Overview of TREC 2004

Sponsored by:
NIST, ARDA, DARPA

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

TREC 2004 Program Committee

Ellen Voorhees, chair
James Allan
Chris Buckley
Gord Cormack
Sue Dumais
Donna Harman
Dave Hawking
Bill Hersh
David Lewis
John Prager
John Prange
Steve Robertson
Mark Sanderson
Karen Sparck Jones
Ross Wilkinson

TREC 2004 Track Coordinators

Genomics: William Hersh

HARD: James Allan

Novelty: Ian Soboroff

Question Answering: Ellen Voorhees

Robust Retrieval: Ellen Voorhees

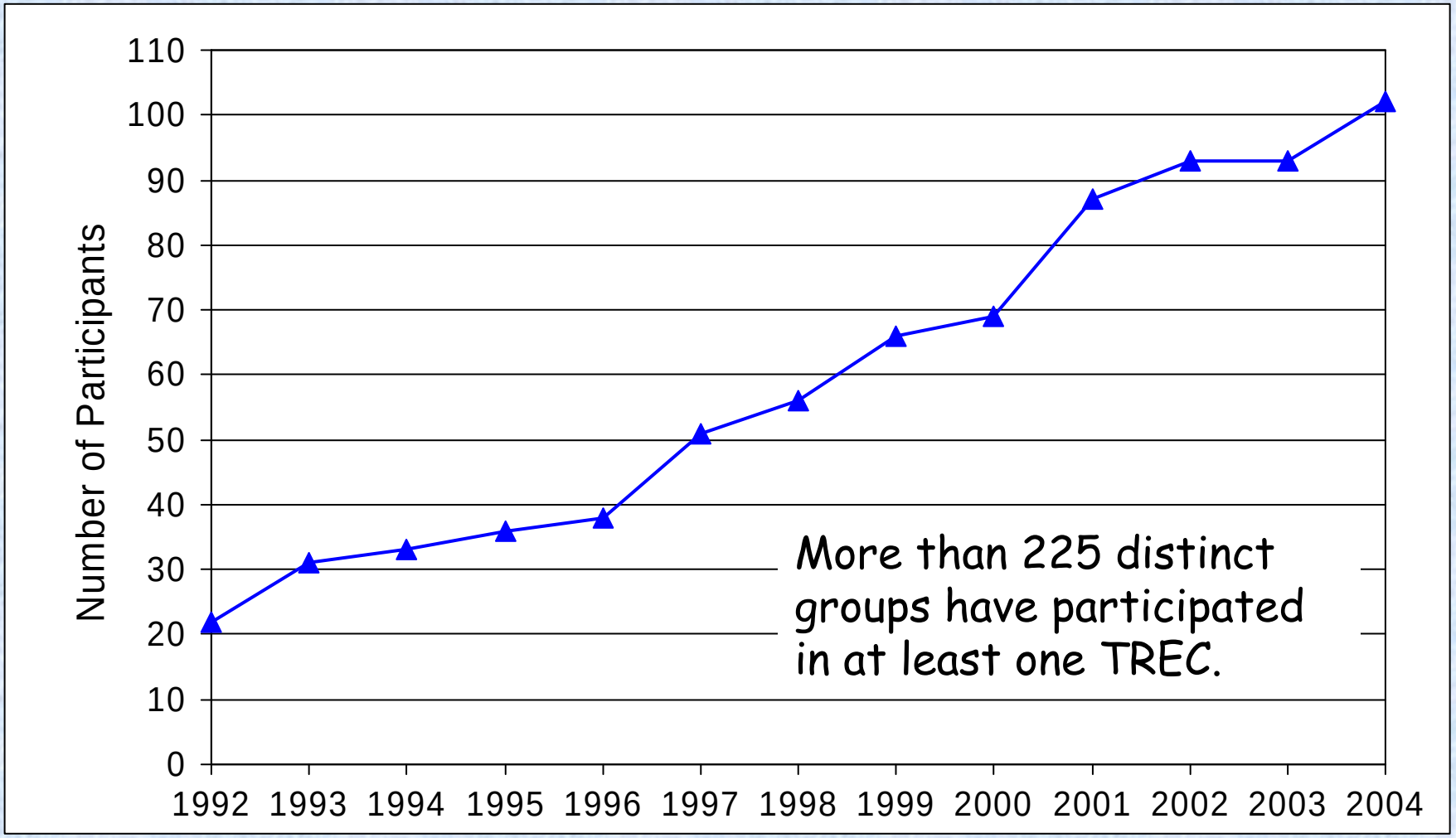
Terabyte: Charles Clarke, Ian Soboroff

Web: David Hawking, Nick Craswell, Ian Soboroff

Alias-i, Inc.
Arizona State U.
Cal. State San Marcos
Carnegie Mellon U.
Chinese Acad. Sci. (3)
Chinese U. Hong Kong
Clairvoyance Corp.
CL Research
Columbia U.
ConverSpeech&Stanford
CSIRO
Dalhousie U.
Decision Aid Team, LAMSADE
Dublin City U.
Etymon
Fondazione Ugo Bordoni
Fudan University (2)
German U. in Cairo
Hong Kong Polytechnic U
Hummingbird
IBM India Research Lab
IBM Research, Haifa
IBM Research, Watson
IDA/CCS
Illinois Inst. of Tech.
Indiana U. (2)
IRIT/SIG
ITC-irst
Johns Hopkins U., APL
Korea University
Language Comp. Corp.
LexiClone, Inc
Macquarie University
Mass. Inst. Tech.
Max-Planck Inst.
Meiji University
Microsoft Research Asia
Microsoft Research Ltd
MITRE Corp.
Monash U.
National Central U.
NSA
National Taiwan U.
Nat'l U. of Singapore
NUS-MIT Alliance
NLM & U Maryland
Oregon Health & Sci. U.
PATOLIS Corp.
Peking U.
Queens College, CUNY
RMIT U.
Rutgers U. (2)
Saarland U.
Sabir Research
Shanghai JiaoTong U.
SUNY Buffalo
Tarragon Consulting
The Robert Gordon U.
TNO & Erasmus MC
Tsinghua University (2)
UC Berkeley
U. Hospital Geneva
U. Lisboa Campo Grande
U. Politcnica Catalunya
U. Paris Sud
U. of Alaska Fairbanks
U. of Alberta
U. of Amsterdam
U. of Chicago
U. of Cincinnati
U. of Edinburgh
U. Edinburgh & Sydney
U. of Glasgow
U. Illinois Chicago
U.Illinois (UIUC)
U. of Iowa
U. of Lethbridge
U. of Limerick
U. of Maryland
U. Massachusetts
U. of Melbourne
U. of Michigan
U. of North Carolina
U. of North Texas
U. of Padova
U. of Pisa
U. of Sheffield
USC-ISI
U. of Tampere
U. of Tokyo
U. of Twente
U. of Wales, Bangor
U. of Waterloo (2)
U. of Wisconsin
Virginia Tech
York University

Text REtrieval Conference (TREC)

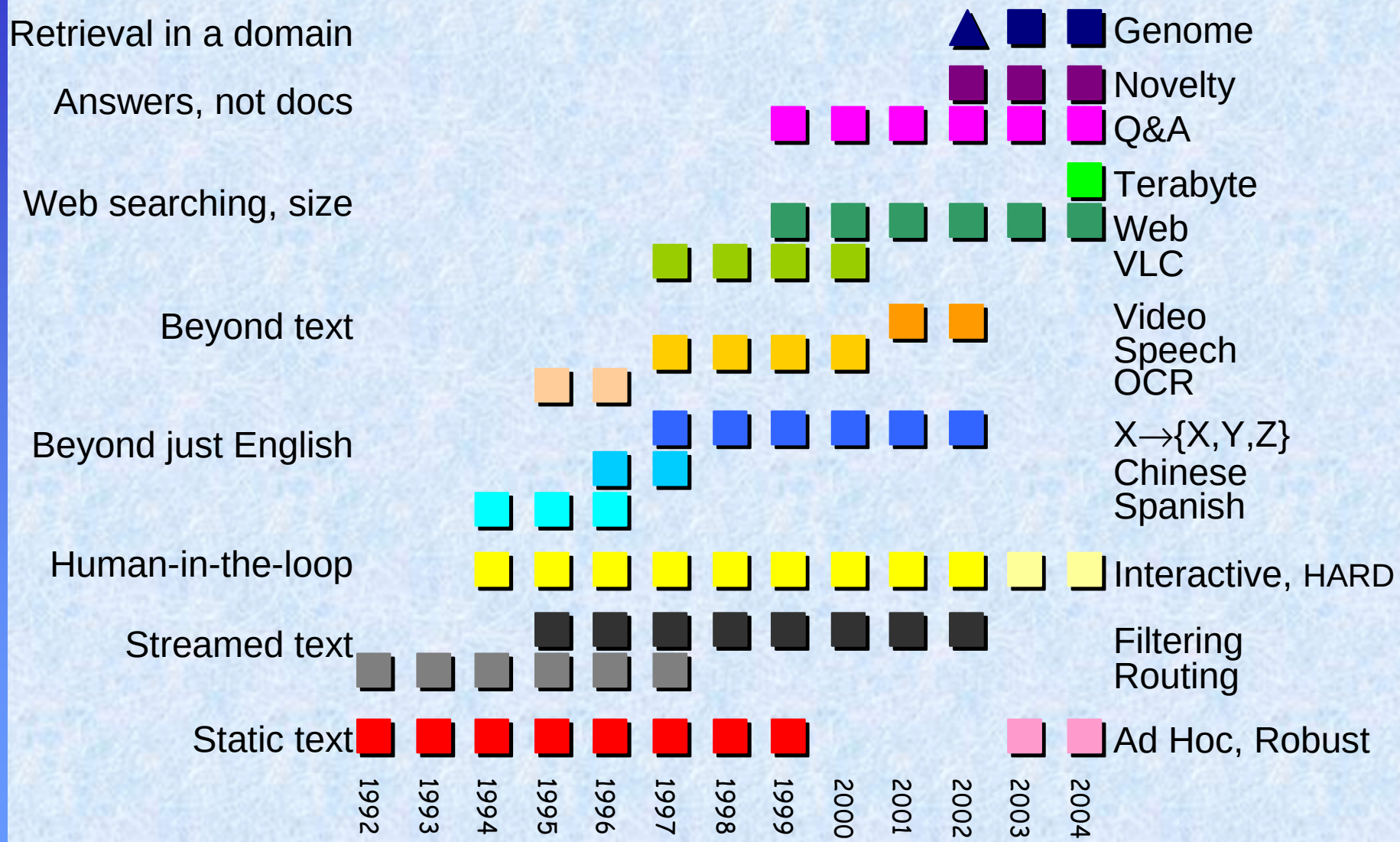
Participant Growth in TREC



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

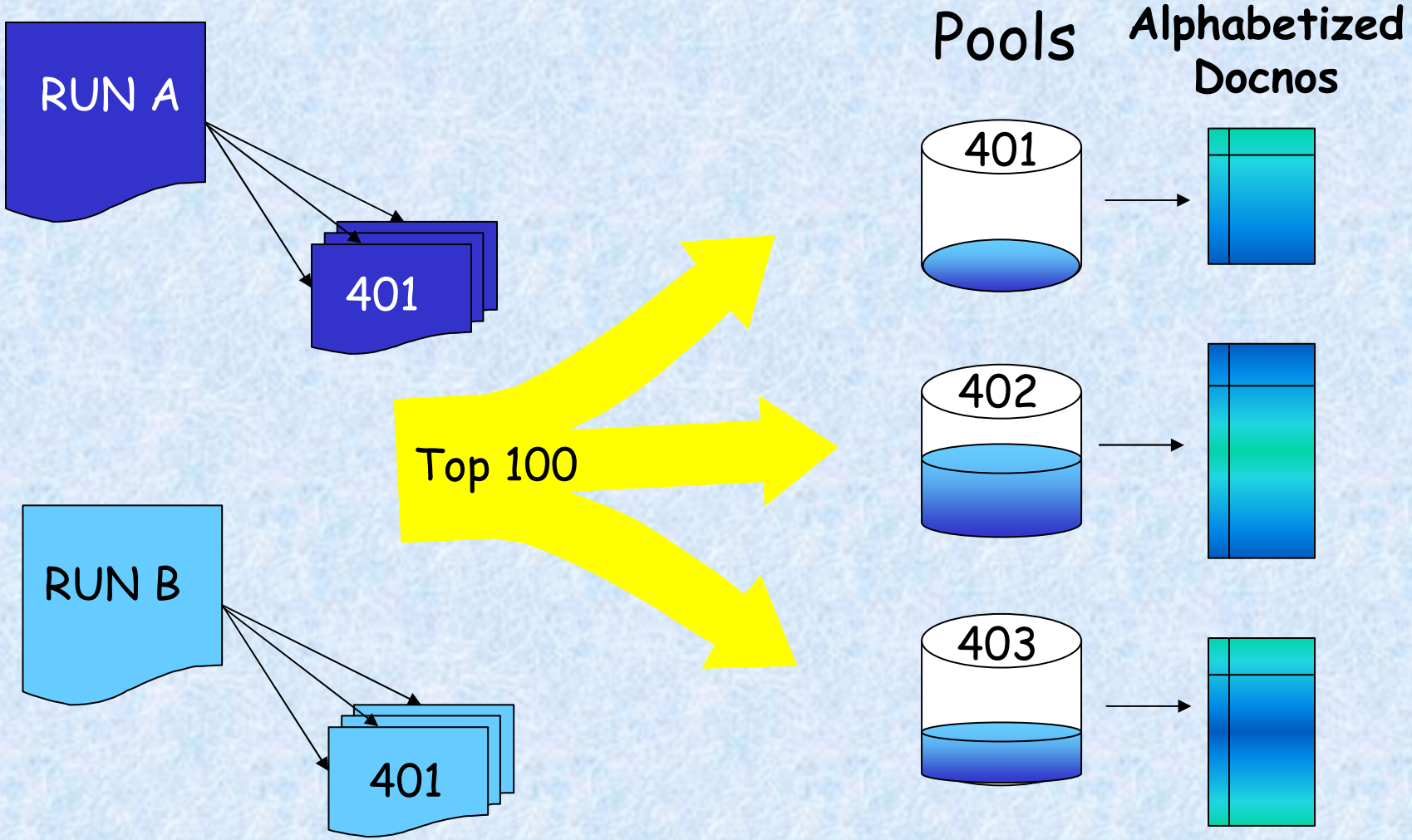
TREC Tracks


Retrieval in a domain
Answers, not docs
Web searching, size
Beyond text
Beyond just English
Human-in-the-loop
Streamed text
Static text
Genome
Novelty
Q&A
Terabyte
Web
VLC
Video
Speech
OCR
X→{X,Y,Z}
Chinese
Spanish
Interactive, HARD
Filtering
Routing
Ad Hoc, Robust
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004

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
  - classification

Creating Relevance Judgments



Text REtrieval Conference (TREC)

TREC 2004 Tracks

- Genomics
  - ad hoc, categorization (triage, annhi, annhiev)
- HARD
- Novelty
  - tasks 1-4
- Question Answering
- Robust Retrieval
- Terabyte
- Web
  - mixed query, categorization

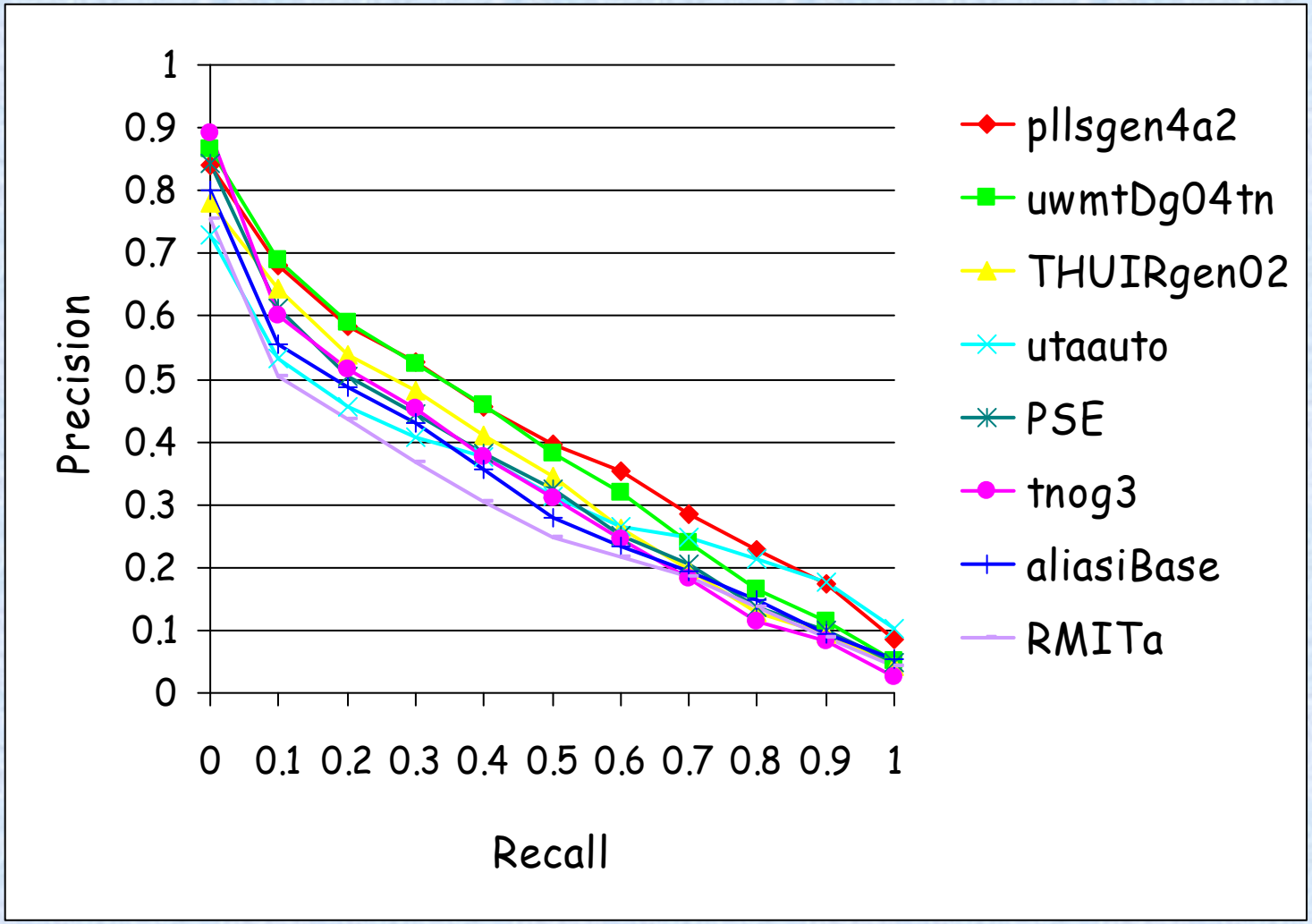
Genomics Track

- Motivation: explore retrieval in a domain
  - with focus on person experienced in the domain
- Two tasks
  - ad hoc: ad hoc retrieval task using MEDLINE records
  - categorization: assist curation process
    - recognize whether documents contain specific kinds of information

Ad Hoc Task

- Documents
  - ~4,600,000 MEDLINE records (~9.5gb) inserted into system between 1994-2003
  - provided to the track by NLM
- Topics
  - 50 topics derived from interviews of biologists with real information needs
  - title, need, and context fields
- Relevance judgments
  - created from pooled results
  - two biologists (1 PhD, 1 undergrad) did judging
  - 3-way judgments: definitely, possibly, not relevant

Top Automatic Ad Hoc Runs



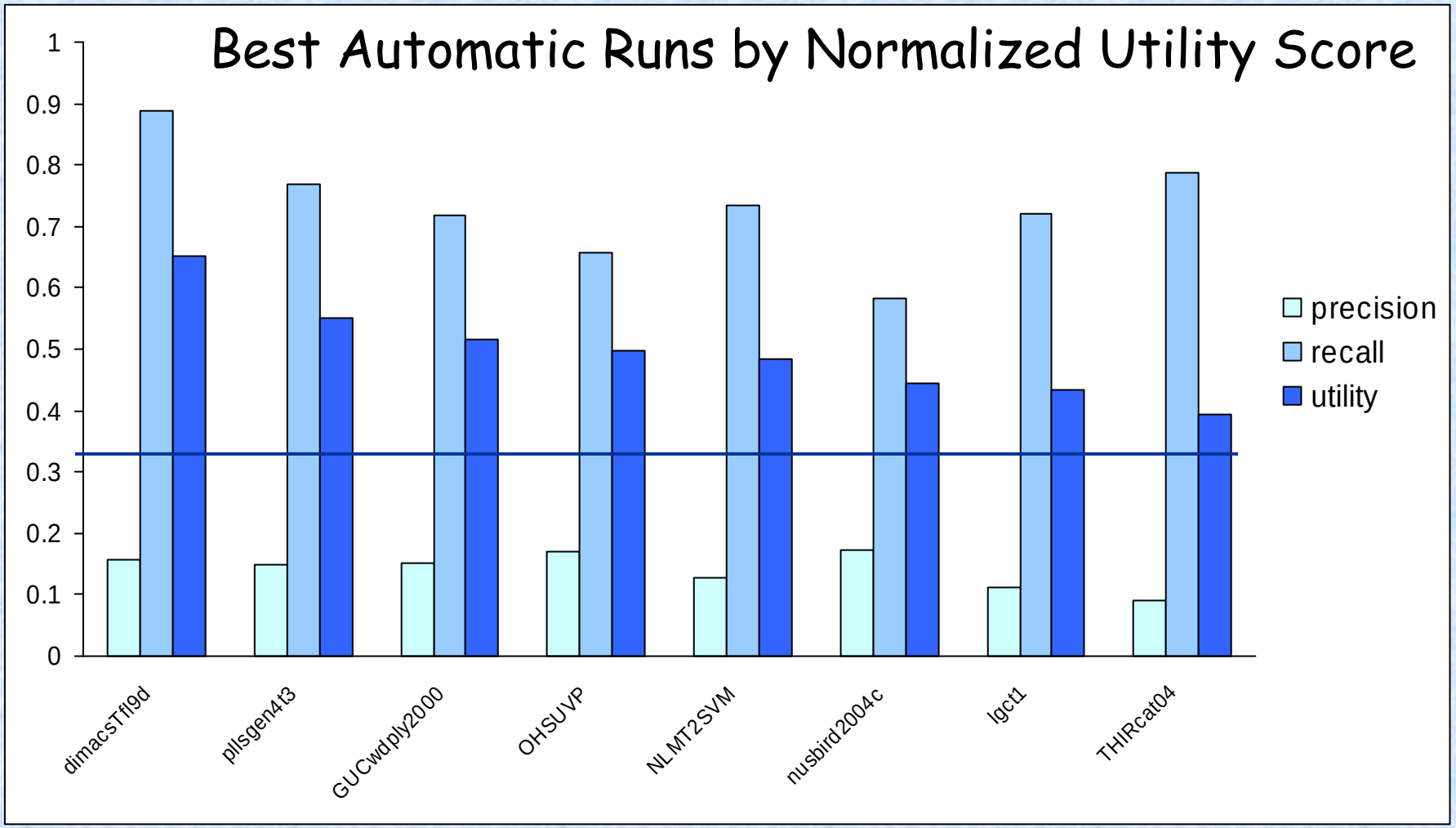
Categorization Tasks

- Genomics field has "model organisms" databases that are manually curated
  - collection of papers regarding target organism with linkages to other resources such as GO
- Classification tasks were abstractions of various tasks currently done by curators
  - triage: find documents that have experimental evidence that requires GO annotation
  - annhi: select the GO hierarchies that contain terms to use in the annotation of this doc
  - annhiev: select which GO evidence codes to use in the annotation of this doc

Categorization Tasks

- Document set
  - full text documents from 2 years of 3 journals
  - text made available by Highwire Press
- Judgments
  - documents were part of the actual curation process of the MGI system
  - used annotations produced in this process as truth

Triage Task Results



HARD Track

- High Accuracy Retrieval from Documents
- 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 additional information
  - metadata supplied in topic statement
  - information collected from *clarifying form*
  - varying unit of retrieval (passage vs. full doc)

HARD Collection

- Documents
  - ~650,000 newswire articles from 2003 (~1.5gb)
  - obtained from LDC
- Topics
  - 50 topics created by LDC; 45 used in doc eval
  - extended version includes metadata, retrieval unit
- Judgments
  - made on pooled results
  - off-topic,on-topic (SOFT-rel),relevant (HARD-rel)
  - "SOFT-rel" = on-topic, but metadata not satisfied
  - passages: selected relevant document extracts

Additional Information

- Metadata from topic statements
  - familiarity [little, much]
  - genre [news-report, opinion-editorial, other, any]
  - geography [US, non-US, any]
  - subject domain [free text]
  - related text (either on-topic or relevant)
- Clarifying forms
  - assessor (surrogate user) spends at most 3 minutes/topic responding to topic-specific form
  - example uses:
    - sense resolution
    - relevance judgments

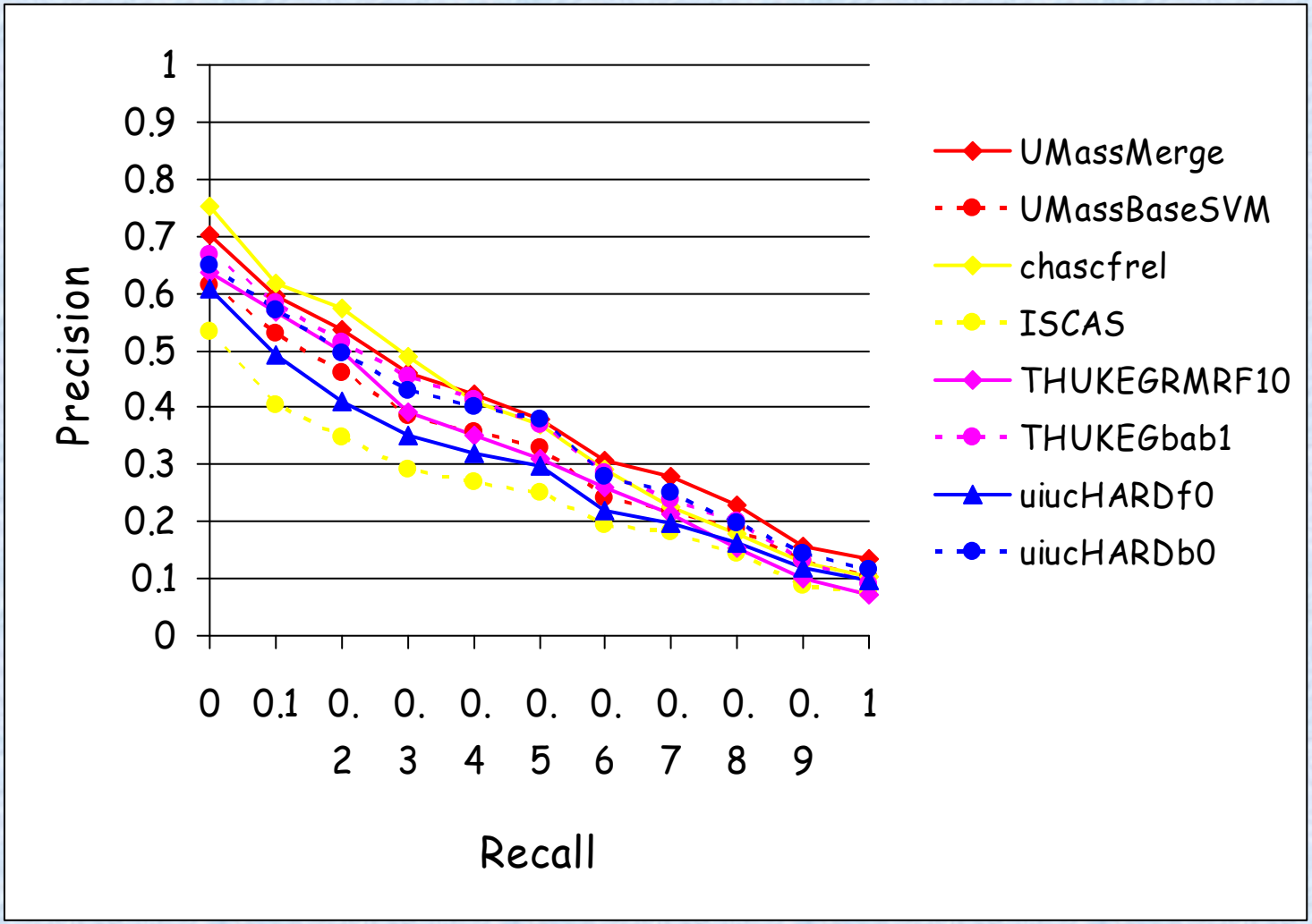
HARD Protocol

- Perform baseline runs using standard topics
- Receive extended topics and/or clarification form responses
- Perform additional (non-baseline) runs exploiting additional info
- Response format based on passage retrieval (where doc is a long passage)

HARD Evaluation

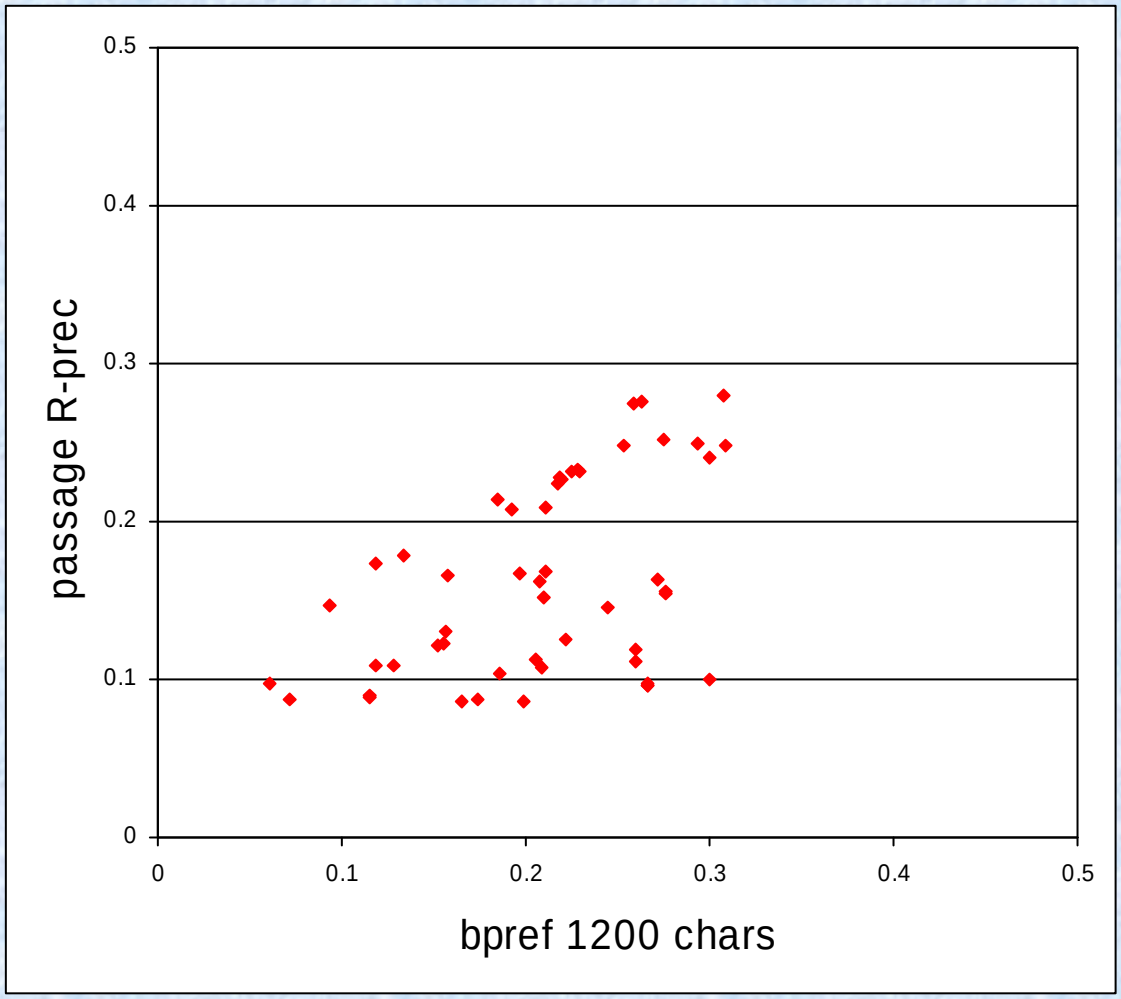
- Document-level
  - standard trec_eval evaluation
  - two evaluation conditions: SOFT-rel documents relevant & SOFT-rel documents not relevant
- Passage-level
  - restricted to 25 topics w/ retrieval unit "passage"
  - two different approaches: character-based & passage-based
  - precision, R-precision, character-based bpref

Top HARD runs vs. Baseline



Sorted by MAP of higher run using HARD-rel judgments

Evaluation by Passages



Scatter plot of bpref 1200 characters vs. passage-based R-prec

Novelty Track

- 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 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
- Tasks same as in 2003 except some topics' document sets may contain irrelevant docs

Novelty Collection

- Documents
  - AQUAINT collection (parallel newswires)
- Topics
  - 50 new topics: 25 events & 25 opinions
- Judgments
  - NIST assessor who created topic manually performed basic task
  - various kinds and amounts of training data defined separate tasks for systems
  - each topic independently judged by second assessor

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

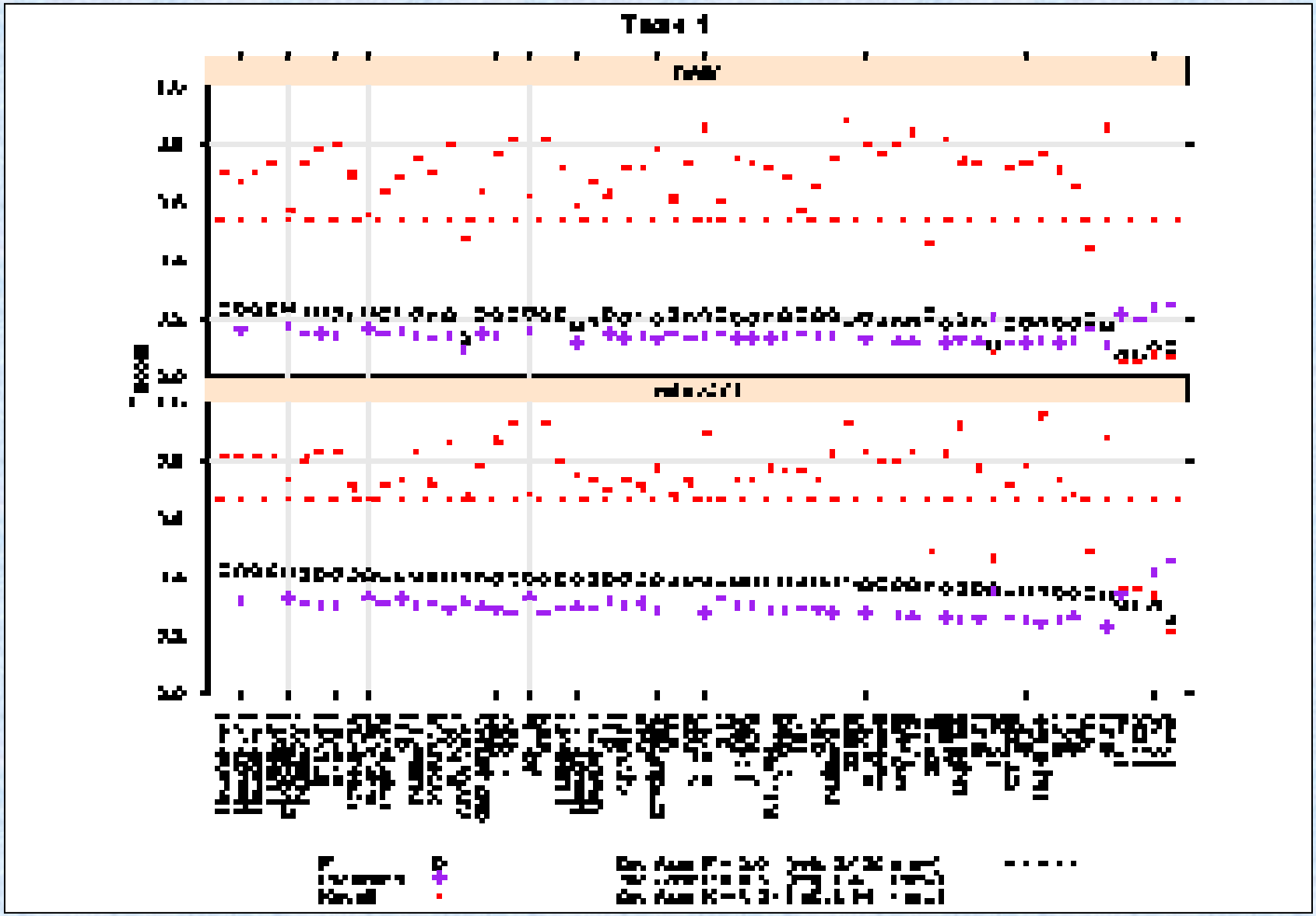
- F score with R and P equally weighted

  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



Question Answering Track

- Goal: return answers, not document lists
- Task:
  - define a target by answering a series of factoid and list questions about that target, plus returning other info not covered by previous questions
  - each question tagged as to type and series
- Used AQUAINT document collection as source of answers
  - 3 GB text; approx. 1,033,000 newswire articles

Question Series

21	Club Med		
	21.1	Factoid	How many Club Med vacation spots are there worldwide?
	21.2	List	List the spots in the United States.
	21.3	Factoid	Where is an adults-only Club Med?
	21.4	Other	

65 series in test set with 4-10 questions per series

- 230 total factoids
- 56 total list questions
- 65 total "other" questions

Factoid Questions

- Response format
  - exactly one response per question
  - since no guarantee that question has answer in collection, a response could be `NIL'
  - else, response was a [docid, answer-string] pair
- Evaluated using accuracy
  - human assessor judged each pair either wrong, unsupported, inexact, or correct
  - NIL response correct iff no known answer
  - accuracy is percentage of 230 questions with a correct response

List Questions

- Questions that ask for instances of a type
  - shorthand for repeatedly asking factoid question
  - may be multiple instances per document & multiple documents with an instance
- 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 instance precision
  - recall and precision equally weighted
  - average F over 55 questions is list component score

`Other' Questions

- Similar to TREC 2003 definition questions
  - additional challenge in recognizing/removing information already returned
- 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
- Assessors matched their facets to system strings
  - could be 0, 1, or multiple matches per string
  - F score with recall weighted 3 times "precision"
  - "precision" is a function of length

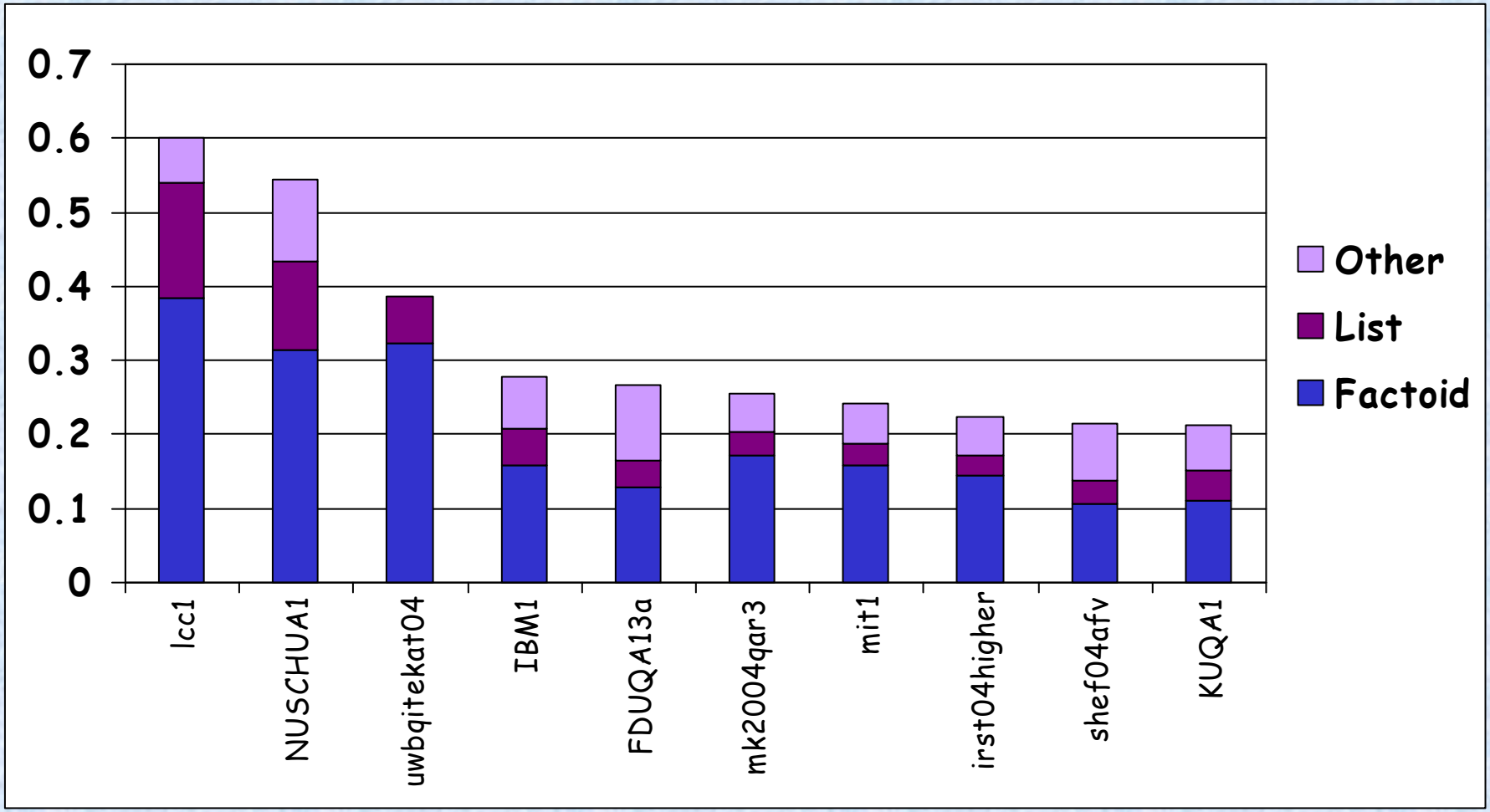
Combined Score

- Final score weighted average of components

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

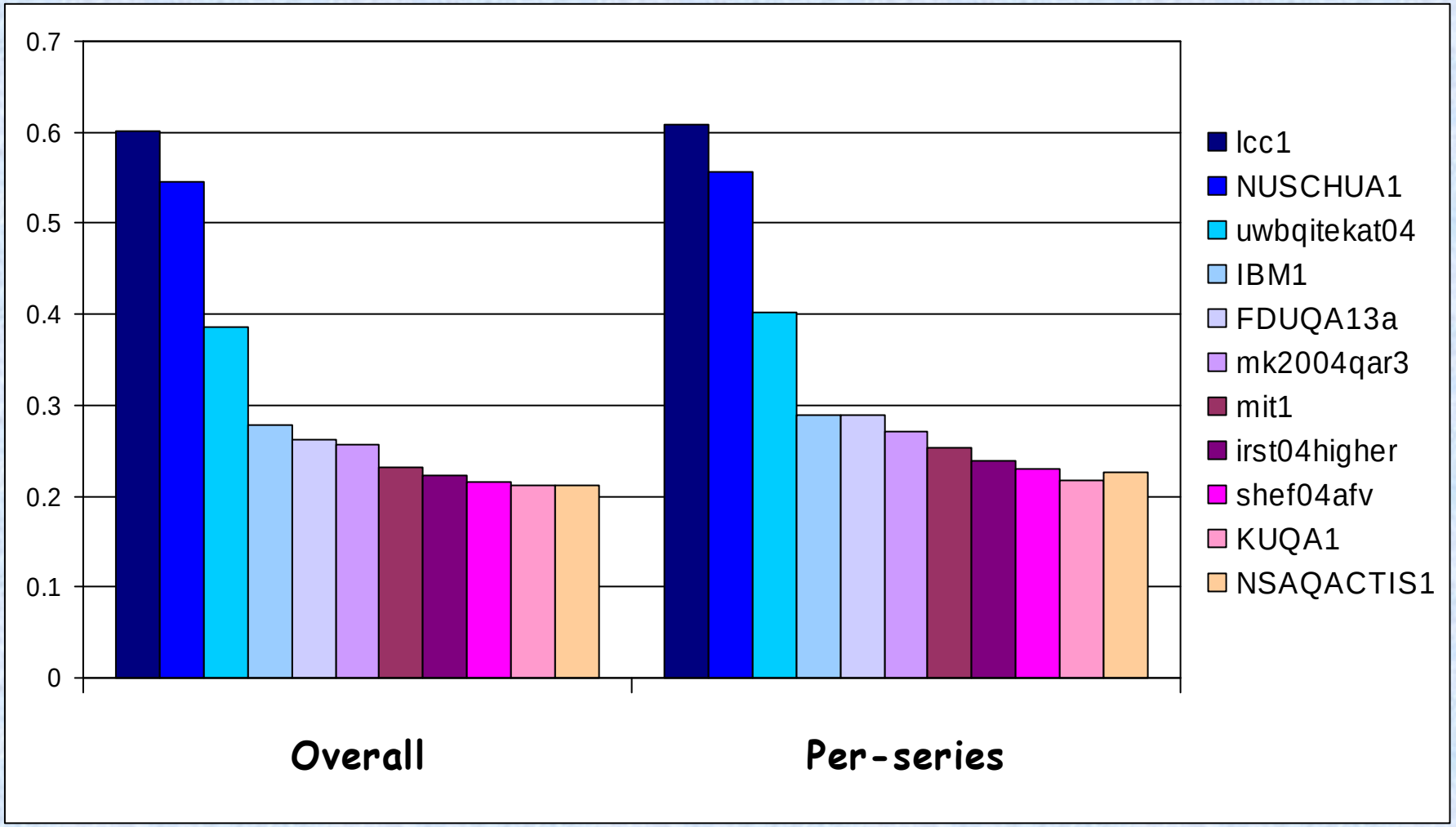
- Can apply same combination function on per-series basis
  - makes QA series more similar to document retrieval topic
  - nicer evaluation properties
  - final scores not equivalent between two methods; little difference in system ranks with current runs

QA Results



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

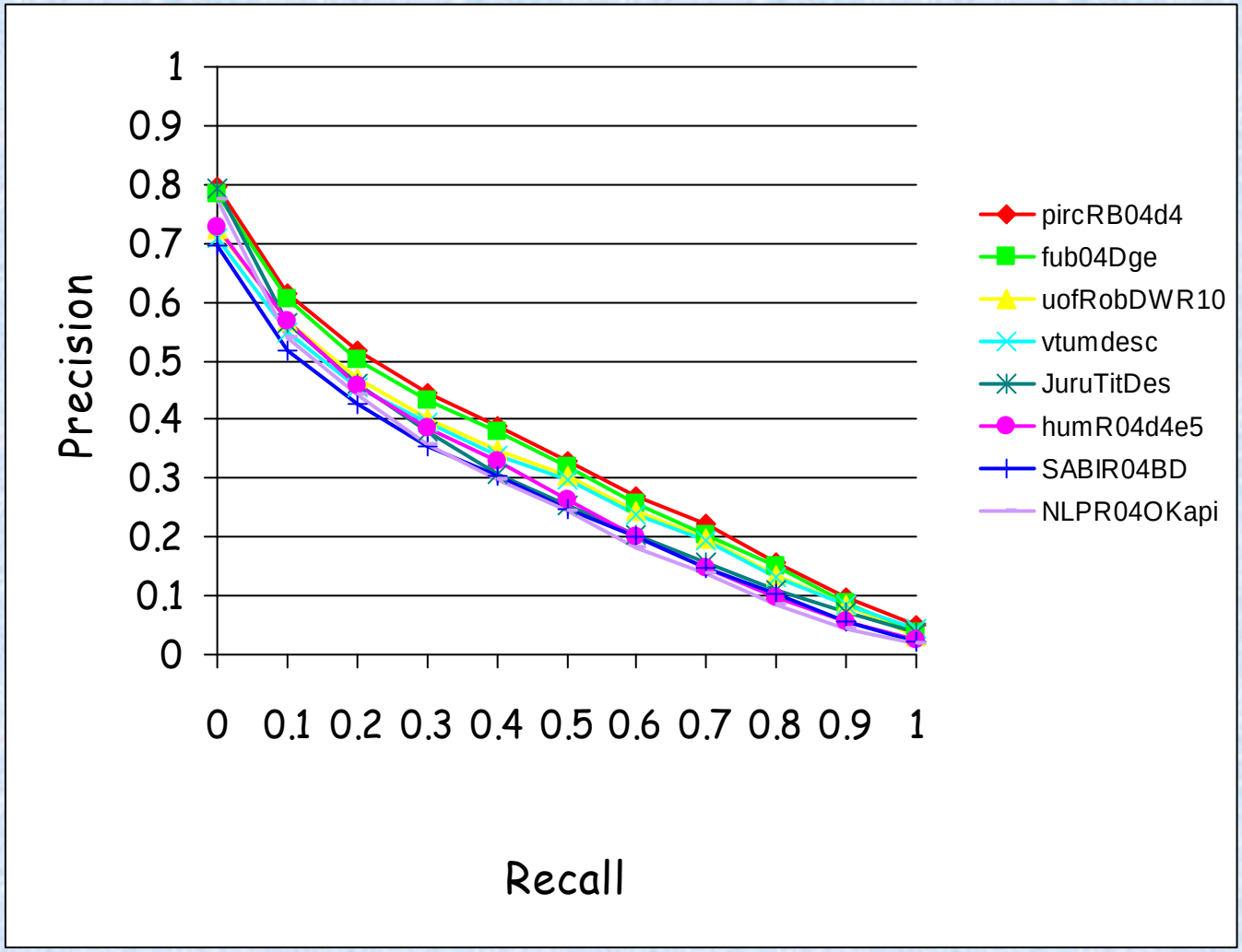
Combined Scores



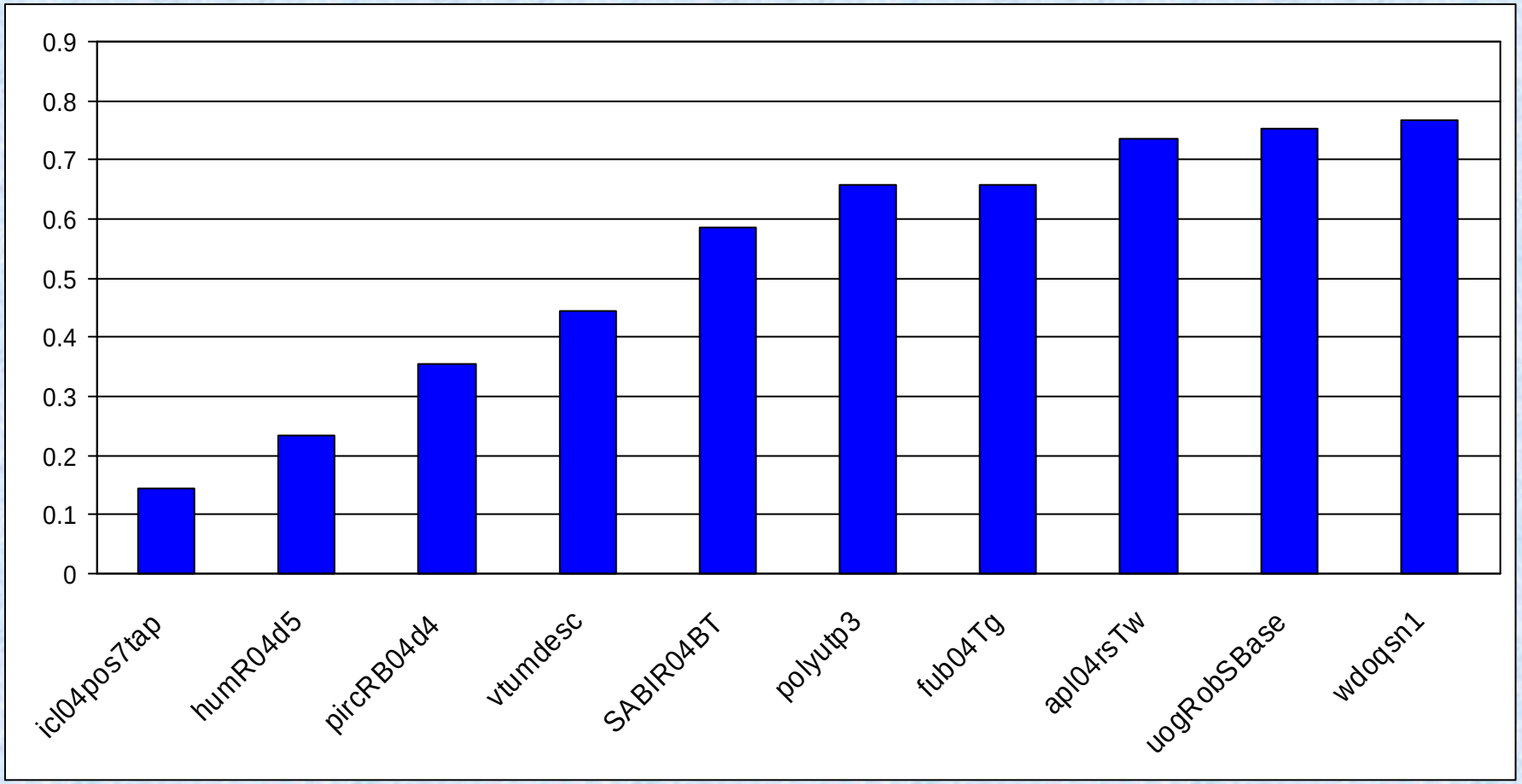
Robust Retrieval Track

- Motivations:
  - focus on poorly performing topics since average effectiveness masks huge variance
  - maintain a traditional ad hoc task in TREC
- Task
  - 250 topics
    - 200 old topics from TRECs 6-8, TREC 2003 robust
    - 50 of the old topics distinguished as difficult
    - 50 new topics created for track by NIST assessors
  - TREC 6-8 document collection: disks 4&5 (no CR)
  - standard trec-eval plus measures from TREC 2003
  - systems also required to predict topic difficulty

Best Description-Only Runs, Combined Topic Set



Predicting Difficulty



Difference in MAP scores between perfect & actual prediction

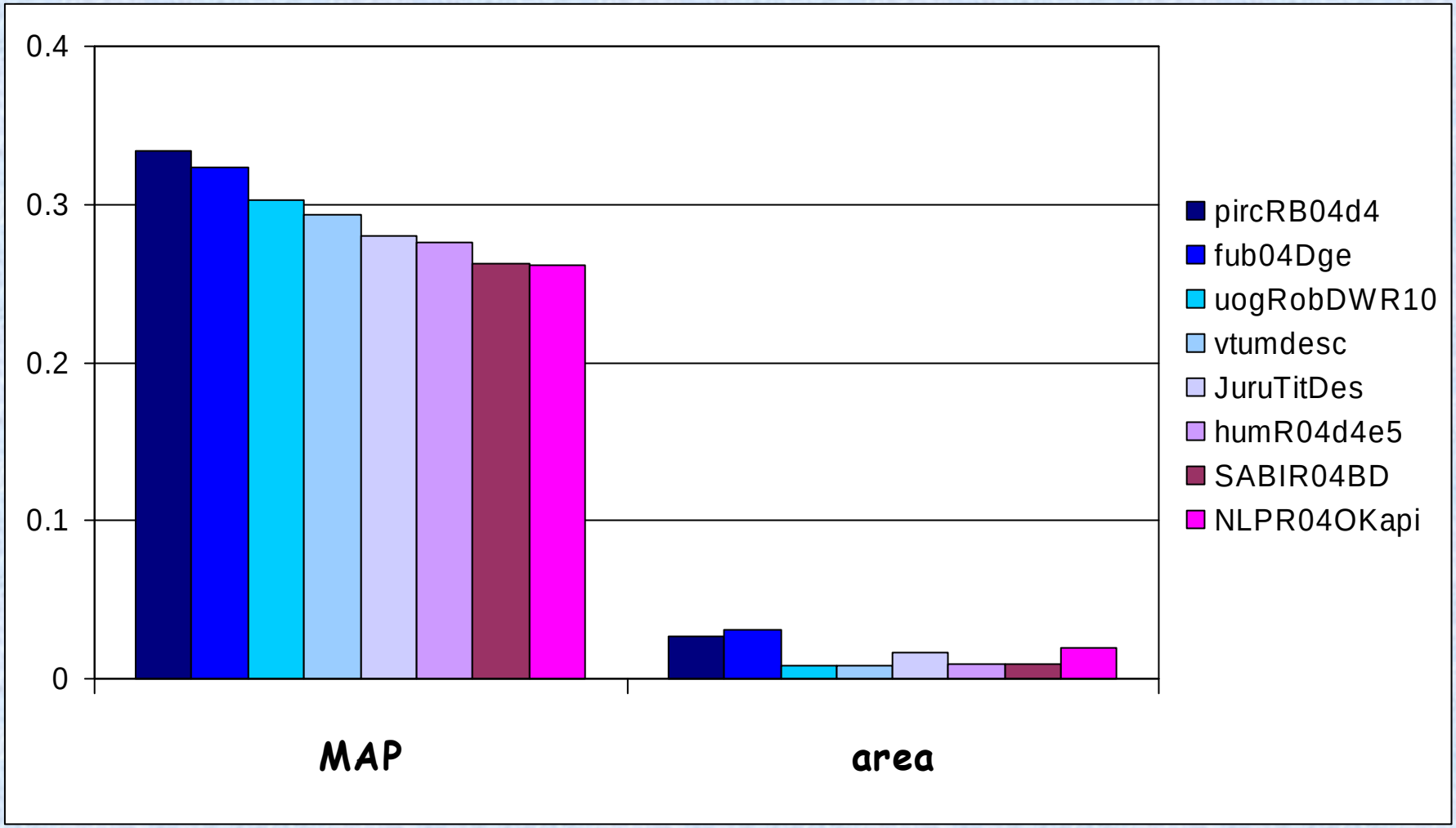
TREC 2003 Findings

- Confirmed that optimizing average effectiveness improves the already-effective topics
- Introduced new measures
  - measures do emphasize the poorly performing topics, but…
  - …measures are unstable with 50 topics

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

Rankings by MAP & Area



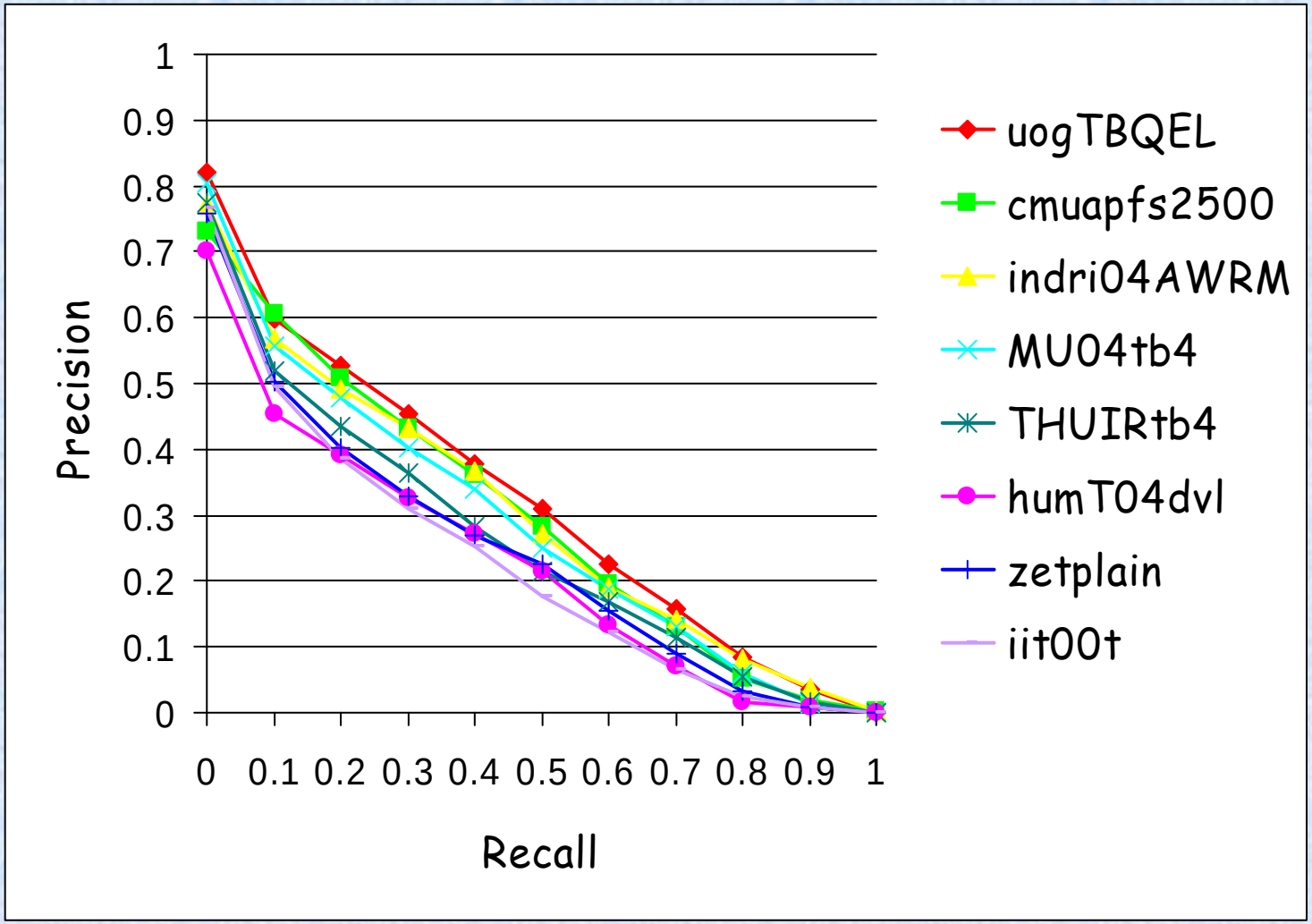
Terabyte Track

- Motivations
  - investigate evaluation methodology for collections substantially larger than existing TREC collections
  - provide test collection for exploring system issues related to size
- Task
  - traditional ad hoc retrieval task
  - systems also required to report various timing and resource statistics

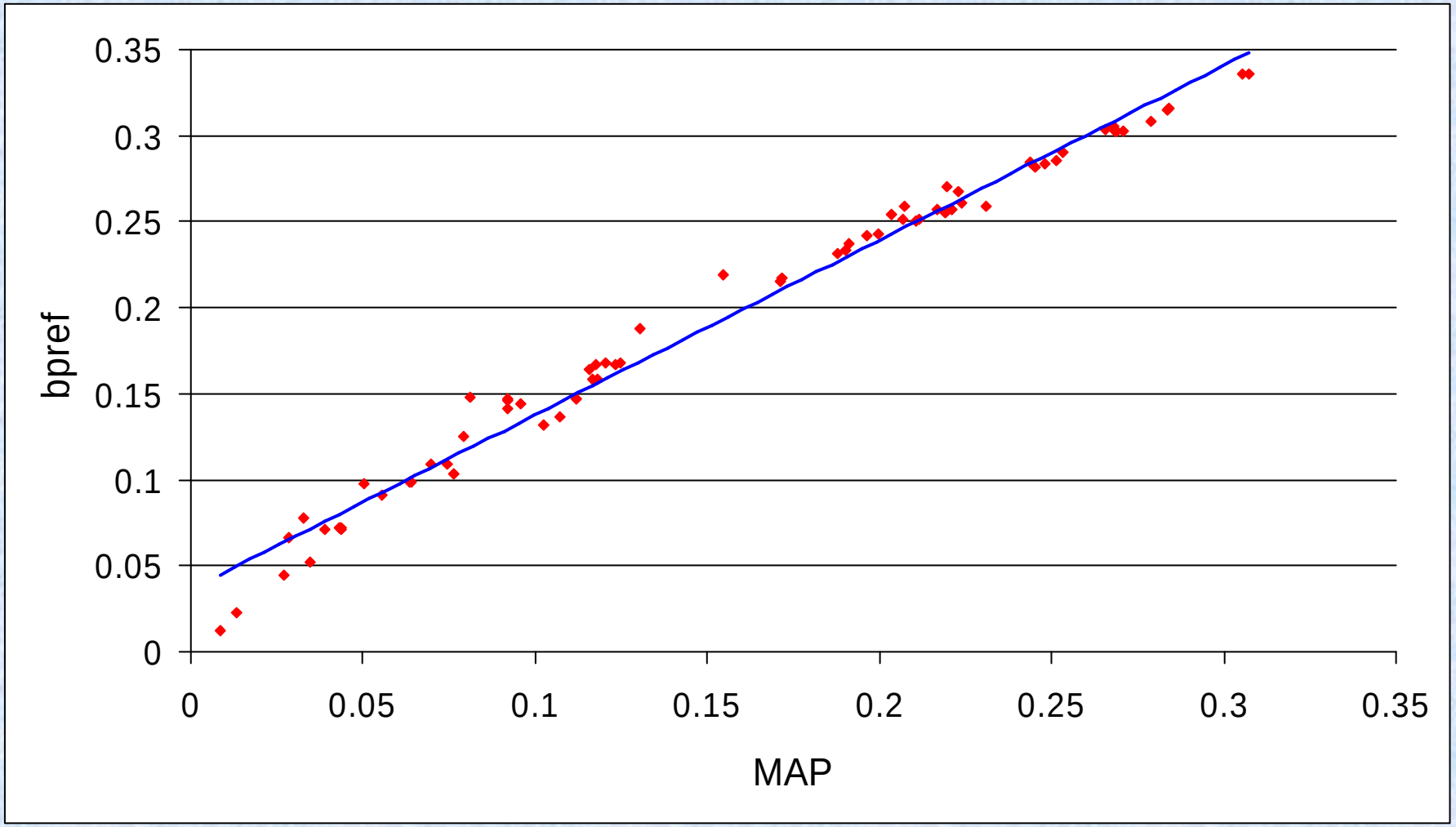
Terabyte Collection

- Documents
  - ~ 25,000,000 web documents (426 gb)
  - spidered in early 2004 from .gov domain
  - includes text from pdf, word, etc. files
- Topics
  - 50 topics created by NIST assessors
  - standard information-seeking requests
- Relevance judgments
  - performed on pooled results (top 85 from 2 runs per group)
  - time consuming!
    - 30 topics judged by Nov 1; in end, 49 topics judged

Top Terabyte Runs



MAP vs. bpref



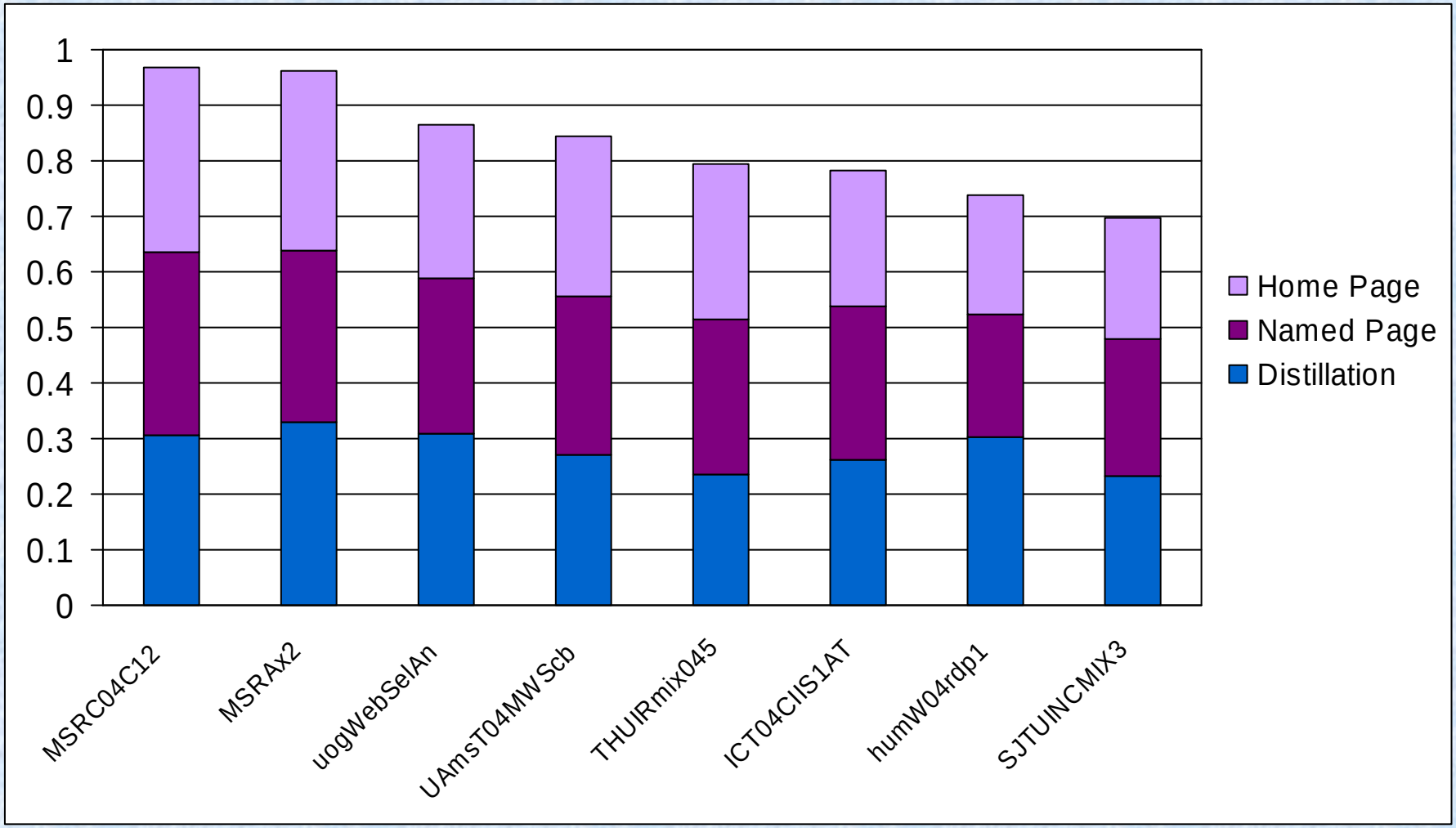
Web Track

- Investigate retrieval behavior on the web
- Two tasks
  - mixed query:
    - 225 queries; 75 each of topic distillation, named page finding, and home page finding
    - systems not told the type of a given query
  - classification: categorize queries by type
- Document set
  - crawl of .GOV created for TREC 2002 web track
  - approx. 18 GB, 1.25 million docs

Mixed Query Task

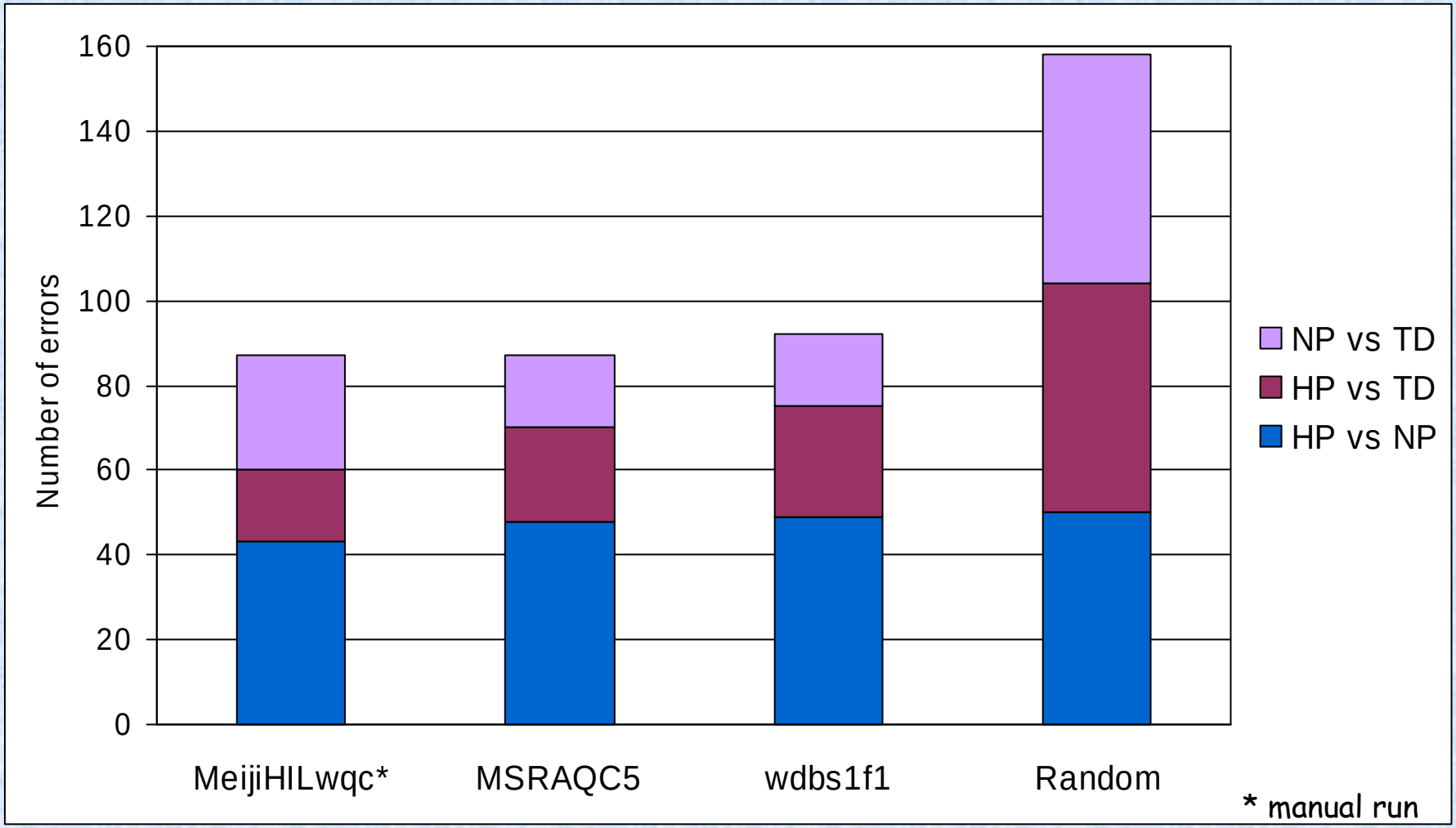
- Combined version of previous years' tasks
- Process:
  - assessors create topic of given type
  - type recorded, but not released
  - systems return ranked list of docs per topic
  - results evaluated at NIST based on recorded type
- Binary judgments by topic author
  - topic distillation: good resource page?
  - home page: correct target page (or alias)?
  - named page: correct target page? [target is not a home page]
- Evaluation
  - MAP (=MRR) & Success@{1,5,10}

Mixed Query Task Results



Normalized Average MAP-MRR Scores

Classification Task



Errors by category type

Future

- TREC will continue
- Tracks selected for TREC 2005 by PC from proposals:
  - genomics, HARD, QA, robust terabyte continuing
  - web track mutates to enterprise search track
  - add new spam track