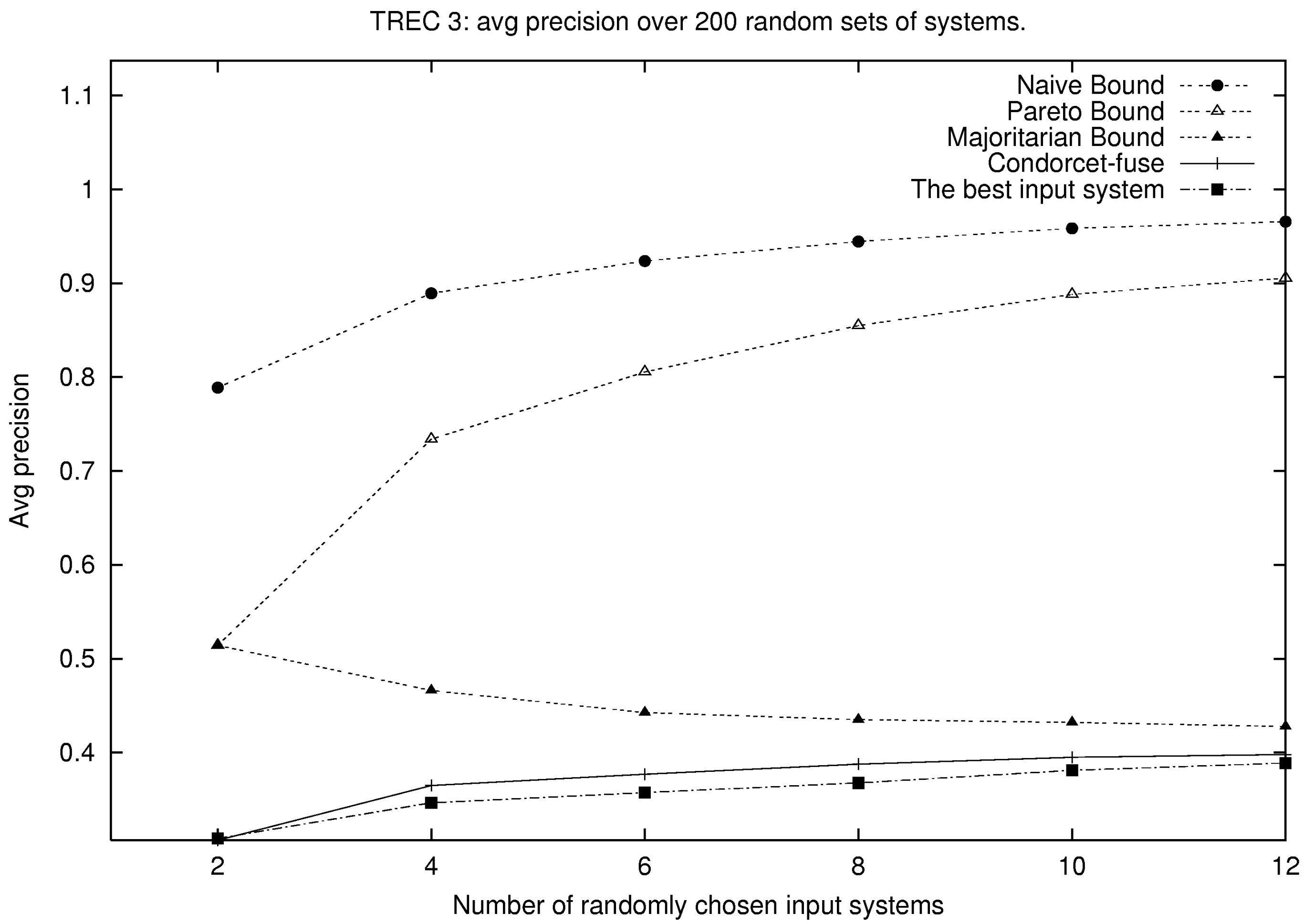
Majoritarian Constraint

- *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.

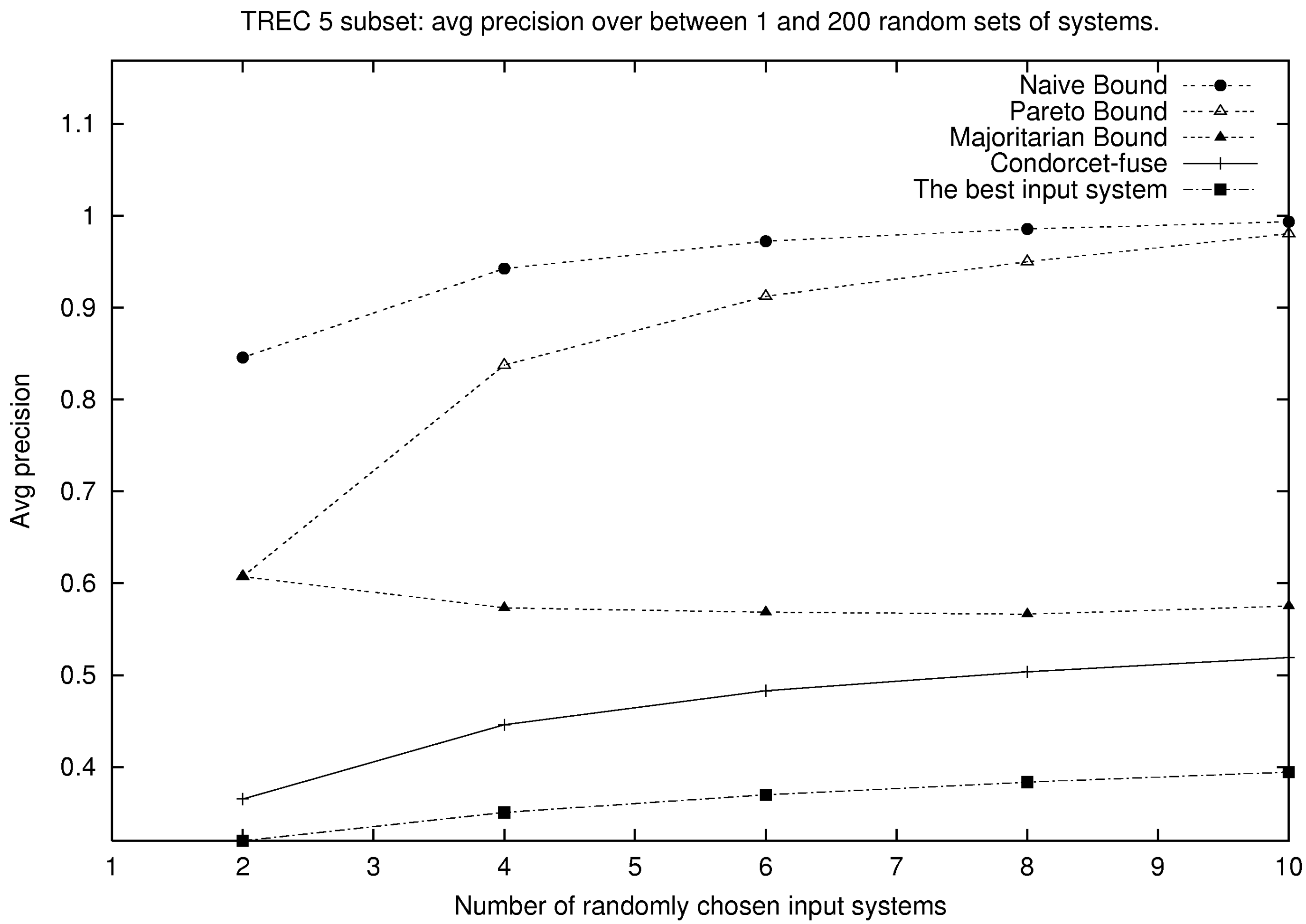
TREC5: Majoritarian Bound



Upper Bounds: TREC3



Upper Bounds: Vogt



Upper Bounds: TREC9