

The TREC Conferences: An Introduction



Ellen Voorhees

NIST

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

Text REtrieval Conference (TREC)

Talk Outline

- General introduction to TREC
 - TREC history
 - TREC impacts
- Cranfield tradition of laboratory tests
 - mechanics of building test collections
 - test collection quality
 - legitimate uses of test collections
- IR evaluation primer

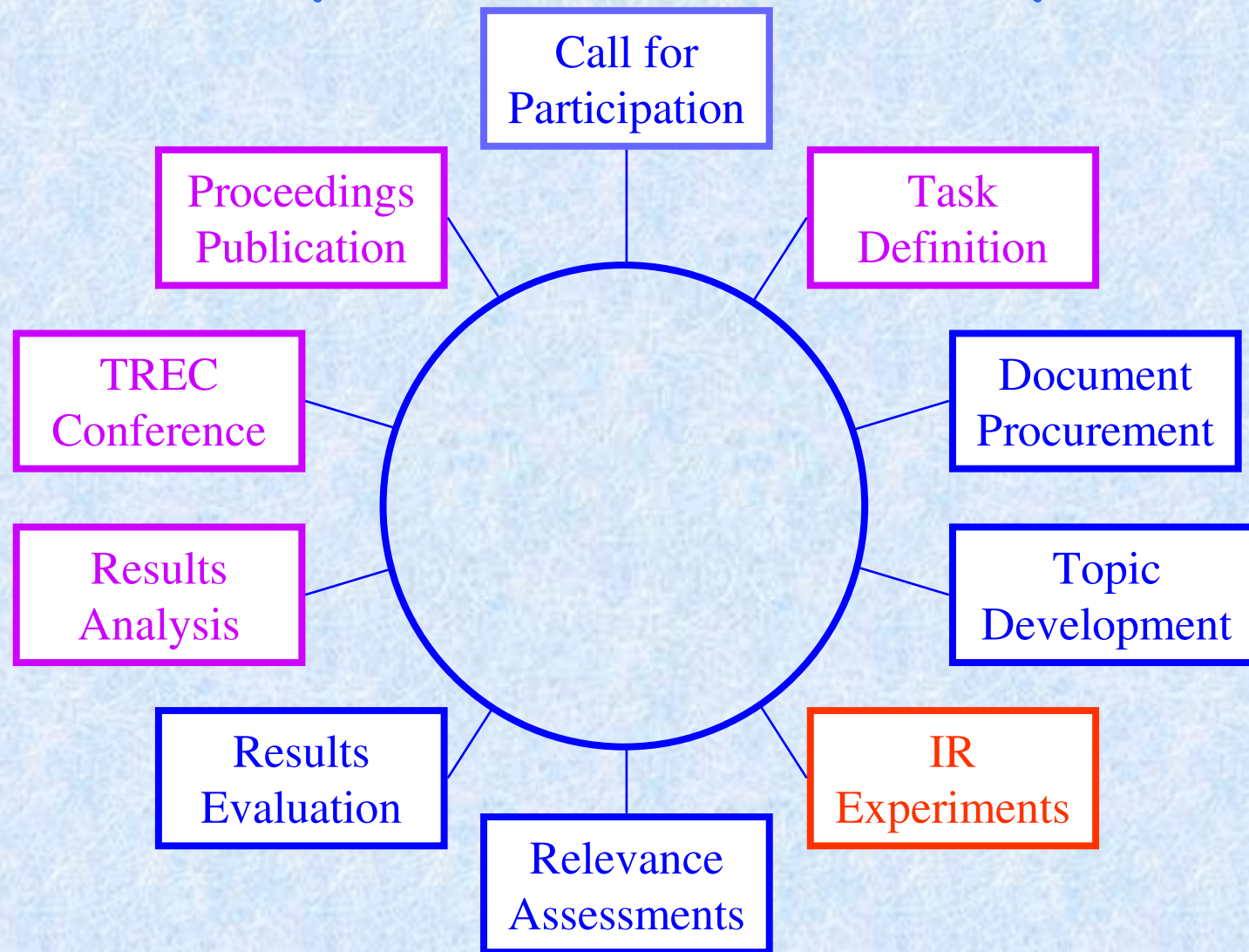
What is TREC?

- A workshop series that provides the infrastructure for large-scale testing of (text) retrieval technology
 - realistic test collections
 - uniform, appropriate scoring procedures
 - a forum for the exchange of research ideas and for the discussion of research methodology

TREC Philosophy

- TREC is a modern example of the Cranfield tradition
 - system evaluation based on test collections
- Emphasis on advancing the state of the art from evaluation results
 - TREC's primary purpose is not competitive benchmarking
 - experimental workshop: sometimes experiments fail!

Yearly Conference Cycle



Text REtrieval Conference (TREC)

TREC 2007 Program Committee

Ellen Voorhees, chair	David Lewis
James Allan	John Prager
Chris Buckley	Steve Robertson
Gord Cormack	Mark Sanderson
Sue Dumais	Ian Soboroff
Donna Harman	Richard Tong
Bill Hersh	Ross Wilkinson

TREC 2007 Track Coordinators

- blog: Craig Macdonald, Iadh Ounis, Ian Soboroff
- enterprise: Bailey, Craswell, de Vries, Soboroff
- genomics: Bill Hersh
- legal: Baron, Oard, Thompson, Tomlinson
- million query: James Allan
- QA: Hoa Dang, Diane Kelly, Jimmy Lin
- spam: Gord Cormack

A Brief History of TREC

- 1992: first TREC conference
 - started by Donna Harman and Charles Wayne as 1 of 3 evaluations in DARPA's TIPSTER program
 - first 3 CDs of documents from this era, hence known as the "TIPSTER" CDs
 - open to IR groups not funded by DARPA
 - 25 groups submitted runs
 - two tasks: ad hoc retrieval, routing
 - 2GB of text, 50 topics
 - primarily an exercise in scaling up systems

A Brief History of TREC

- 1993 (TREC-2)
 - true baseline performance for main tasks
- 1995 (TREC-4)
 - official beginning of TREC track structure
- 1998 (TREC-7)
 - routing dropped as a main task
- 2000 (TREC-9)
 - ad hoc main task dropped; first all-track TREC

TREC Tracks

- Task that focuses on a particular subproblem of text retrieval
- Tracks invigorate TREC & keep TREC ahead of the state-of-the-art
 - specialized collections support research in new areas
 - first large-scale experiments debug what the task really is
 - provide evidence of technology's robustness

TREC Tracks

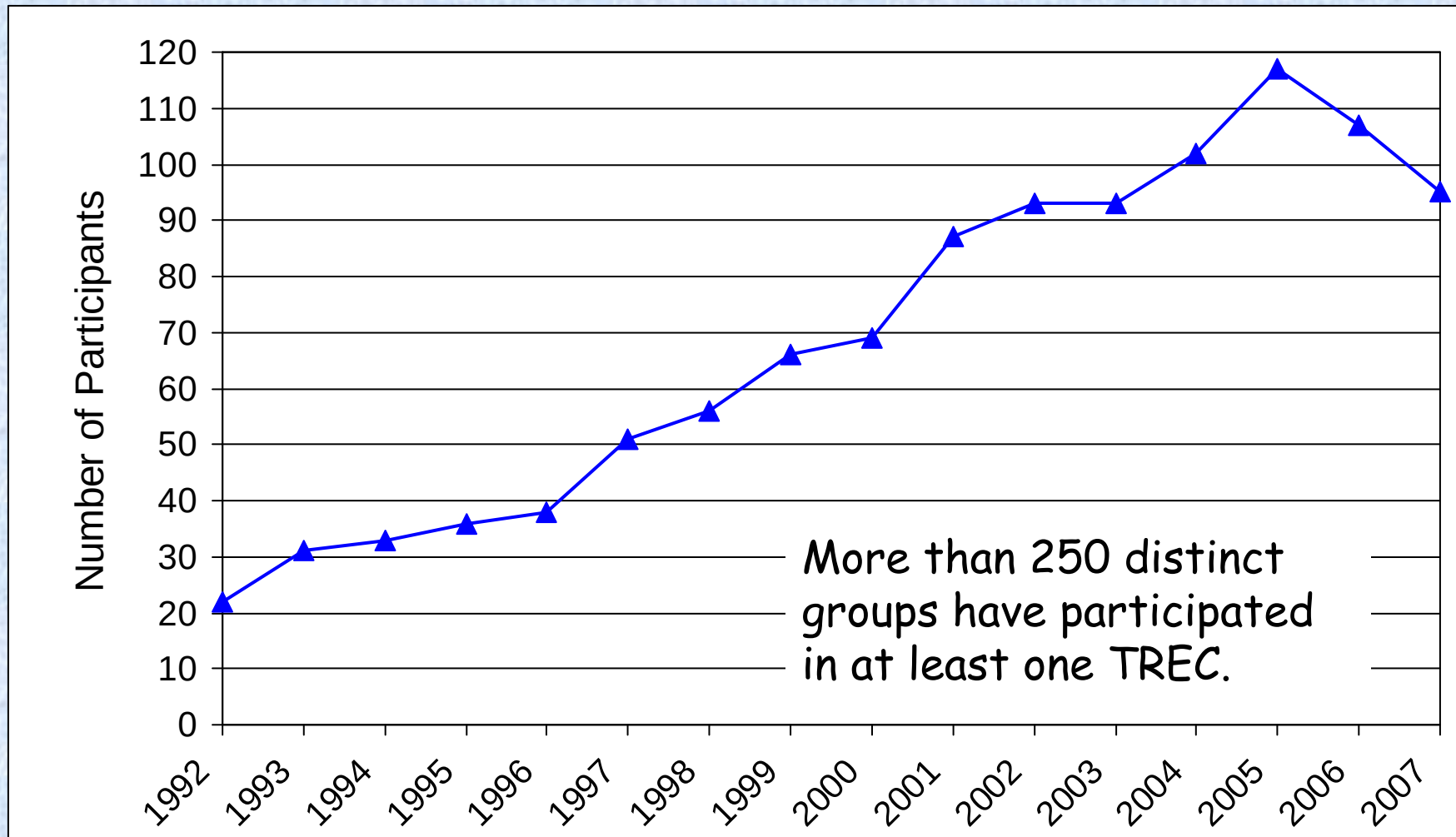
- Set of tracks in a particular TREC depends on:
 - interests of participants
 - appropriateness of task to TREC
 - needs of sponsors
 - resource constraints
- Need to submit proposal for new track in writing to NIST
 - new brainstorming session Thursday morning

The TREC Tracks

[illegible]

Text REtrieval Conference (TREC)

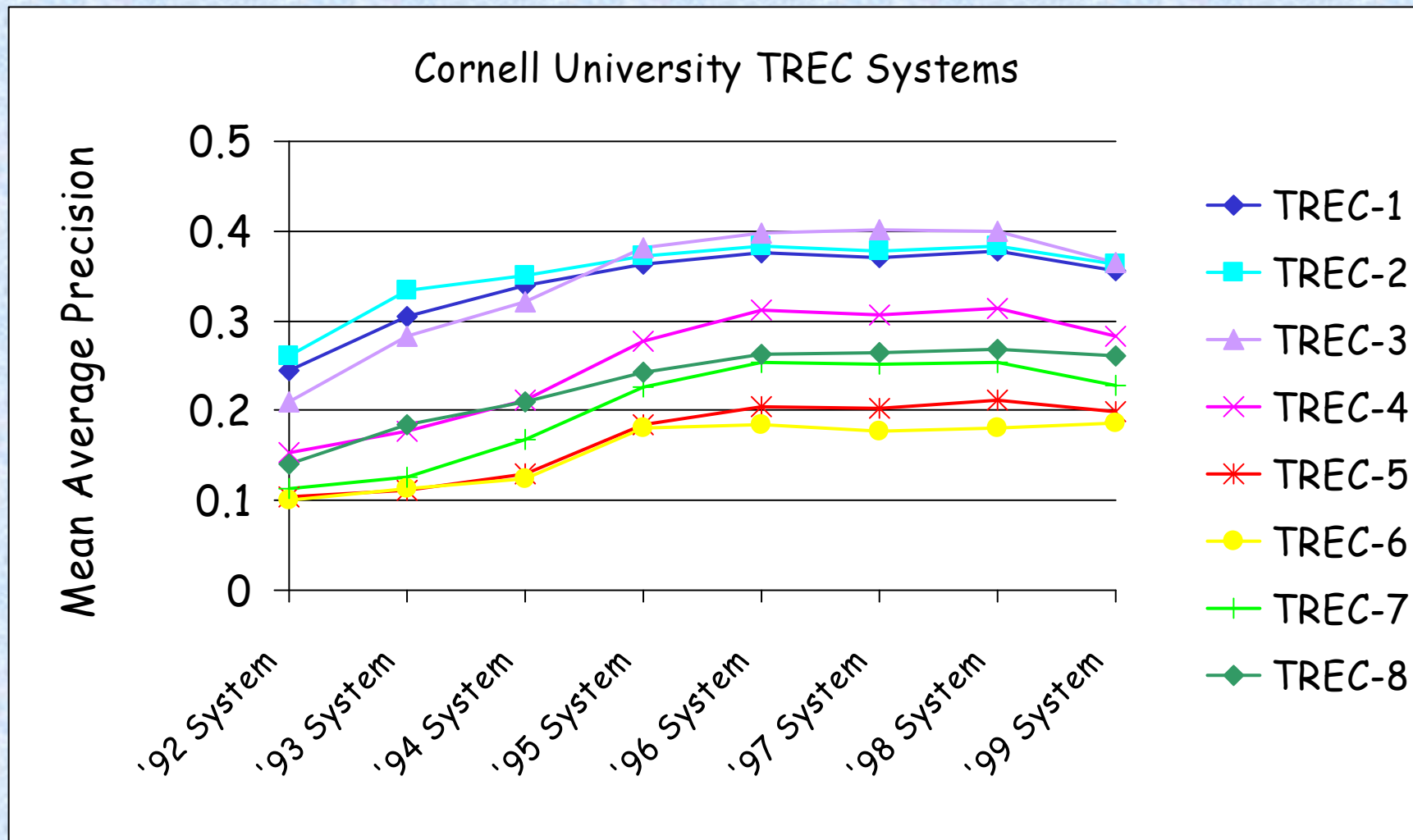
Participation in TREC



TREC Impacts

- Test collections
- Incubator for new research areas
- Common evaluation methodology and improved measures for text retrieval
- Open forum for exchange of research
- Technology transfer

TREC Impacts



Ad Hoc Technologies

	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6	TREC-7
Term weights	baseline start of Okapi wts	Okapi perfects “BM25” algorithm	new wts for SMART, INQUERY, PIRCS	Okapi/ SMART wts used by others	adaptations of Okapi/SMART algorithm in most systems	new Twente and BBN models
Passages	use of subdocs by PIRCS	heavy use of passages/ subdocs	decline in use of passages		use of passages in relevance feedback	multiple uses of passages
Auto query expansion		start of expansion using top X documents	heavy use of expansion using top X documents	start of more complex expansion	more sophisticated expansion experiments by many groups	
Manual query mods		manual expansion using other sources	experiments in manual editing/user- in-the-loop	extensive user-in-the- loop experiments	simpler user-specific strategies tested	
Other new areas		initial use of data fusion		start of concentration on initial topic	more complex use of data fusion continued focus on initial topic, especially the title	

TREC Impacts

- Test collections
 - papers in general IR literature use TREC collections
- Common evaluation methodology and improved measures for text retrieval
 - documents best practices in IR research methodology for new researchers
- Incubator for new research areas
 - PhD theses resulting from CLIR, SDR, QA participation

TREC Impacts

- Open forum for exchange of research
 - TREC papers figure prominently in IR syllabi on the web
 - publication of all results prevents unsuccessful research from being duplicated
- Technology transfer
 - impact is far greater than just those who actually participate

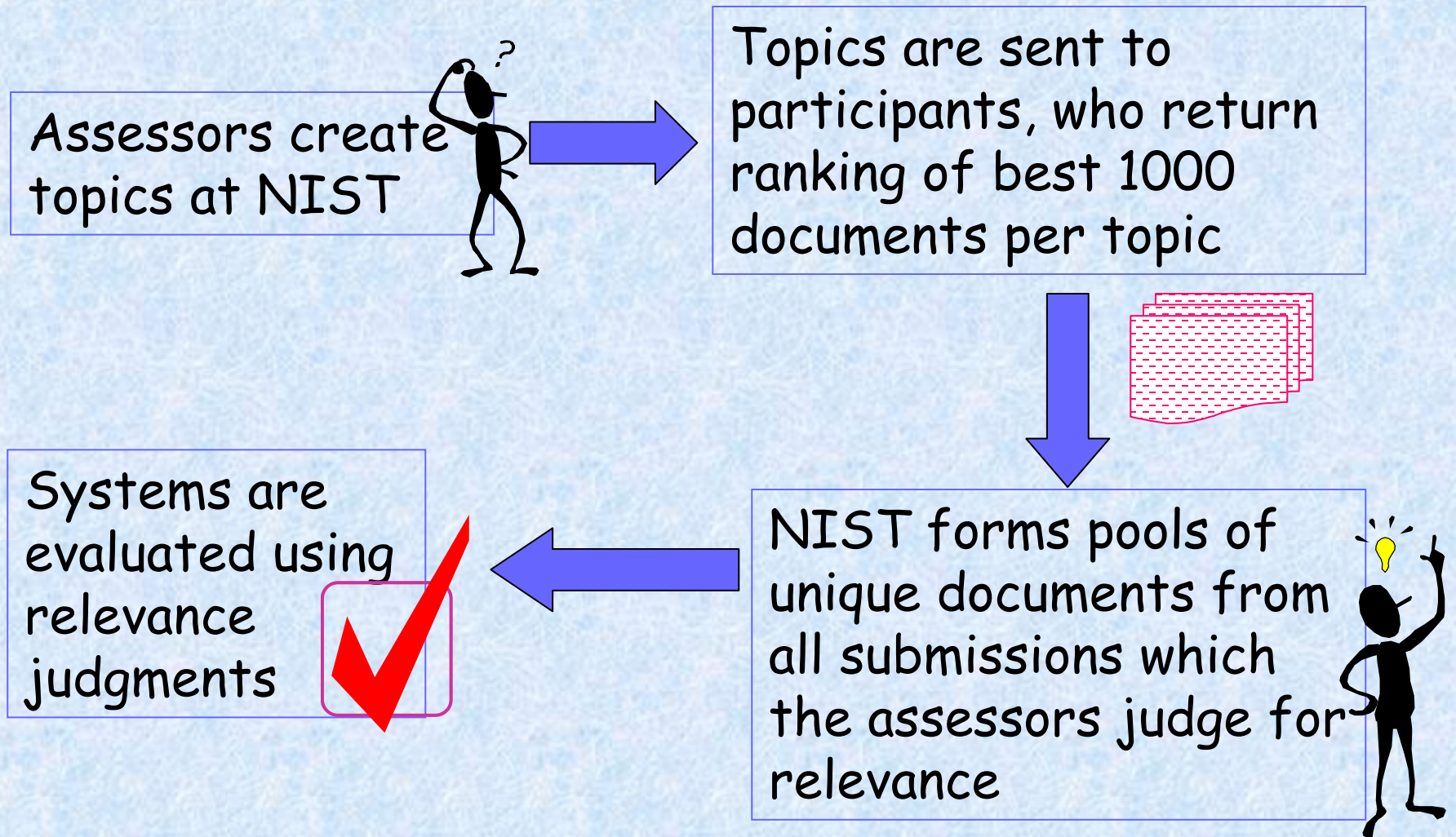
Talk Outline

- General introduction to TREC
 - TREC history
 - TREC impacts
- ➡ Cranfield tradition of laboratory tests
 - mechanics of building test collections
 - test collection quality
 - legitimate uses of test collections
- IR evaluation primer

Cranfield Tradition

- Laboratory testing of system components
 - fine control over variables
 - abstraction from operational setting
 - comparative testing
- Test collections
 - set of documents
 - set of questions
 - relevance judgments

TREC approach



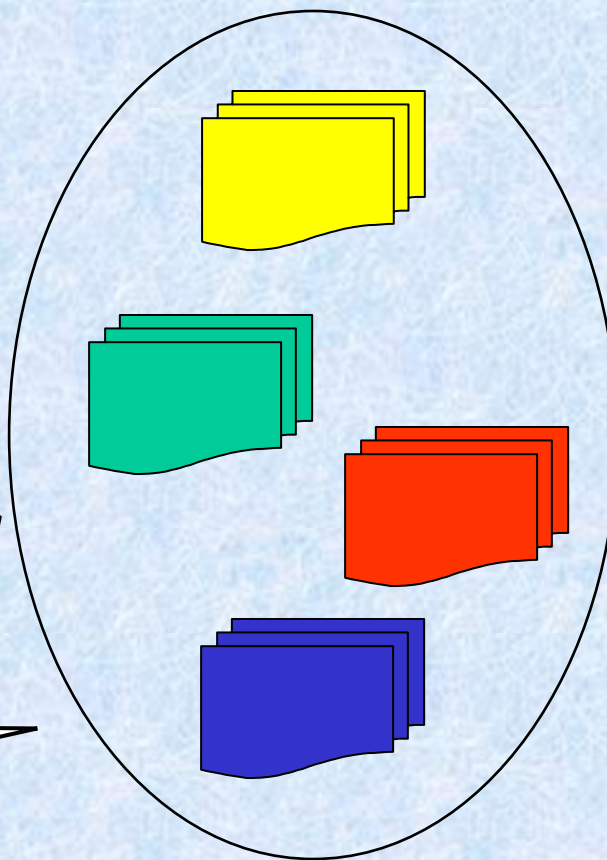
Creating a test collection for an ad hoc task

topic statements

Automatic: no manual intervention

Manual: everything else, including interactive feedback

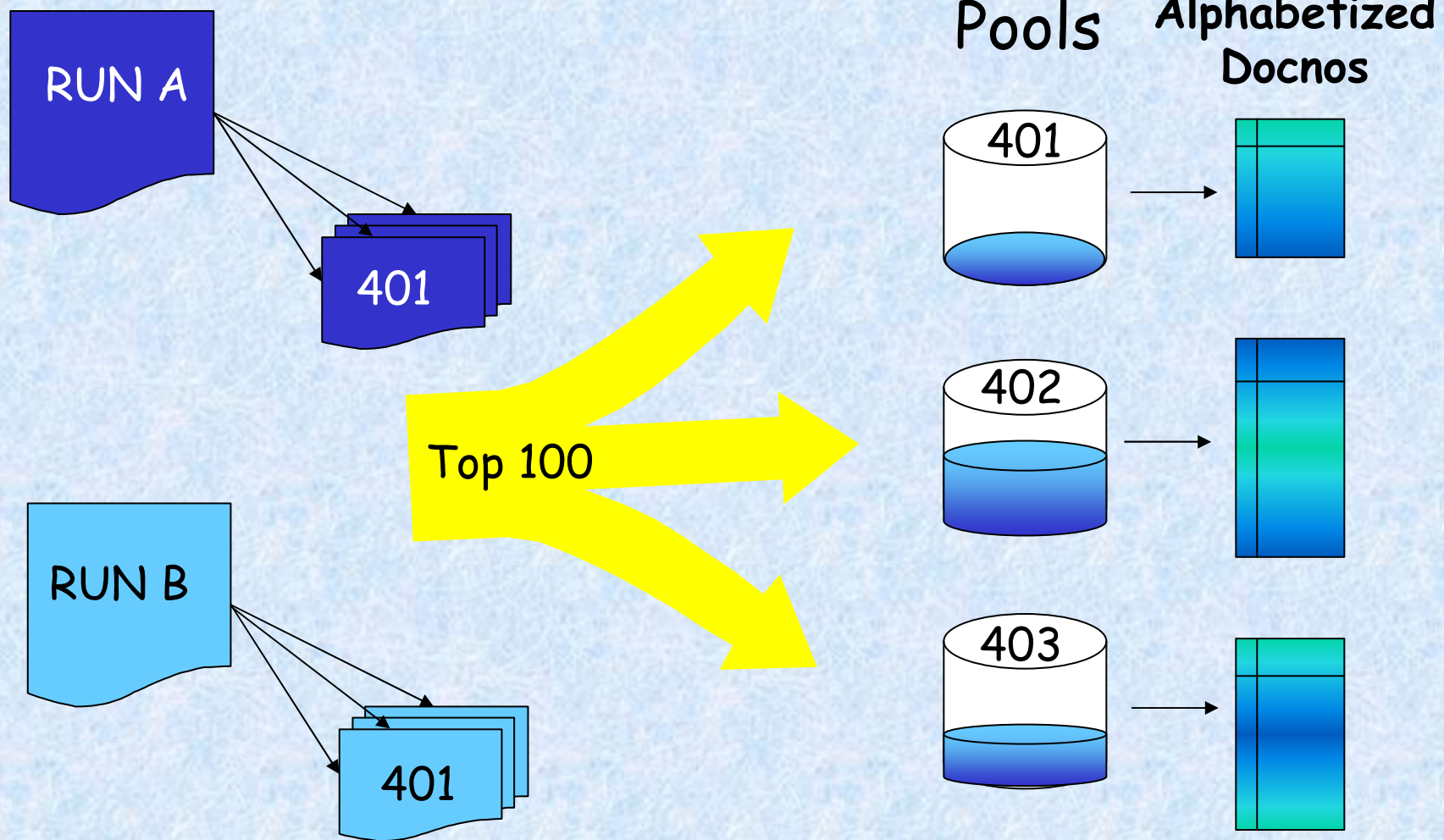
queries



representative
document set

ranked
list

Creating Relevance Judgments



Documents

- Must be representative of real task of interest
 - genre
 - diversity (subjects, style, vocabulary)
 - amount
 - full text vs. abstract
- TREC
 - generally newswire/newspaper
 - general interest topics
 - fulltext

Topics

- Distinguish between stmt of user need (topic) & system data structure (query)
 - topic gives criteria for relevance
 - allows for different query construction techniques
- TREC topics are NOT all created equal
 - 1-150: very detailed, rich content
 - 151-200: method of topic creation resulted in focused, easy topics
 - 201-250: single sentence only
 - 301-450: title is set of hand-picked keywords

Relevance Judgments

- Main source of criticism of Cranfield tradition
 - In test collections, judgments are usually binary, static, and assumed to be complete.
 - But...
 - "relevance" is highly idiosyncratic
 - relevance does not entail utility
 - documents have different degrees of relevance
 - relevance can change over time for the same user
 - for realistic collections, judgments cannot be complete

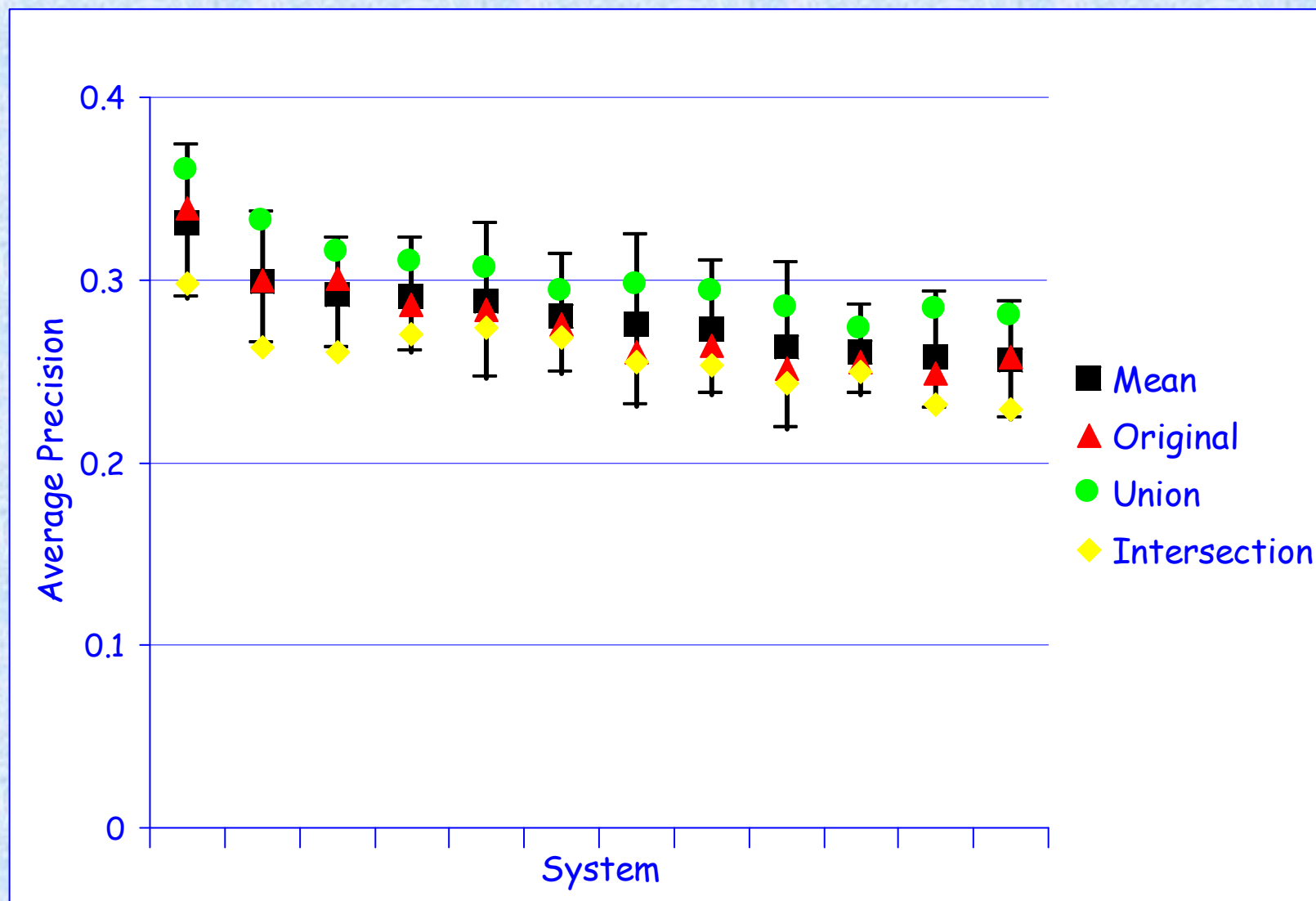
Relevance Judgments

- Consistency
 - idiosyncratic nature of relevance judgments does not affect comparative results
- Incompleteness
 - the important issue is that relevant judgments be unbiased
 - complete judgments must be unbiased
 - TREC pooling has been adequate to produce unbiased judgments (until recently)

Consistency

- Mean Kendall τ between system rankings produced from different qrel sets: .938
- Similar results held for
 - different query sets
 - different evaluation measures
 - different assessor types
 - single opinion vs. group opinion judgments

Average Precision by Qrel



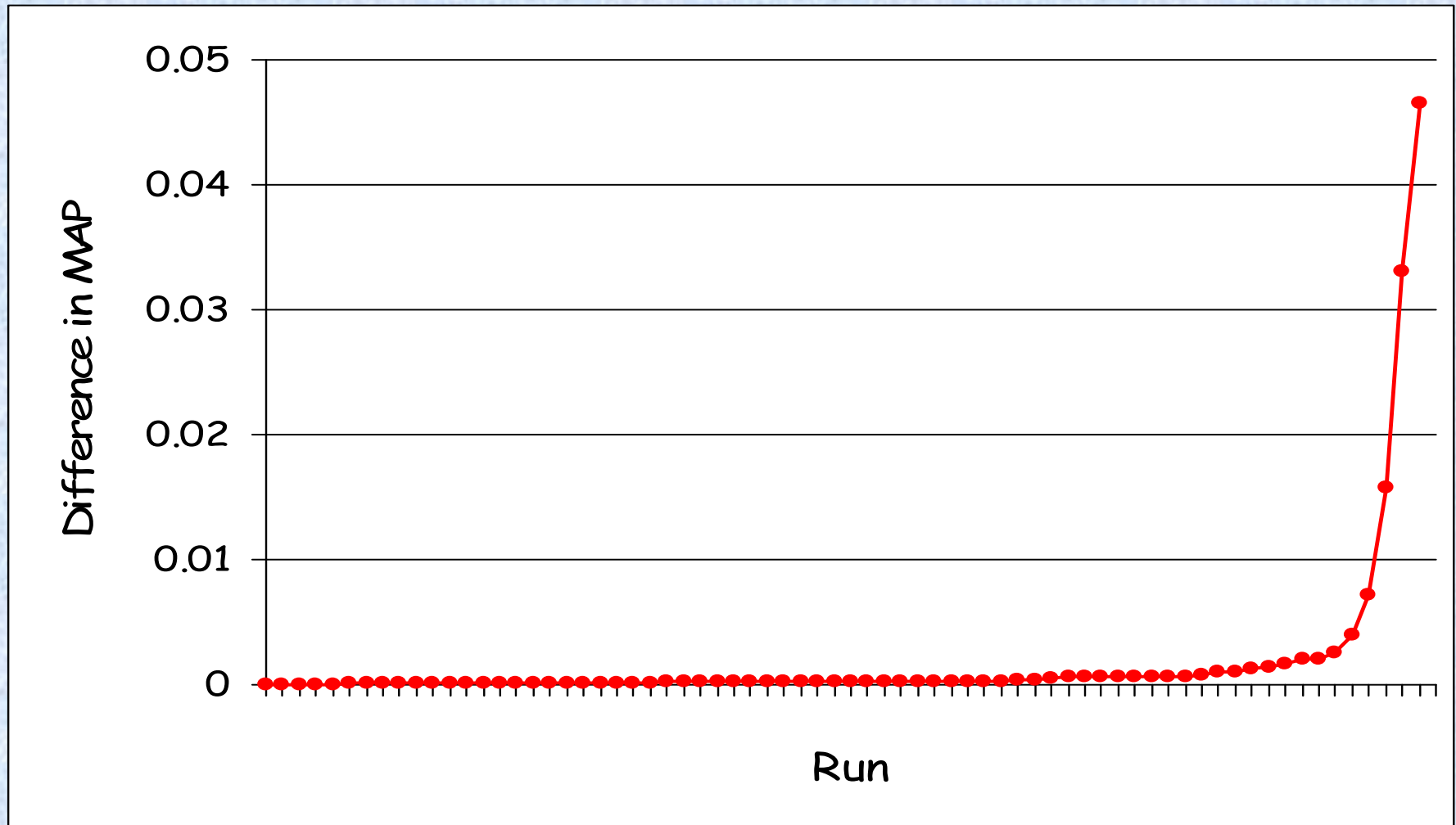
QA Judgments

- Judging correctness, not relevance
- Assessors have differences of opinions as to what constitutes a correct answer
 - granularity of names, dates
 - assumed context
- Comparative evaluation stable despite those differences

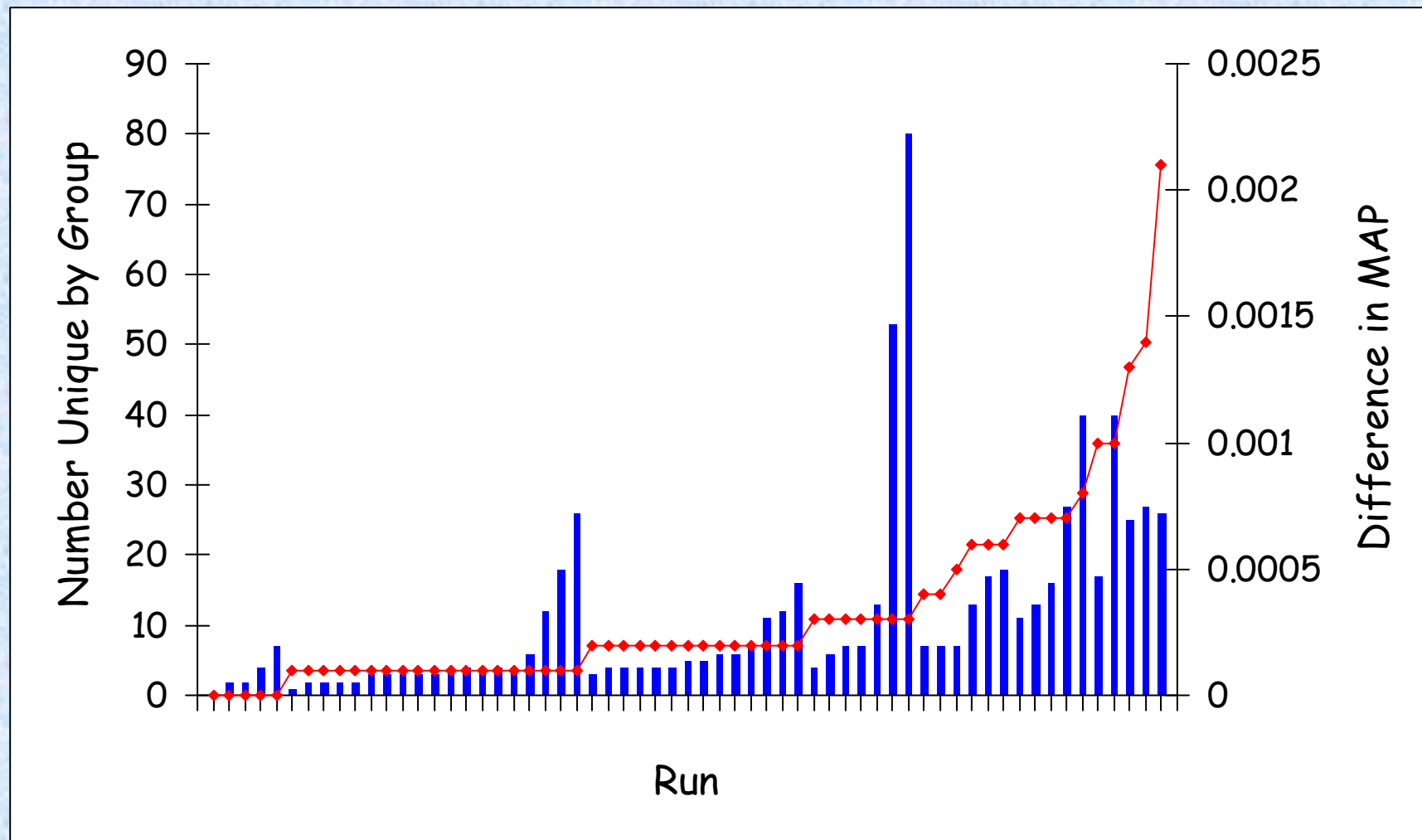
Incompleteness

- Study by Zobel [SIGIR-98]:
 - Quality of relevance judgments does depend on pool depth and diversity
 - TREC ad hoc collections not biased against systems that do not contribute to the pools
 - TREC judgments not complete
 - additional relevant documents distributed roughly uniformly across systems but highly skewed across topics

Uniques Effect on Evaluation



Uniques Effect on Evaluation: Automatic Only



Pool Size

- "Adequate" pool size is proportional to collection size
 - implies that ever larger collections need ever more documents judged
 - infeasible (given resource constraints) for collections the size of the terabyte collection
 - terabyte, million query tracks looking to develop new pooling methodology

Note: collection built in TREC 2005 Robust/HARD tracks also affected; legal track has taken different tack

Cranfield Tradition

- Test collections are abstractions, but laboratory tests are useful nonetheless
 - evaluation technology is predictive (i.e., results transfer to operational settings)
 - different relevance judgments almost always produce the same comparative results
 - adequate pools allow unbiased evaluation of unjudged runs

Cranfield Tradition

- Note the emphasis on comparative !!
 - absolute value of effectiveness measures not meaningful
 - absolute value changes as relevance judgments change
 - theoretical maximum of 1.0 for both recall and precision not obtainable by humans (inter-assessor judgments suggest 65% precision at 65% recall)
 - evaluation results are only comparable when they are from the same collection
 - a subset of a collection is a different collection
 - comparisons between different TREC collections are invalid

Sensitivity Analysis

- With archive of TREC results, have empirically determine relationship between number of topics, Δ of scores, & error rate [Voorhees & Buckley, 2002]
 - error rates generally larger than accounted for in literature
 - confidence increases with topic set size
 - confidence also increases with larger Δ , but then power of comparison reduced
 - confidence can be increased by repeating experiment on multiple collections

Talk Outline

- General introduction to TREC
 - TREC history
 - TREC impacts
- Cranfield tradition of laboratory tests
 - mechanics of building test collections
 - test collection quality
 - legitimate uses of test collections

➡ IR evaluation primer

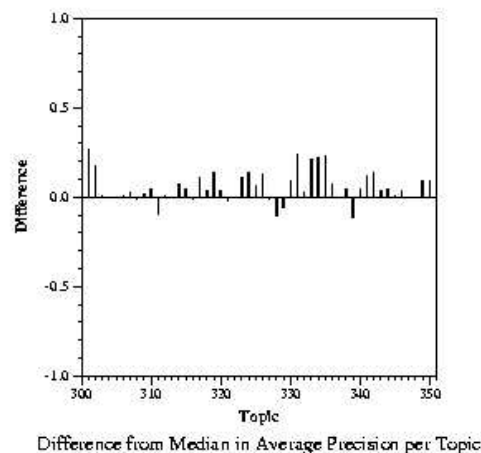
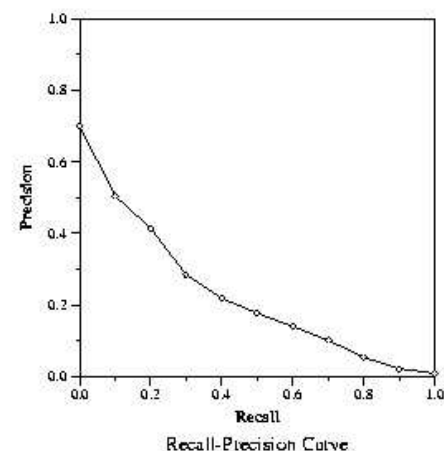
trec_eval Evaluation Report

Ad hoc results — Cornell University

Summary Statistics	
Run Number	Cor6A3cdl
Run Description	Category A, Automatic, long
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4611
Rel-ret:	2590

Recall Level Precision Averages	
Recall	Precision
0.00	0.7013
0.10	0.5050
0.20	0.4150
0.30	0.2846
0.40	0.2187
0.50	0.1775
0.60	0.1402
0.70	0.1015
0.80	0.0538
0.90	0.0224
1.00	0.0091
Average precision over all relevant docs	
non-interpolated	0.2139

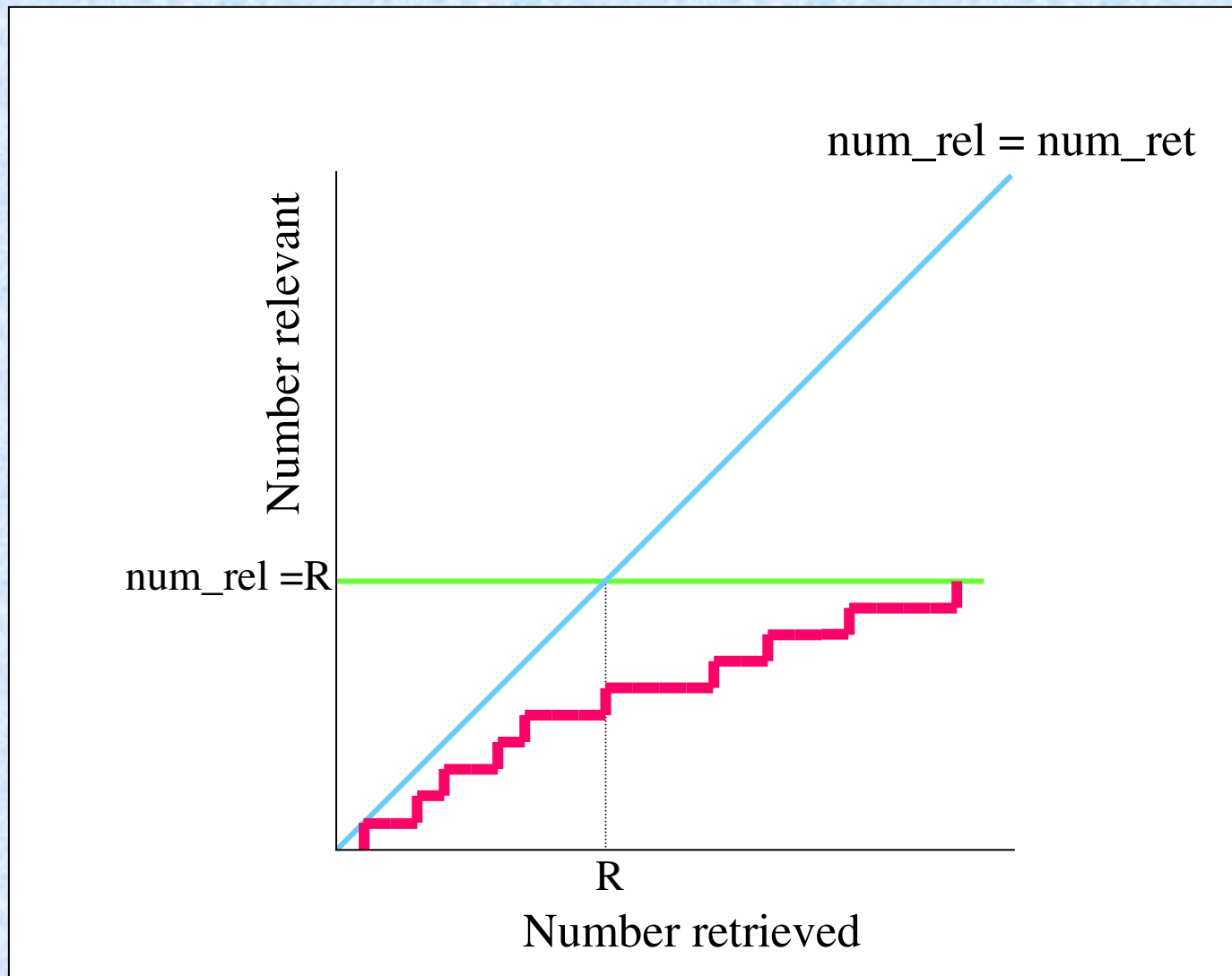
Document Level Averages	
	Precision
At 5 docs	0.4480
At 10 docs	0.4260
At 15 docs	0.4013
At 20 docs	0.3630
At 30 docs	0.3200
At 100 docs	0.2010
At 200 docs	0.1418
At 500 docs	0.0823
At 1000 docs	0.0518
R-Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.2415



Evaluation Measure Criteria

- Related to a user satisfaction
- Interpretable
- Able to average or collect
- Have high discrimination power
- Able to be analyzed

Ranked Retrieval Chart



Evaluation Contingency Table

	Relevant	Non-Relevant
Retrieved	r	$n-r$
Non-Retrieved	$R-r$	$N-n-R+r$

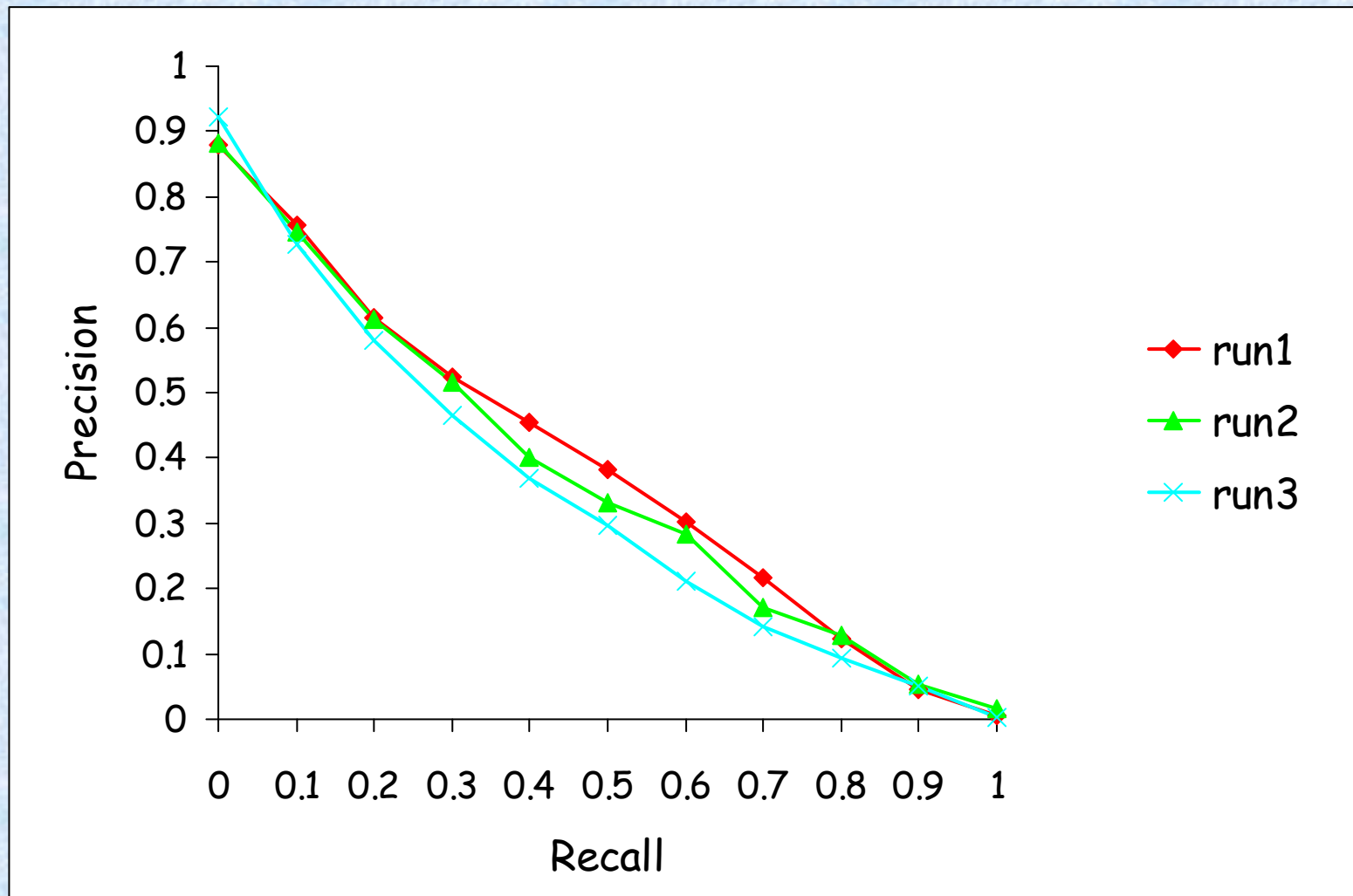
N = number docs in collection

n = number docs retrieved

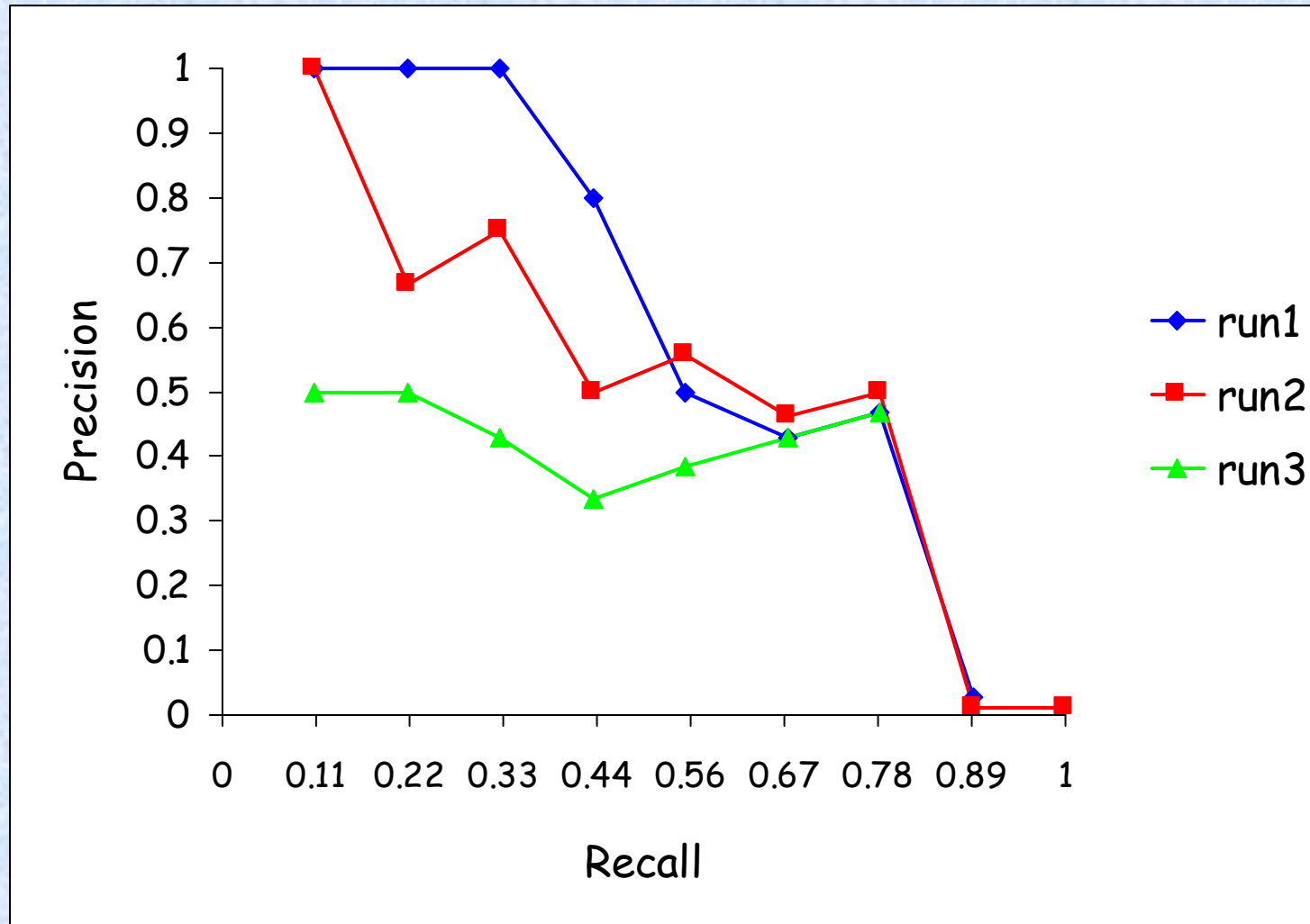
R = number relevant docs

r = number relevant retrieved

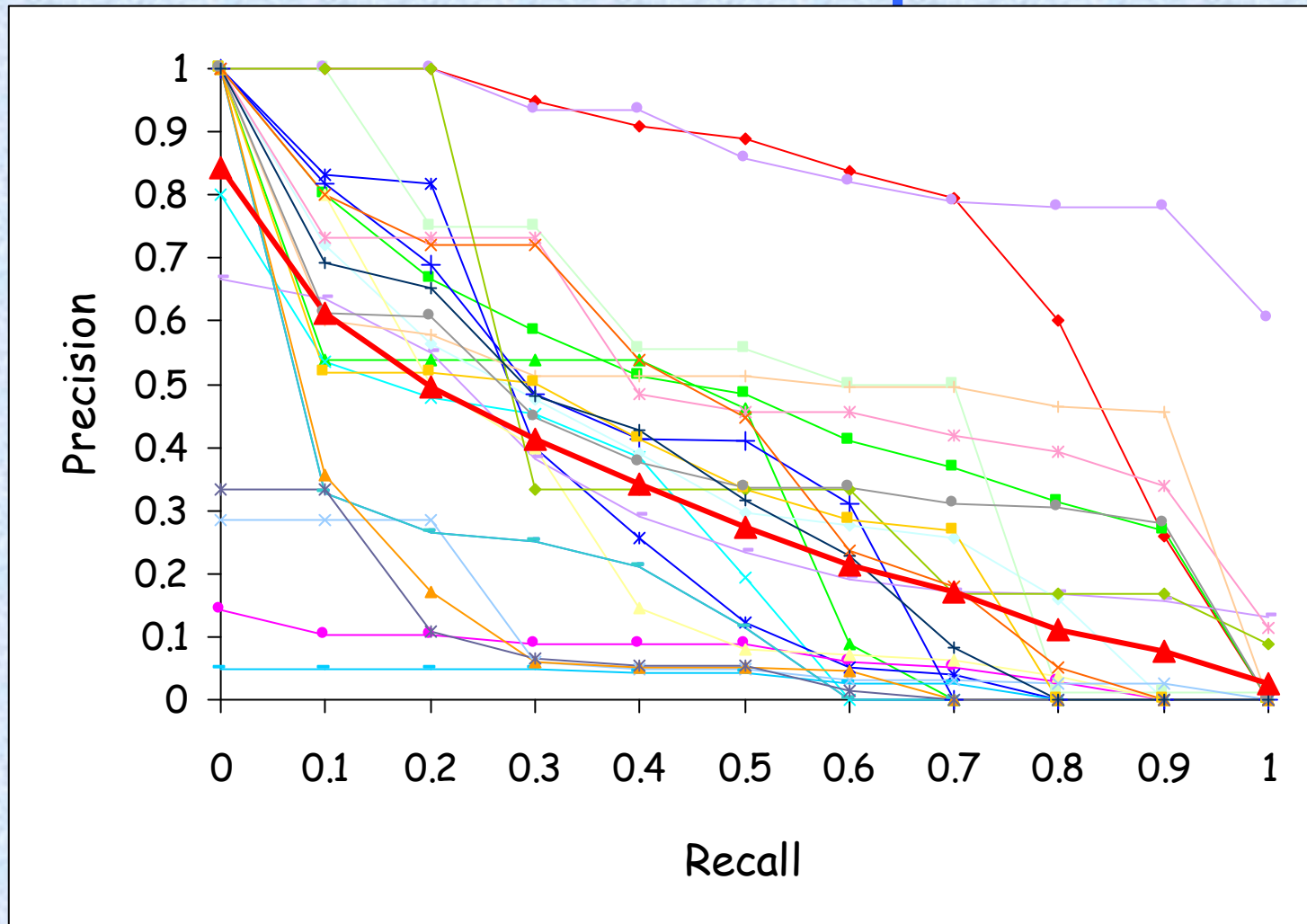
Recall-Precision Graph



Uninterpolated R-P Curve for Single topic



Interpolated R-P Curves for Individual Topics



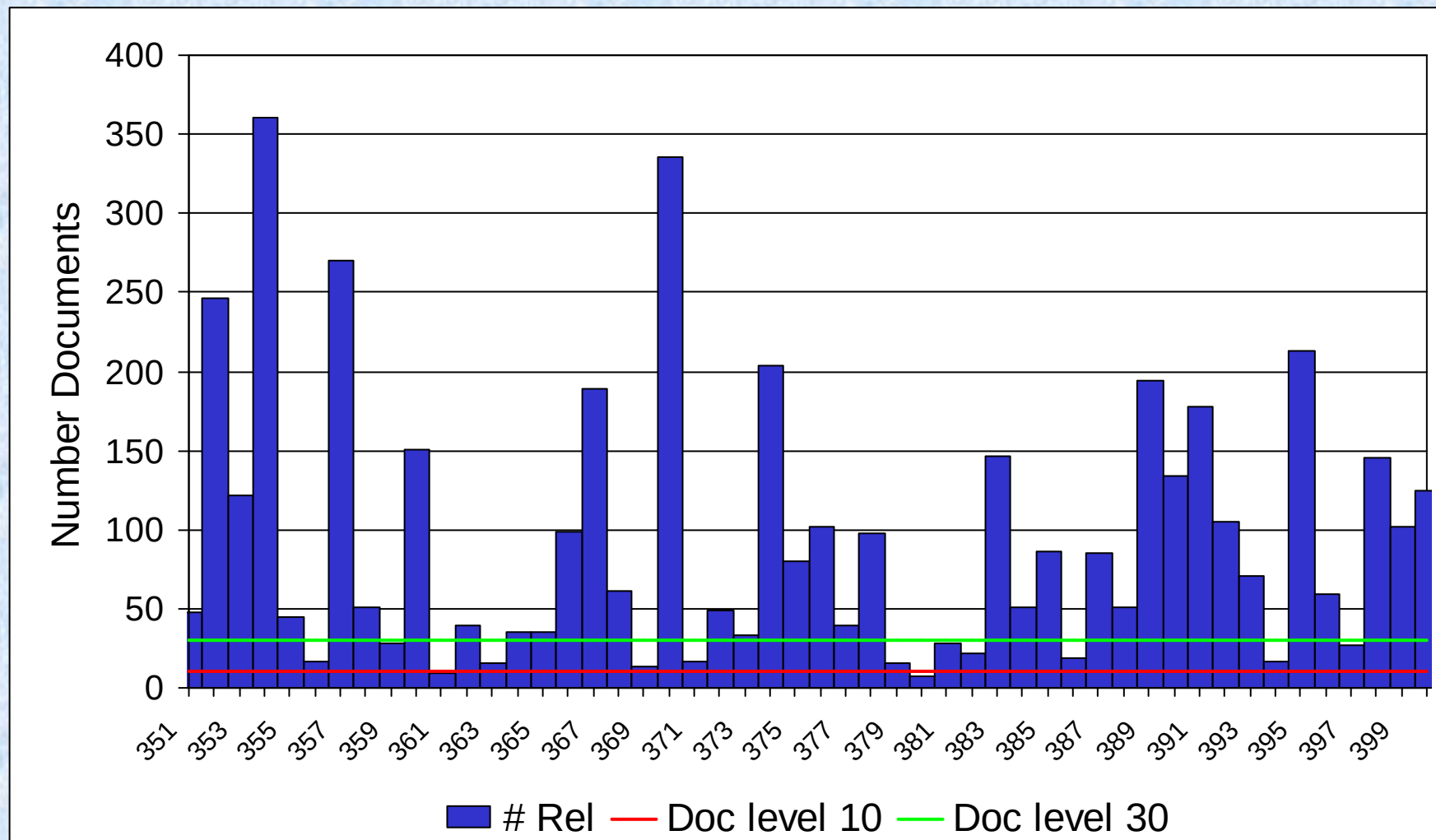
Single Number Summary Scores

- Precision (n): r / n
- Recall(n): r / R
- Average precision: $\text{Avg}_{rd} (\text{Prec}(\text{rank of } rd))$
- R-Precision: $\text{Prec}(R)$
- Recall at .5 precision
 - use $\text{Prec}(10)$ if precision $< .5$ in top 10
- Rank of first relevant (expected search length)

Document Level Measures

- Advantage
 - immediately interpretable
- Disadvantages
 - don't average well
 - different number of relevant implies topics are in different parts of recall-precision curve
 - theoretical maximums impossible to reach
 - insensitive to ranking: only # rels that cross cut-off affect ranking
 - less useful for tuning a system

Number Relevant



Average Precision

- Advantages
 - sensitive to entire ranking: changing a single rank will change final score
 - stable: a small change in ranking makes a relatively small change in score
 - has both precision- and recall-oriented factors
 - ranks closest to 1 receive largest weight
 - computed over all relevant documents
- Disadvantages
 - less easily interpreted

Runs Ranked by Different Measures

P(10)	P(30)	R-Prec	Ave Prec	Recall at .5 Prec	Recall (1000)	Total Rel	Rank of 1 st Rel
INQ502 ok7ax att98atdc att98atde INQ501 nect'chall nect'chdes ok7am mds98td INQ503 Cor7A3rrf tno7tw4 MerAbtnd acsys7al iowacuhk1	INQ502 ok7ax INQ501 att98atdc nect'chall att98atde ok7am nect'chdes INQ503 bbn1 tno7exp1 mds98td pirc8Aa2 Cor7A3rrf ok7as	ok7ax INQ502 ok7am att98atdc att98atde INQ501 bbn1 mds98td nect'chdes nect'chall ok7as tno7exp1 acsys7al pirc8Aa2 Cor7A3rrf	ok7ax att98atdc att98atde ok7am INQ502 mds98td bbn1 tno7exp1 INQ501 pirc8Aa2 Cor7A3rrf acsys7al ok7as nect'chdes nect'chall	att98atdc ok7ax mds98td ok7am INQ502 att98atde INQ501 ok7as bbn1 nect'chall tno7exp1 Cor7A3rrf acsys7al Cor7A2rrd INQ503	ok7ax tno7exp1 att98atdc att98atde Cor7A3rrf ok7am bbn1 pirc8Aa2 INQ502 pirc8Ad INQ501 nect'chdes nect'chall acsys7al mds98td	ok7ax tno7exp1 att98atdc bbn1 att98atde INQ502 INQ501 ok7am Cor7A3rrf pirc8Aa2 nect'chdes mds98td acsys7al nect'chall pirc8Ad	tno7tw4 bbn1 INQ502 nect'chall tnocbm25 MerAbtnd att98atdc acsys7al mds98td ibms98a Cor7A3rrf ok7ax att98atde Brkly25 nect'chdes

Ranked by measure averaged over 50 topics

Text REtrieval Conference (TREC)

Correlations Between Rankings

	P(30)	R Prec	Ave Prec	Recall at .5 P	Recall (1000)	Total Rels	Rank 1 st Rel
P(10)	.8851	.8151	.7899	.7855	.7817	.7718	.6378
P(30)		.8676	.8446	.8238	.7959	.7915	.6213
R Prec			.9245	.8654	.8342	.8320	.5896
Ave Prec				.8840	.8473	.8495	.5612
R at .5 P					.7707	.7762	.5349
Recall(1000)						.9212	.5891
Total Rels							.5880

Kendall's τ computed between pairs of rankings

Binary Preference (bpref)

- Measure designed to be robust in the face of incomplete judgments
 - tracks MAP very closely with full judgments
 - more stable than MAP when judgments noticeably incomplete
 - [Buckley and Voorhees, SIGIR 2004]
- Makes use of judged documents only
 - normalized count of the times known relevant documents are ranked before known nonrelevant

$$\frac{1}{R} \sum_r \left(1 - \frac{\ln \text{ranked before } rl}{\min(R, N)} \right)$$

Summary

- TREC emphasizes individual experiments evaluated on a benchmark task
 - leverages modest government investment into substantially more R&D than could be funded directly
 - improves state-of-the-art
 - accelerates technology transfer