

Overview of TREC 2003



Sponsored by:
NIST, ARDA, DARPA

Ellen Voorhees

NIST

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Text REtrieval Conference (TREC)

TREC 2002 Program Committee

Ellen Voorhees, chair

James Allan

Nick Belkin

Chris Buckley

Jamie Callan

Gord Cormack

Sue Dumais

Fred Gey

Donna Harman

Dave Hawking

Bill Hersh

Jim Mayfield

John Prange

Steve Robertson

Karen Sparck Jones

Ross Wilkinson

TREC 2003 Track Coordinators

Genomics: Bill Hersh

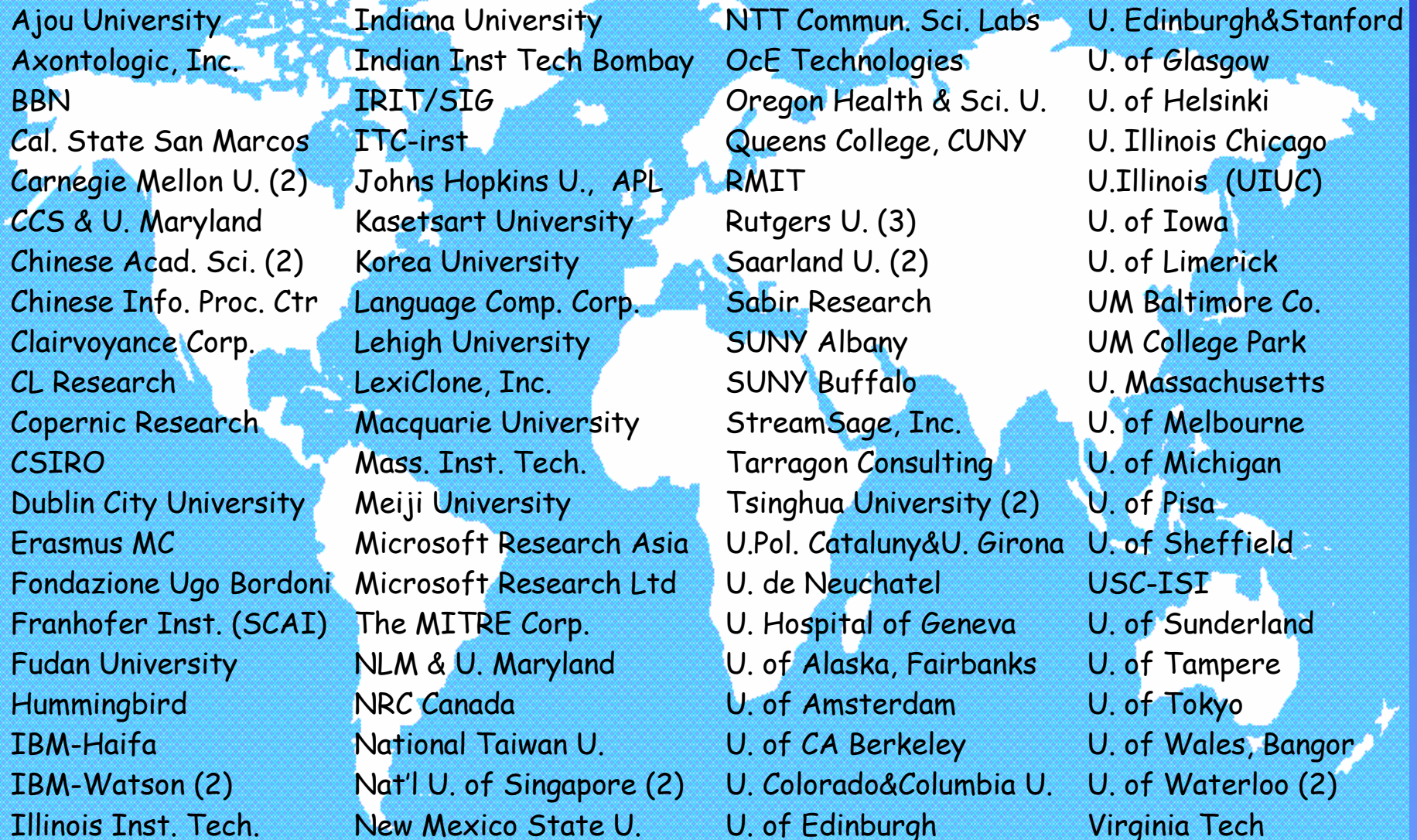
HARD: James Allan

Novelty: Ian Soboroff, Donna Harman

Question Answering: Ellen Voorhees

Robust Retrieval: Ellen Voorhees

Web: David Hawking, Nick Craswell, Ian Soboroff

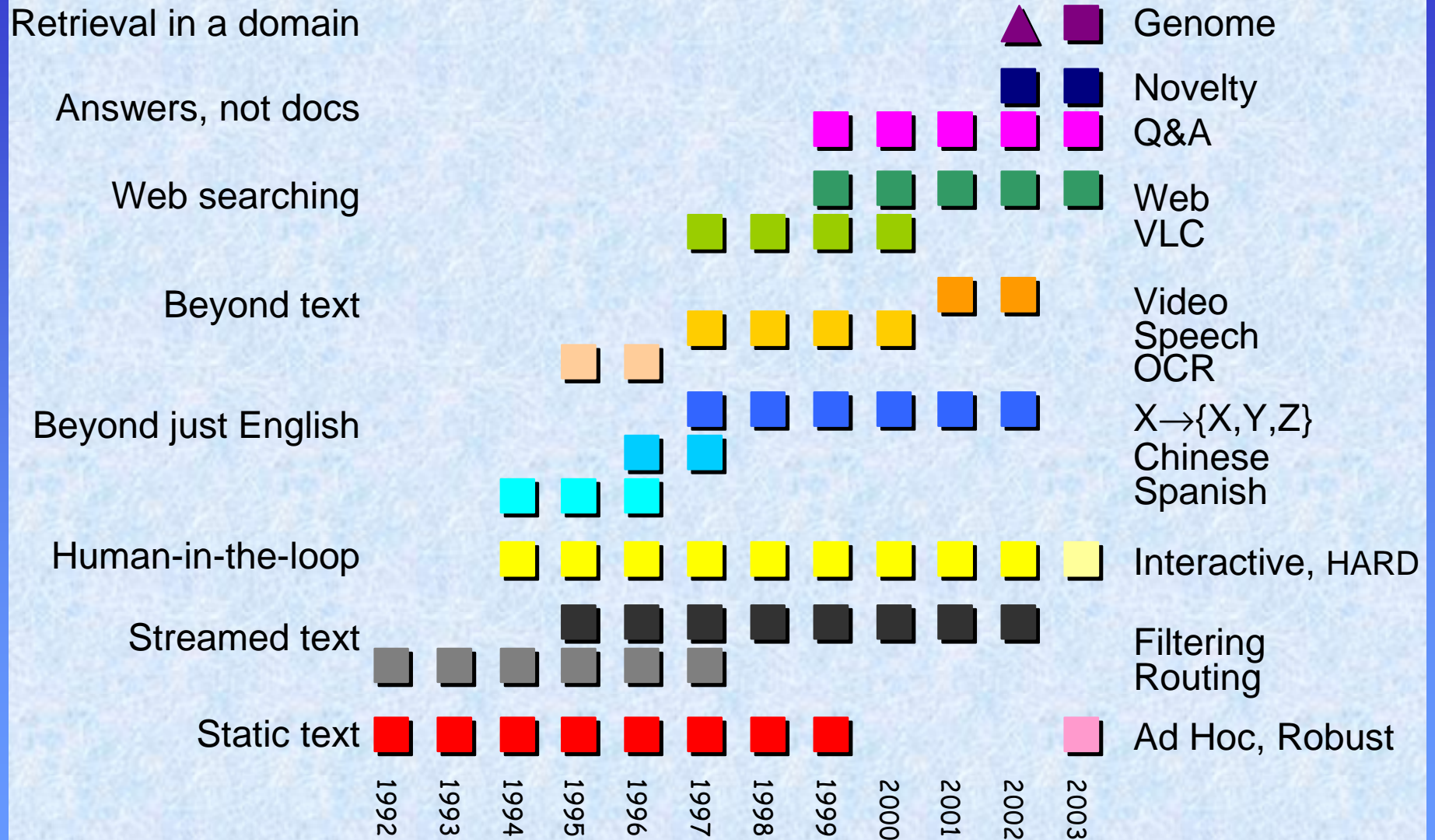


Ajou University	Indiana University	NTT Commun. Sci. Labs	U. Edinburgh&Stanford
Axontologic, Inc.	Indian Inst Tech Bombay	OcE Technologies	U. of Glasgow
BBN	IRIT/SIG	Oregon Health & Sci. U.	U. of Helsinki
Cal. State San Marcos	ITC-irst	Queens College, CUNY	U. Illinois Chicago
Carnegie Mellon U. (2)	Johns Hopkins U., APL	RMIT	U.Illinois (UIUC)
CCS & U. Maryland	Kasetsart University	Rutgers U. (3)	U. of Iowa
Chinese Acad. Sci. (2)	Korea University	Saarland U. (2)	U. of Limerick
Chinese Info. Proc. Ctr	Language Comp. Corp.	Sabir Research	UM Baltimore Co.
Clairvoyance Corp.	Lehigh University	SUNY Albany	UM College Park
CL Research	LexiClone, Inc.	SUNY Buffalo	U. Massachusetts
Copernic Research	Macquarie University	StreamSage, Inc.	U. of Melbourne
CSIRO	Mass. Inst. Tech.	Tarragon Consulting	U. of Michigan
Dublin City University	Meiji University	Tsinghua University (2)	U. of Pisa
Erasmus MC	Microsoft Research Asia	U.Pol. Cataluny&U. Girona	U. of Sheffield
Fondazione Ugo Bordoni	Microsoft Research Ltd	U. de Neuchatel	USC-ISI
Franhofer Inst. (SCAI)	The MITRE Corp.	U. Hospital of Geneva	U. of Sunderland
Fudan University	NLM & U. Maryland	U. of Alaska, Fairbanks	U. of Tampere
Hummingbird	NRC Canada	U. of Amsterdam	U. of Tokyo
IBM-Haifa	National Taiwan U.	U. of CA Berkeley	U. of Wales, Bangor
IBM-Watson (2)	Nat'l U. of Singapore (2)	U. Colorado&Columbia U.	U. of Waterloo (2)
Illinois Inst. Tech.	New Mexico State U.	U. of Edinburgh	Virginia Tech

TREC Goals

- To increase research in information retrieval based on large-scale collections
- To provide an open forum for exchange of research ideas to increase communication among academia, industry, and government
- To facilitate technology transfer between research labs and commercial products
- To improve evaluation methodologies and measures for information retrieval
- To create a series of test collections covering different aspects of information retrieval

TREC Tracks

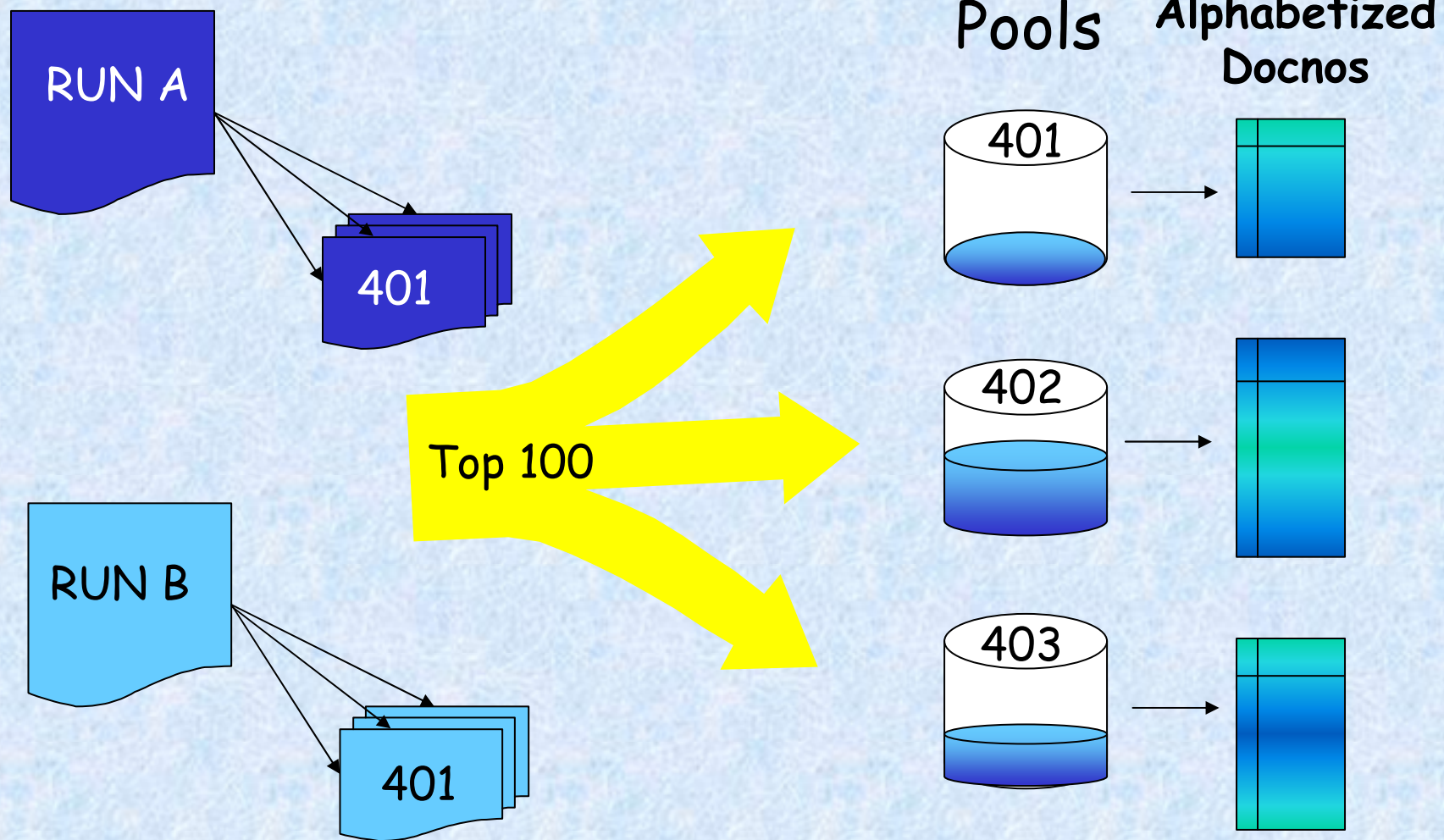


Text REtrieval Conference (TREC)

Common Terminology

- "Document" broadly interpreted
 - page in a web search
 - MEDLINE record in genomics track
- Different types of tasks
 - ad hoc search
 - known-item search
 - answer extraction

Creating Relevance Judgments





Text REtrieval Conference (TREC)

TREC 2003 Tracks

- Genomics
 - primary, secondary
- HARD
- Novelty
 - tasks 1-4
- Question Answering
 - main, passages
- Robust Retrieval
- Web
 - topic distillation, navigational, interactive

Genomics Track

- New track for 2003
 - "pre-track" in TREC 2002
 - first year of a 5-year plan
- Motivation: explore retrieval in a domain
- Two tasks
 - primary: ad hoc task of MEDLINE records
 - secondary: information extraction task to locate descriptive text

Genomics Resources

- MEDLINE

- NLM bibliographic database of medical literature
- a record consists of bibliographic data about an article, the abstract, and MESH headings

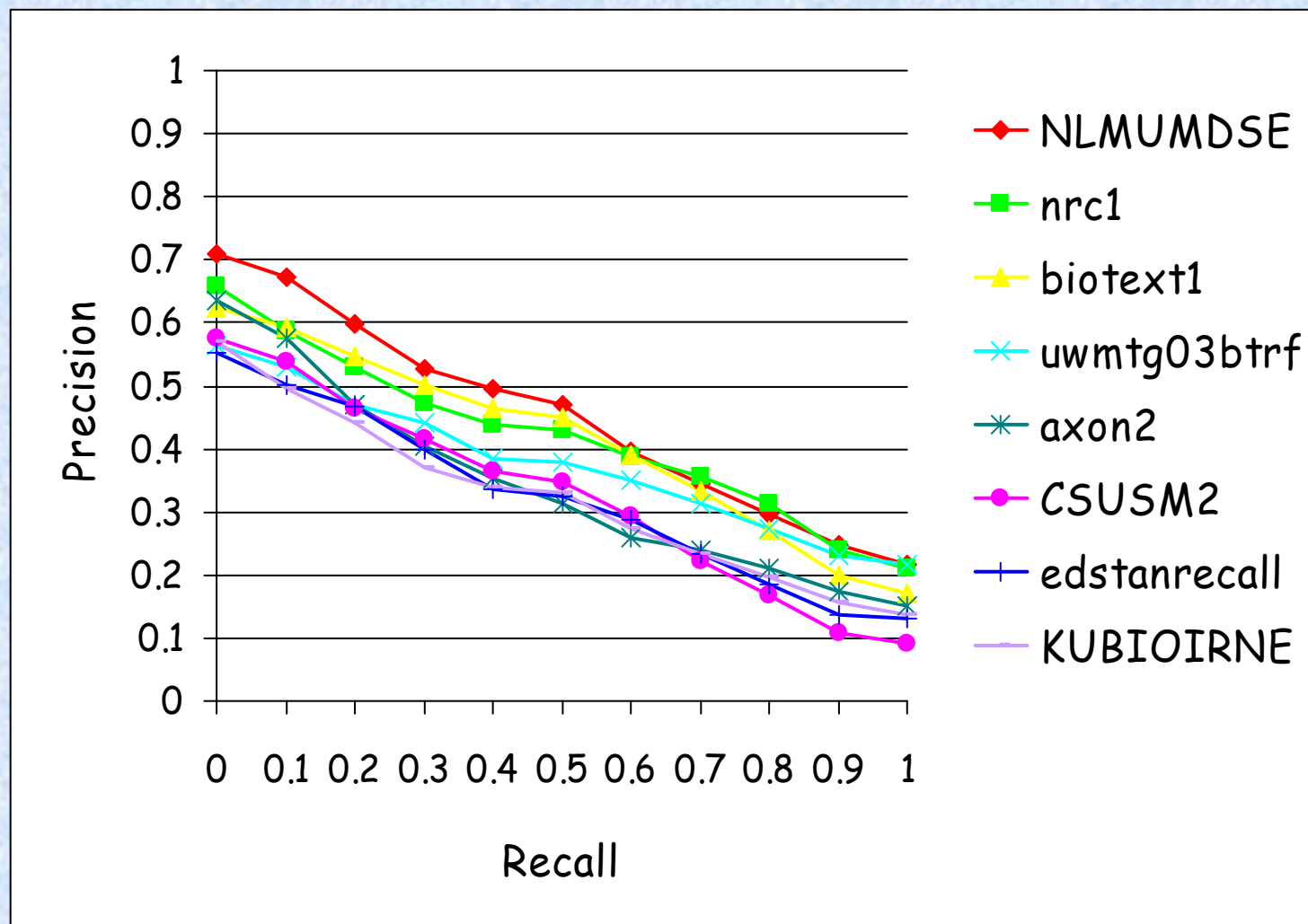
- GeneRIF

- Gene Reference into Function
- consists of a statement about the function of a gene, plus a pointer to the MEDLINE record that documents that function
 - e.g.: *"role in potentiating hematopoietic cell migration"*
- part of LocusLink, a database for biotechnology information

Primary Task

- Documents
 - 525,938 MEDLINE records indexed between April 1, 2002 and April 1, 2003
 - provided to the track by NLM
- Topics
 - 50 gene names
 - interpreted as "find MEDLINE records that focus on the basic biology of the gene or its protein products in the designated organism"
- Relevance judgments
 - GeneRIF data used as surrogate judgments

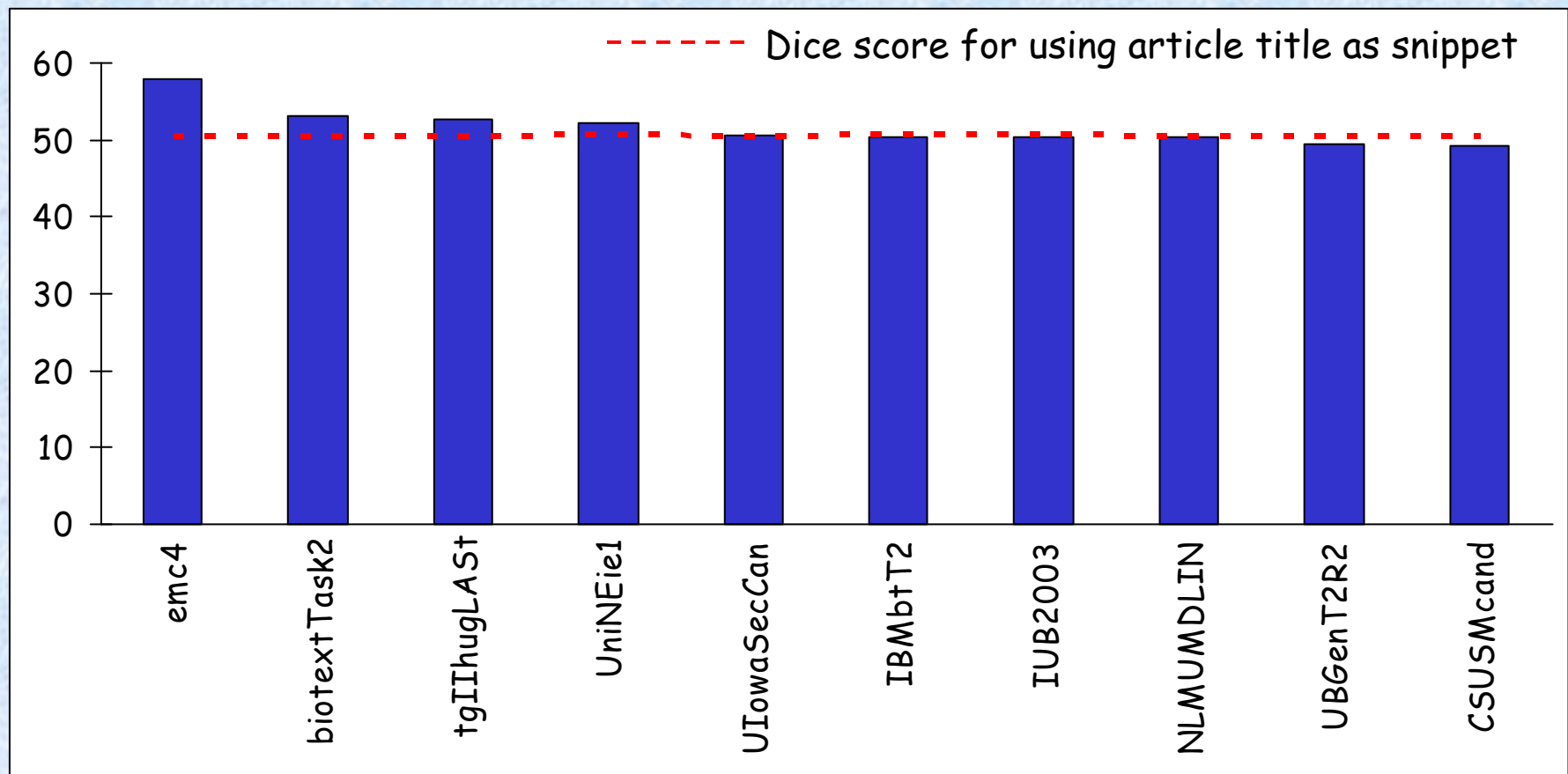
Top Primary Task Runs



Secondary Task

- Information extraction task
 - given full text of article, extract its GeneRIF statement
 - article text obtained Highwire Press
 - 139 articles in test set
- Evaluation
 - overlap between system's candidate statement and actual GeneRIF statement
 - 4 variants of Dice coefficient to measure overlap
 - preliminary analysis suggested 95% of actual GeneRIF statements contain some text from title or abstract

Secondary Task Results



Best run per group for classic Dice score

HARD Track

- High Accuracy Retrieval from Documents
- New track for 2003
- Goal: improve ad hoc retrieval by customizing the search to the user
 - current systems return results for "average" user
 - necessarily limits effectiveness of system for particular user
- Ad hoc task with metadata
 - metadata gathered at topic creation
 - metadata collected from *clarifying form*

HARD Collection

- Documents
 - 1999 subset of AQUAINT collection
 - 1999 government docs (FR, CR)
 - 372,219 documents; 1.7GB text
- Topics
 - created by LDC using standard protocol
 - captured additional metadata
- Relevance Judgments
 - binary judgments by topic author
 - SOFT-rel: on-topic, but constraints not satisfied
 - passages: selected relevant document extracts

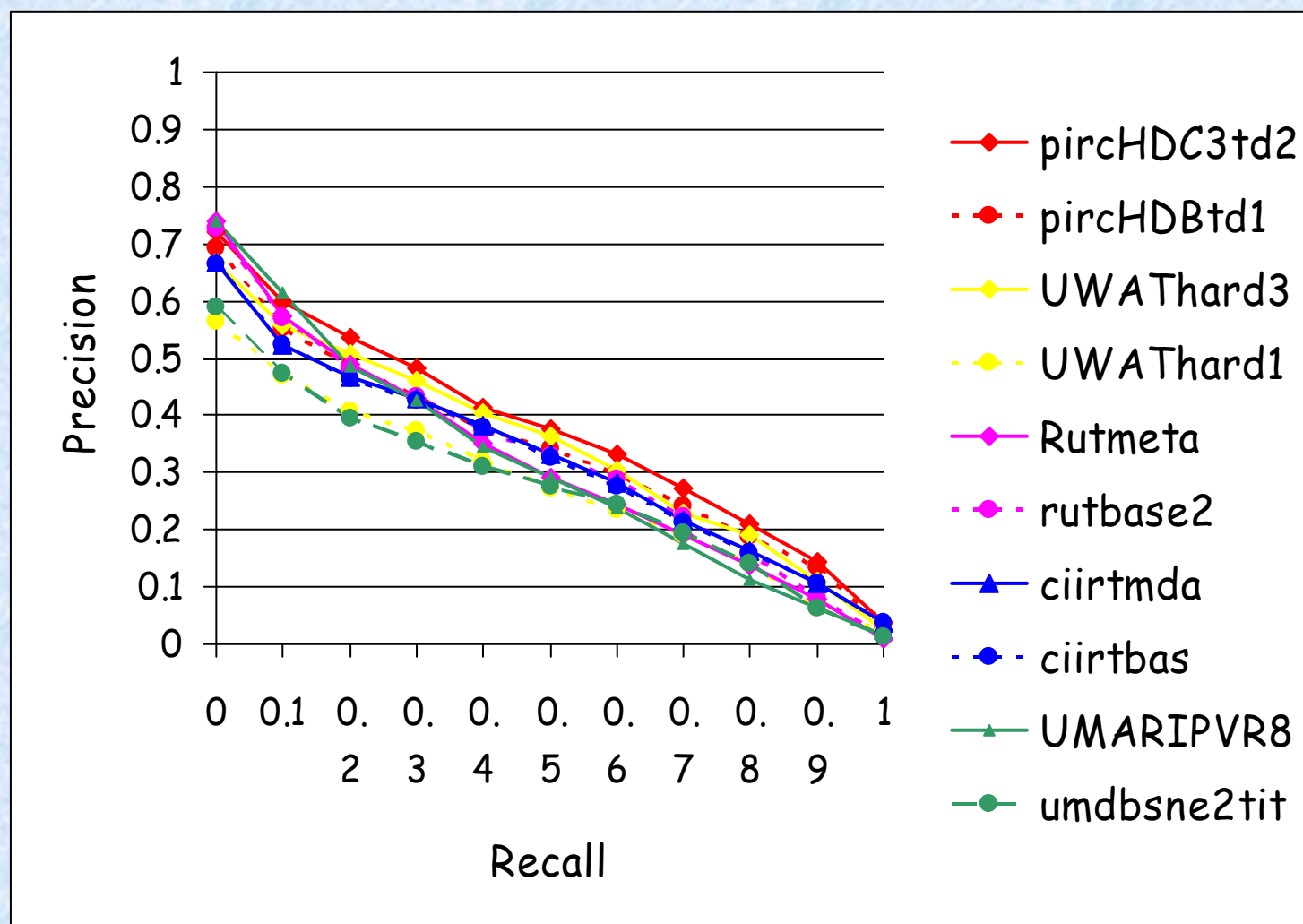
Metadata

- Metadata from topic statements
 - the purpose of the search
 - the genre of the desired response
 - the user's familiarity with the subject matter
 - the granularity of the desired response
 - biographical data about user (age, sex, etc.)
- Clarifying forms
 - assessor (surrogate user) spends at most 3 minutes/topic responding to topic-specific form
 - example uses:
 - sense resolution
 - relevance judgments

HARD Evaluation

- Document-based
 - standard trec_eval evaluation
 - two evaluation conditions: SOFT-rel documents relevant & SOFT-rel documents not relevant
- Passage-based
 - restrict to cut-off based evaluation measures
 - documents treated as (potentially long) passages
 - recall and precision based on character positions
 - recall: proportion of relevant characters retrieved
 - precision: proportion of retrieved characters relevant
 - when retrieved and relevant granularity is documents, recall is equivalent to document-based, precision is not

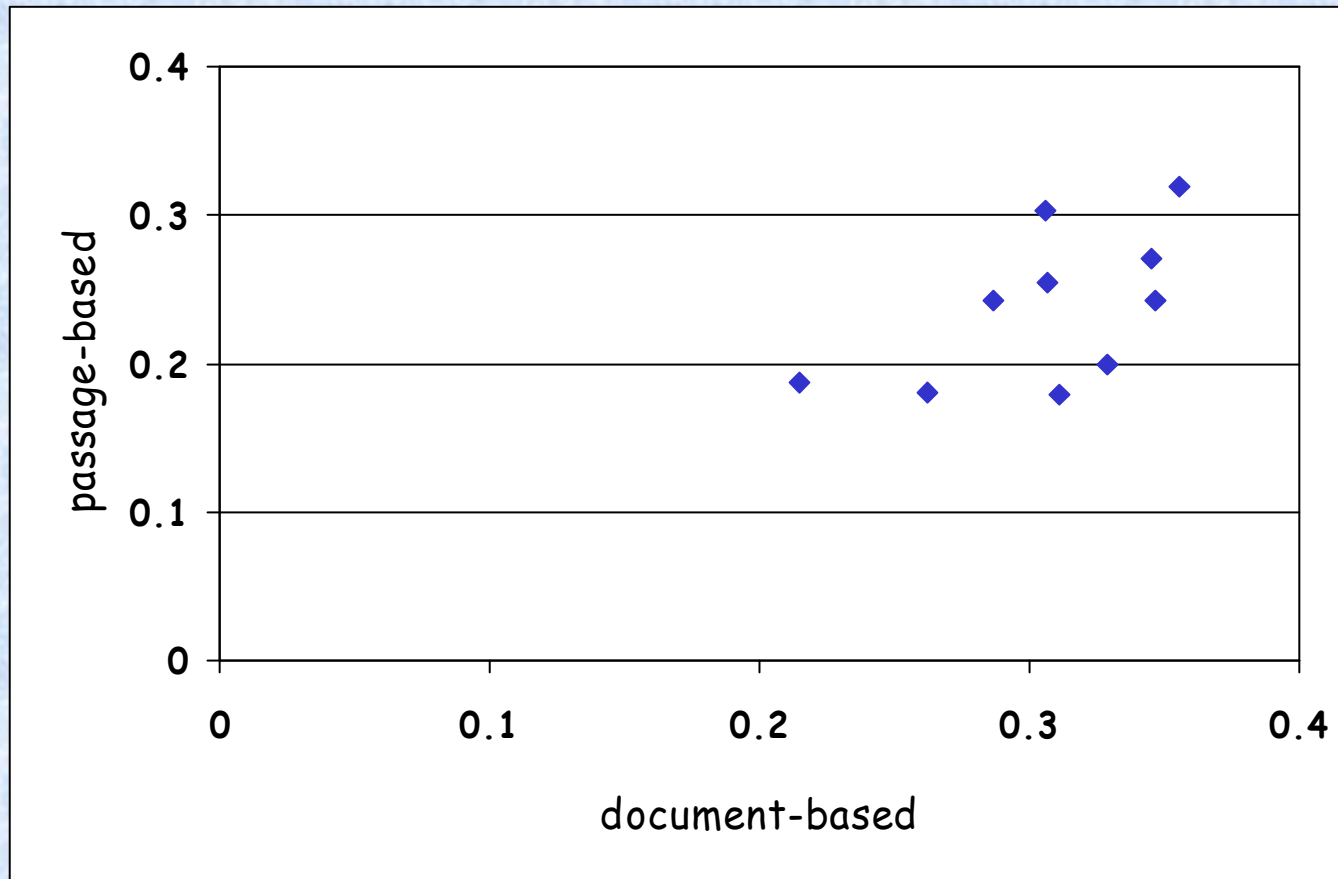
Top HARD runs vs. Baseline



Sorted by MAP of non-baseline run using HARD-rel judgments

Text REtrieval Conference (TREC)

Evaluation by Passages



Scatter plot of passage-based R-prec vs. document-based R-prec

Text REtrieval Conference (TREC)

Novelty Track

- Track started in TREC 2002
- Goal: investigate systems' abilities to locate relevant and non-redundant information within an ordered set of docs
- Motivation: reduce user's workload by eliminating extraneous information from system response

Novelty Track

- Task
 - given is a time-ordered set of relevant docs segmented into sentences & a topic statement
 - return
 - 1) the set of sentences containing relevant information
 - 2) a subset of the relevant sentences such that redundant information is eliminated

Changes in Novelty Task

- In first year, almost no sentences were relevant & virtually all relevant were novel
- Changes for year 2:
 - document collection changed to AQUAINT collection (parallel newswires)
 - created new topics; topic author did judging
 - 25 event topics & 25 opinion topics
 - various kinds and amounts of training data defined separate tasks

Novelty Track Tasks

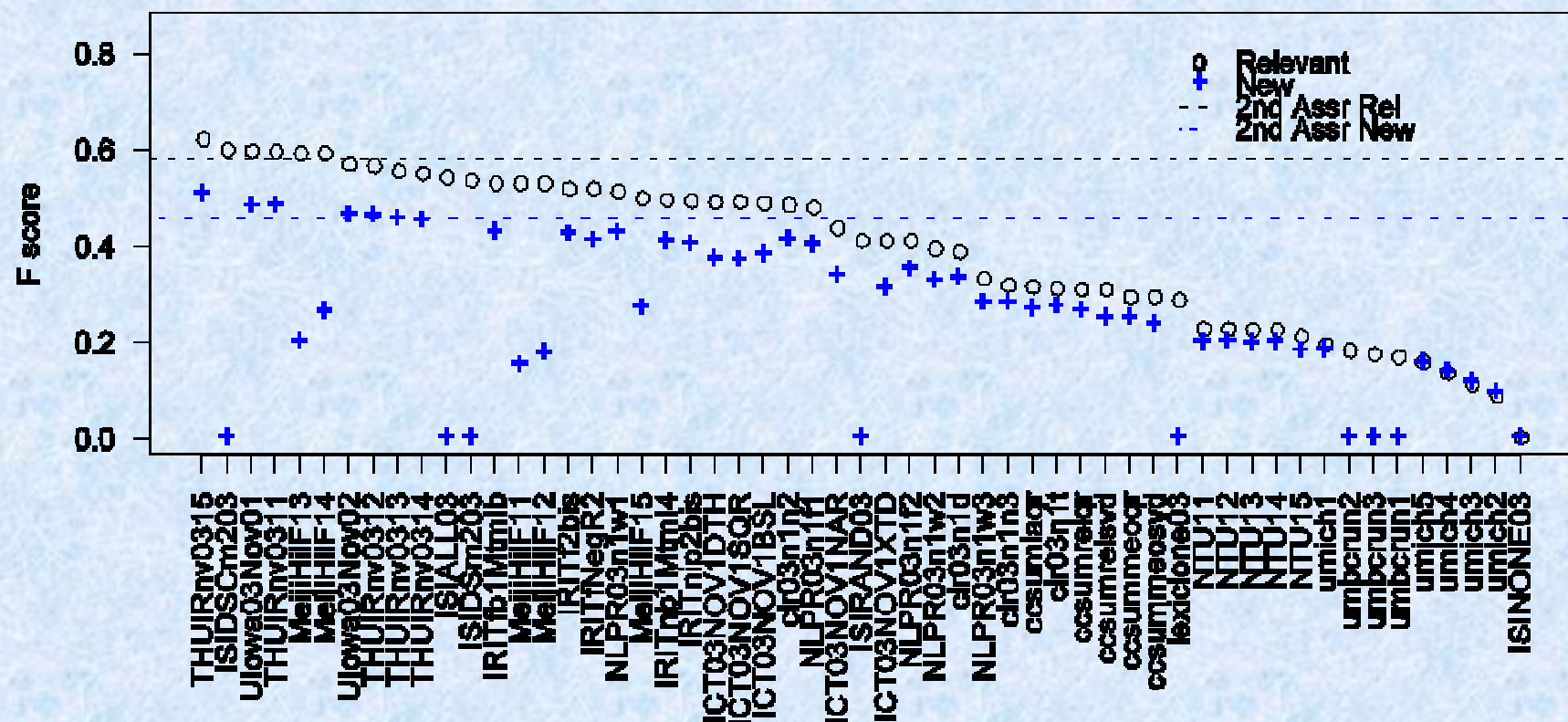
- **Task 1:** Find all relevant and new sentences in 25 documents per topic
- **Task 2:** Given all relevant sentences, find all new sentences
- **Task 3:** Given relevant and new sentences for first 5 documents, find relevant and new sentences in remaining 20 documents
- **Task 4:** Given all relevant sentences and new sentences in first 5 documents, find new sentences in remaining 20 documents

Novelty Evaluation

- Reference data created by assessors
 - performed Task 1 manually
 - each topic independently judged twice, but evaluation based on sentence sets of author
- Measure
 - F score with R and P equally weighted
 - M = number of matched sentences
 - A = number of sentences assessor chose
 - S = number of sentences returned
$$R = M/A \quad P = M/S$$
$$F = (2 \times P \times R) / (P + R)$$

Novelty Track Results

Task 1, Relevant and Novel F Scores



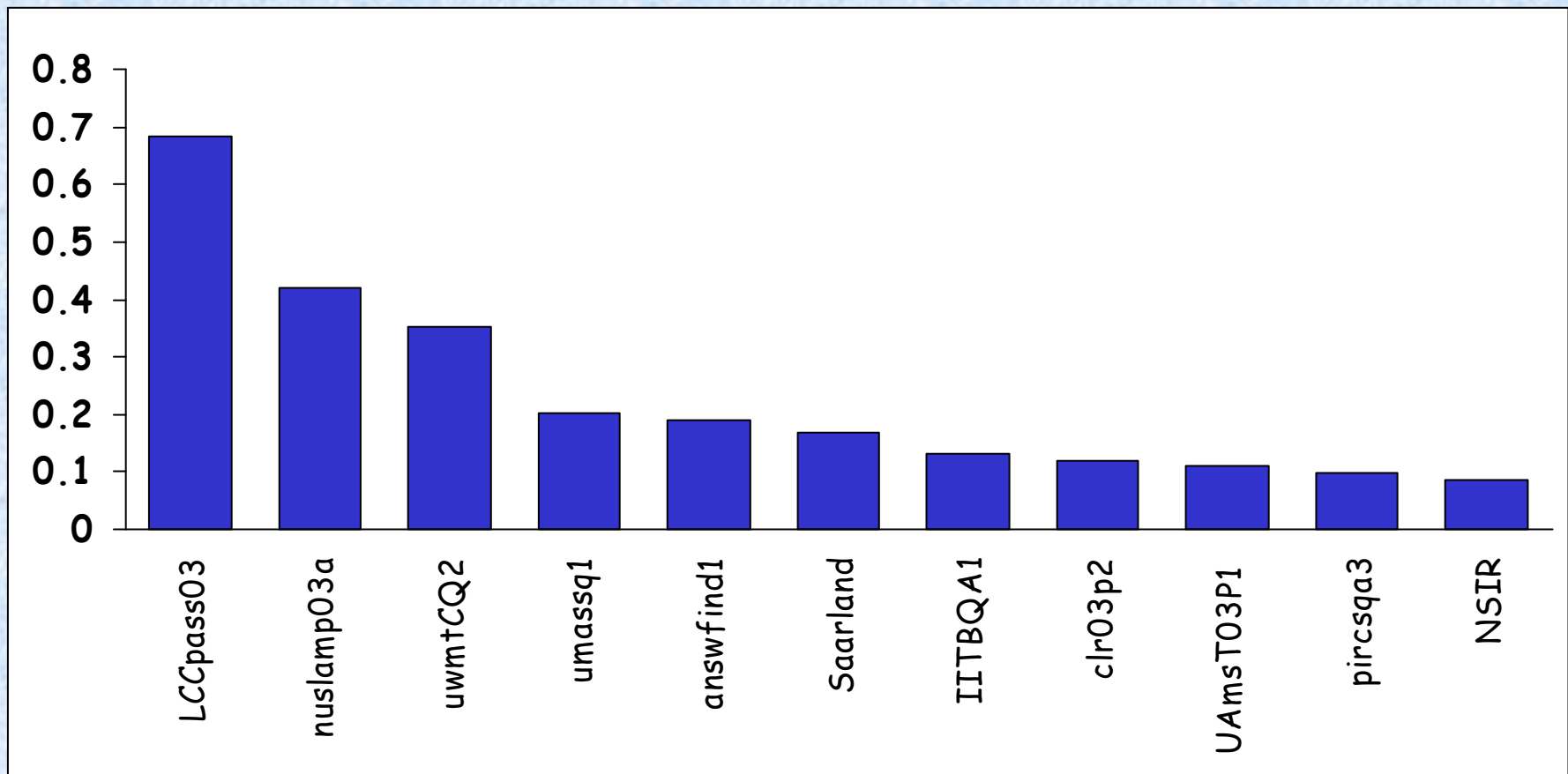
Question Answering Track

- Goal: return answers, not document lists
- Two tasks:
 - passages: return single document extract that contains an answer to a factoid question
 - main: combination of three question types; each type was tagged and has its own evaluation method. Final score a weighted average of components.
- Both tasks used AQUAINT document collection as source of answers
 - 3 GB text; approx. 1,033,000 newswire articles

QA Passages Task

- 413 factoid questions
 - drawn from new MSNSearch and AOL logs
 - no guarantee that question has answer in collection, so a response could be 'NIL'
 - else, response was a single document extract no longer than 250 characters
- Evaluated using accuracy
 - answers judged correct/unsupported/wrong
 - accuracy is percentage of correct responses

QA Passages Task Results



Accuracy for best passages task run per group

Main Task

- Three question types
 - 413 **factoids**: same as passages task except must be exact answer, not document extract
 - 37 **lists**: assemble set of instances where each instance is a factoid question answer
 - 50 **definitions**: return text strings that together define target of question

- Final score weighted average of components

$$\text{FinalScore} = \frac{1}{2}\text{FactoidScore} + \frac{1}{4}\text{ListScore} + \frac{1}{4}\text{DefScore}$$

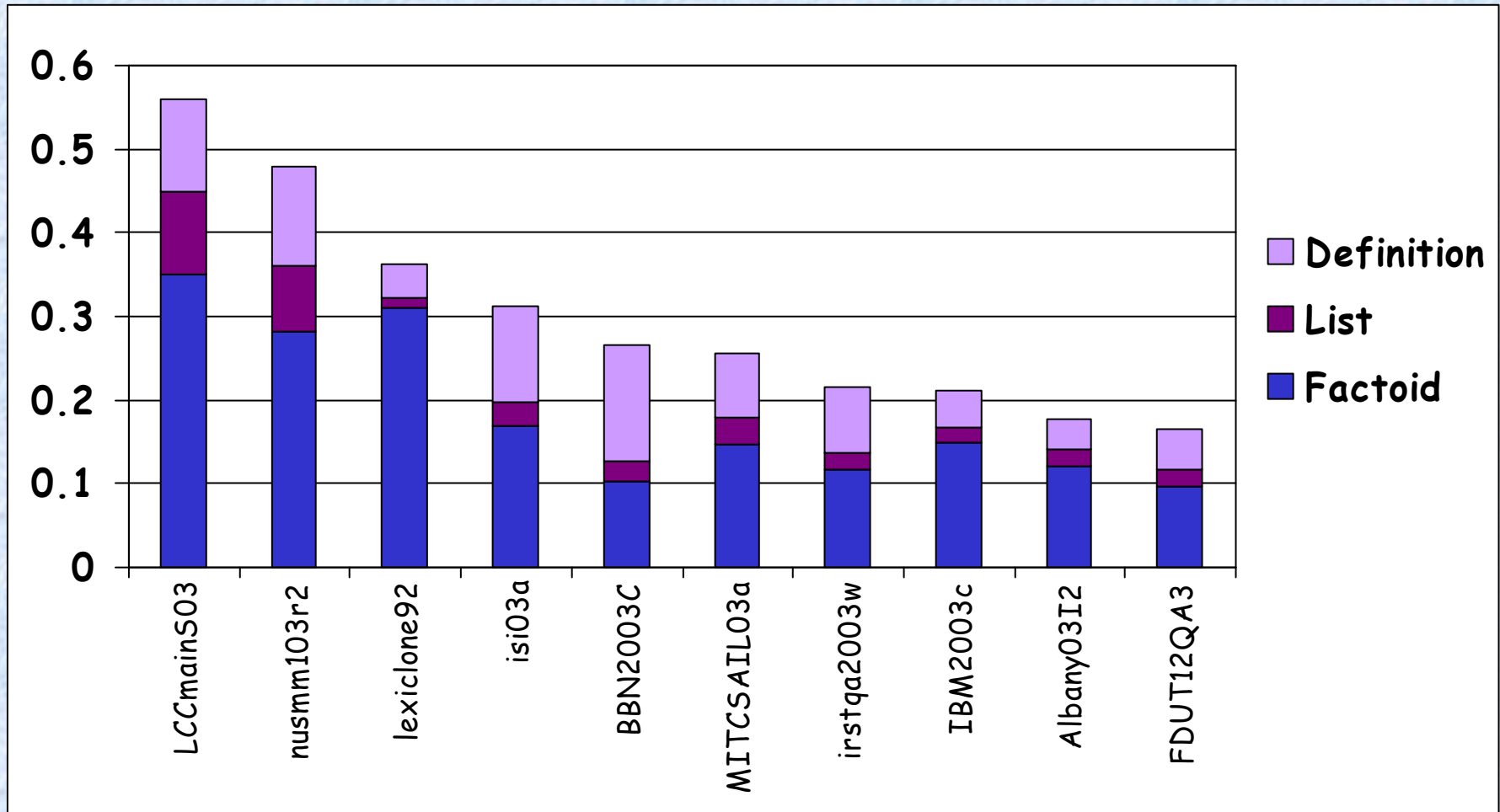
QA List Component

- 37 questions that ask for instances of a type
 - *What Chinese provinces have a McDonald's restaurant?*
 - questions created by NIST assessors
 - may be multiple instances per document & multiple documents with an instance
 - no target number to retrieve specified in question
- Response is an unordered set of instances
 - an instance is a single [doc, string] pair
 - answer-string required to be exact
- Evaluated using F score on instance recall and precision
 - recall and precision equally weighted
 - average F over 37 questions is list component score

QA Definition Component

- 50 questions asking for a definition of a term or biographical data for a person
 - *Who is Vlad the Impaler? What is pH in chemistry?*
 - questions drawn from same logs as factoids
 - assessor created definition by searching docs
- System response is an unordered set of strings
 - each string represents different facet of def
 - no limit on length of strings or number of strings
- Assessor matched his facets to system strings
 - could be 0, 1, or multiple matches per string
 - F score with recall weighted 5 times "precision"
 - "precision" is a function of length

QA Main Task Results



Final combined scores for best main task run per group for top 10 groups

Robust Retrieval Track

- New track in 2003
- Motivations:
 - focus on poorly performing topics since average effectiveness usually masks huge variance
 - bring traditional ad hoc task back to TREC
- Task
 - 100 topics
 - 50 old topics from TRECs 6-8
 - 50 new topics created by 2003 assessors
 - TREC 6-8 document collection: disks 4&5 (no CR)
 - standard trec_eval evaluation plus new measures

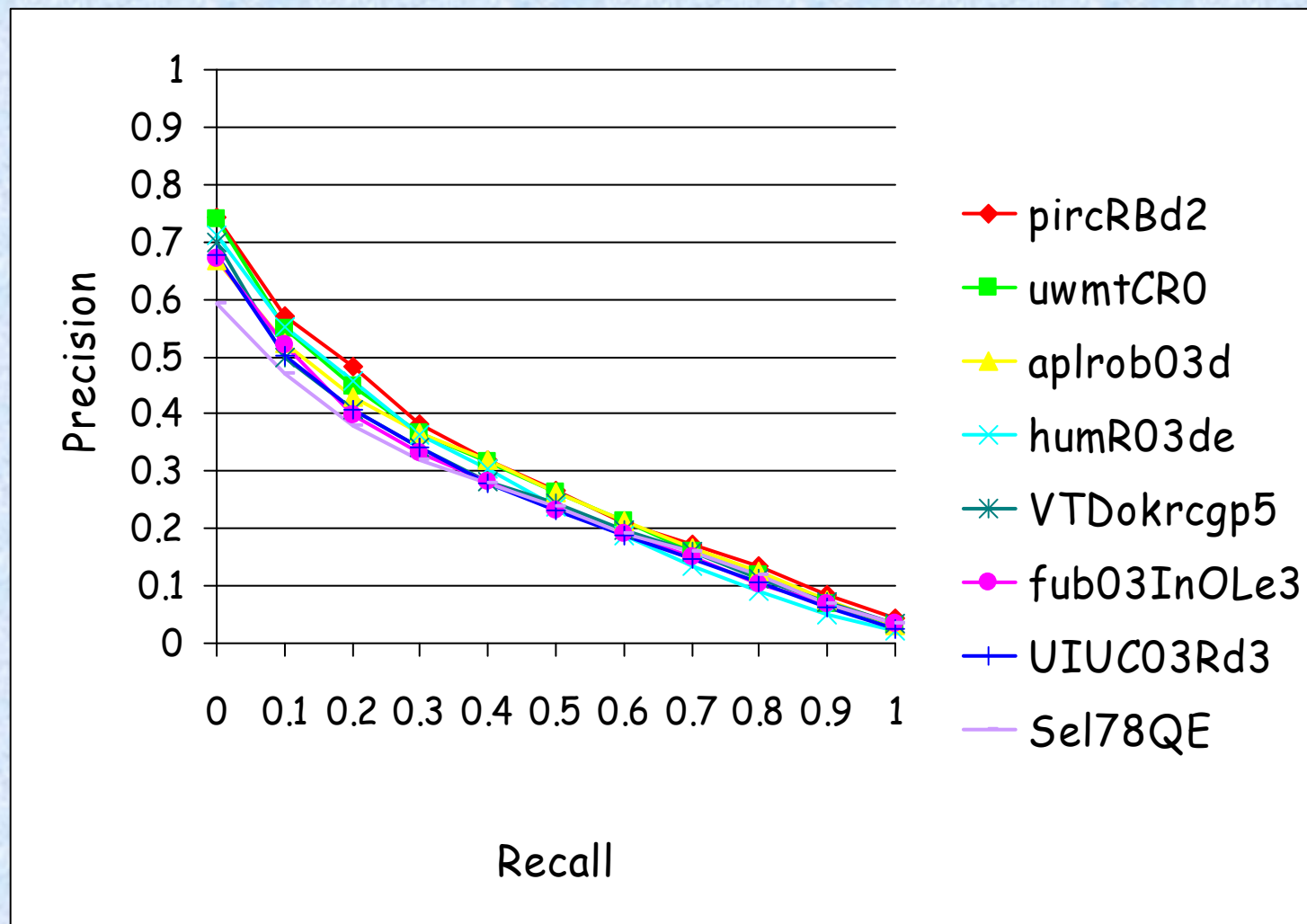
Robust Topic Selection

- 50 old topics are difficult
 - median average precision score low with at least one high outlier in previous TREC
 - systems could train using relevance data; relevance data not used in run itself
- 50 new topics intended as control group
 - created using standard topic development process
 - no relevance data existed until results submitted
 - judged on 3-point scale to create an ad hoc collection with multiple levels of relevance; multiple levels not used for track evaluation

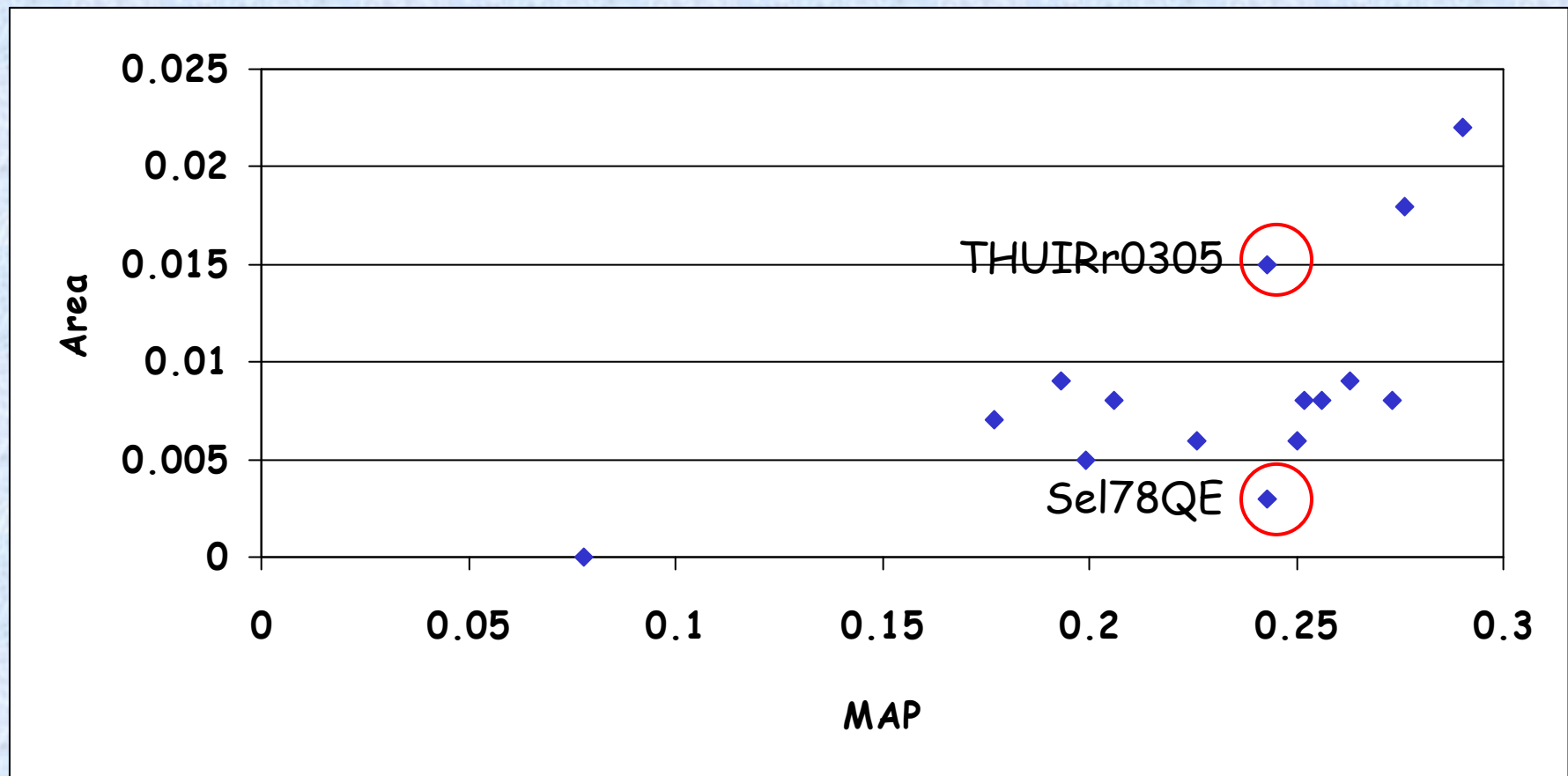
Measures for Robust Retrieval

- Percentage of topics with no relevant retrieved in top 10
 - direct, intuitive measure of behavior of interest
 - very coarse measure
- Area under $MAP(X)$ vs. X curve
 - much more sensitive but far less intuitive measure
 - compute MAP over worst X topics & plot value as a function of X ; use $X \leq \frac{1}{4}N$ when there are N topics total; calculate area underneath this curve
 - emphasizes the worst topics
 - different systems have different worst topics, so measure computed over different set per system

Best Description-Only Runs, Combined Topic Set



Differences in Measures



Scatter plot of area measure vs. MAP for description runs

Text REtrieval Conference (TREC)

Web Track

- Investigate retrieval behavior on the web
- Three tasks
 - topic distillation:
 - both interactive & non-interactive versions
 - similar to ad hoc task but goal is to find homepages of credible sites devoted to subject matter of topic
 - navigational: known-item task to find page specified by topic
- Document set
 - crawl of .GOV developed for last year's web track
 - approx. 18 GB, 1.25 million docs

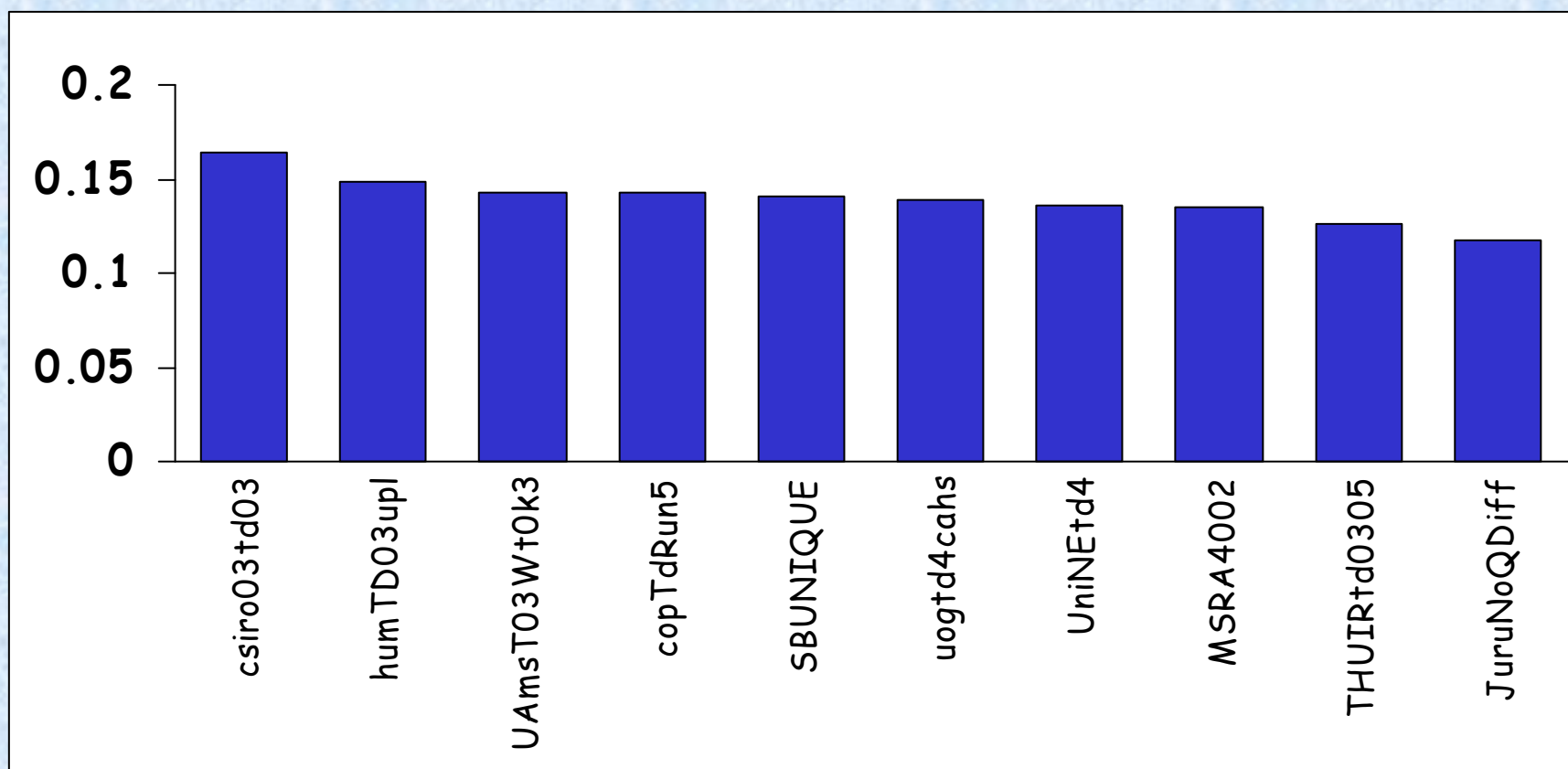
Web Interactive Task

- Goal: explore the role of the human user in the topic distillation task
- 2 groups (CSIRO, Rutgers) participated
 - both investigated whether more structured retrieval results help humans do task
 - found little/no significant difference in quality of final retrieved sets, but users reported liking structured results better & were slightly more efficient when using structured results

Web Topic Distillation Task

- Topics were broad ad hoc topics
 - 50 topics created by NIST assessors
 - target content for which .GOV has good resources
- Binary judgments by topic author
 - requires assessor to understand structure of a site to identify its homepage & to judge the quality of the information the site offers
- Evaluation by R-precision
 - many topics had fewer than 10 relevant homepages

Topic Distillation Results

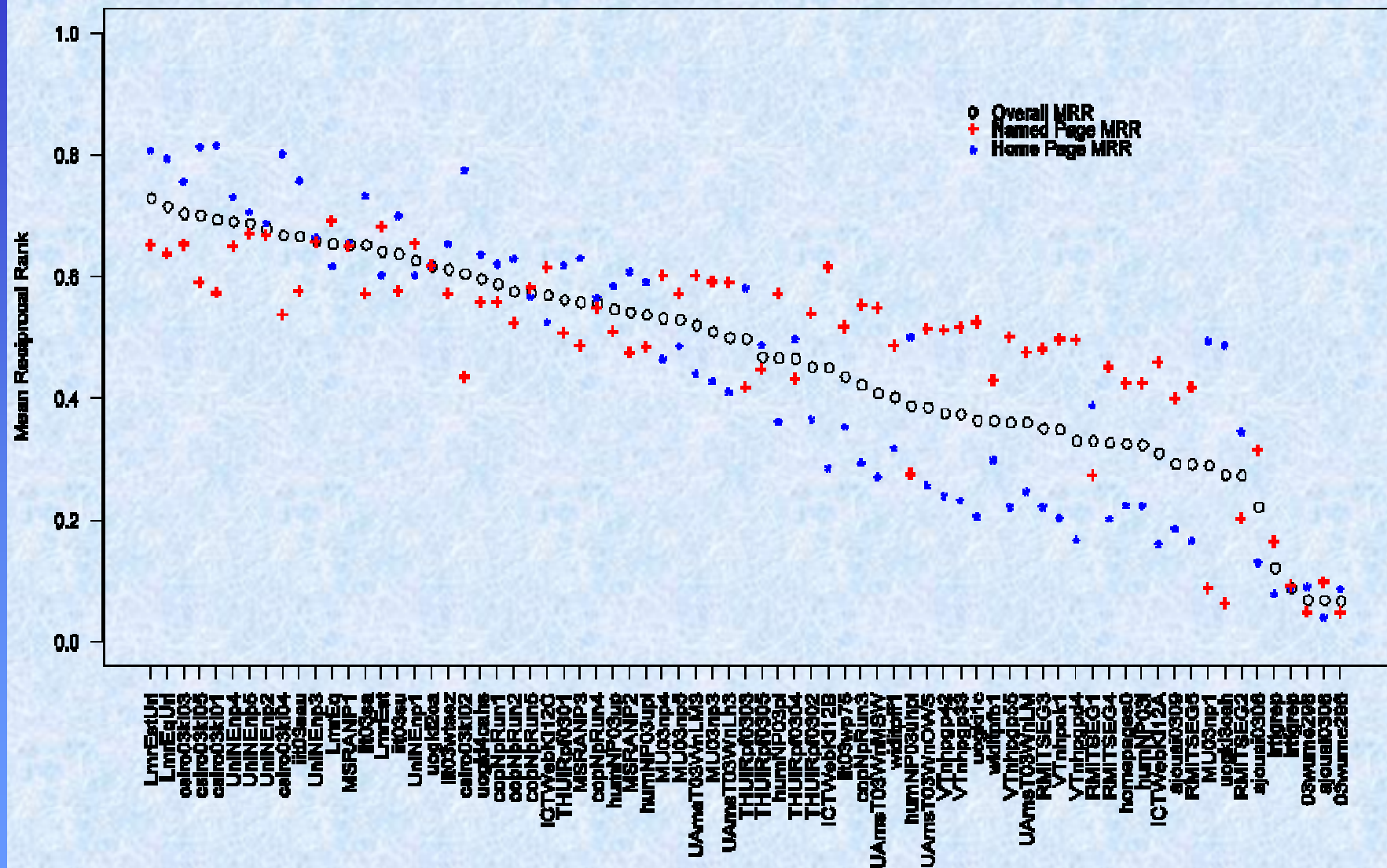


Top 10 groups by average R-prec of best run

Web Navigation Task

- Generalization of homepage finding task: return the page described in the topic
 - *Tennessee Valley authority* → www.tva.gov
 - 300 topics created by the assessors
 - half of topics targeted homepages
- Small pools to find aliases, mirror sites
- Evaluation:
 - MRR of first correct page
 - percentage of topics that do not return correct page in top 10 retrieved

Web Navigational Results



Text REtrieval Conference (TREC)

Future

- TREC will continue
 - This year's tracks likely (not guaranteed!) to continue
 - genomics, HARD, robust retrieval: tracks have always run for at least two years
 - QA: strong sponsor interest
 - web, novelty: interest remains high; some question as to what's next with regard to tasks
 - One new track proposal
 - investigate ad hoc evaluation methodologies for terabyte scale collections
 - outgrowth of SIGIR 2003 workshop