

Overview of TREC 2013



Ellen Voorhees

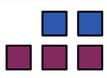
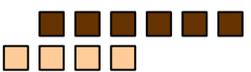
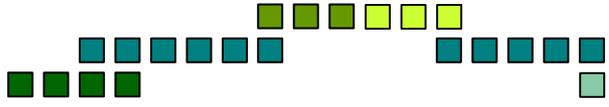
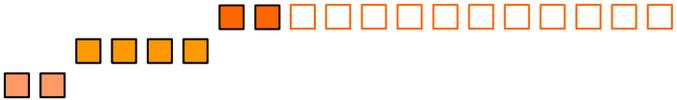
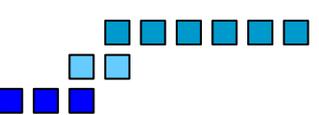
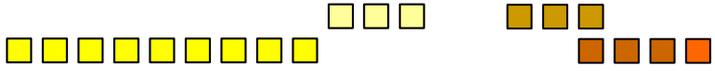
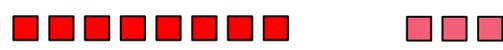
NIST

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

Back to our roots, writ large

- KBA, Temporal Summarization, Microblog
 - original TIPSTER foci of detection, extraction, summarization
 - TDT, novelty detection
- Federated Web Search
 - federated search introduced in Database Merging track in TRECs 4-5
- Web
 - web track in various guises for ~15 years
 - risk-minimization recasts goal of Robust track
- Crowdsourcing
 - re-confirmation of necessity of human judgments to distinguish highly effective runs

TREC TRACKS

		 Contextual Suggestion Crowdsourcing
Personal documents		Blog, Microblog Spam
Retrieval in a domain		Chemical IR Genomics, Medical Records
Answers, not documents		Novelty, Temporal Summary QA, Entity
Searching corporate repositories		Legal Enterprise
Size, efficiency, & web search		Terabyte, Million Query Web VLC, Federated Search
Beyond text		Video Speech OCR
Beyond just English		Cross-language Chinese Spanish
Human-in-the-loop		HARD, Feedback Interactive, Session
Streamed text		Filtering, KBA Routing
Static text		Ad Hoc, Robust
	1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013	

TREC 2013 Track Coordinators

- **Contextual Suggestion:** Adriel Dean-Hall, Charlie Clark, Jaap Kamps, Nicole Simone, Paul Thomas
- **Federated Web Search:** Thomas Demeester, Djoerd Hiemstra, Dong Nguyen, Dolf Trieschnigg
- **Crowdsourcing:** Gabriella Kazai, Matt Lease, Mark Smucker
- **Knowledge-Base Population:** John Frank, Steven Bauer, Max Kleiman-Weiner, Dan Roberts, Nilesh Tripuraneni
- **Microblog:** Miles Efron, Jimmy Lin
- **Session:** Ashraf Bah, Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas,
- **Temporal Summarization:** Javed Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Virgil Pavlu, Tetsuya Sakai
- **Web:** Paul Bennett, Charlie Clarke, Kevyn Collins-Thompson, Fernando Diaz

TREC 2013 Program Committee

Ellen Voorhees, chair

James Allan

David Lewis

Chris Buckley

Paul McNamee

Ben Carterette

Doug Oard

Gord Cormack

John Prager

Sue Dumais

Ian Soboroff

Donna Harman

Arjen de Vries

Diane Kelly

TREC 2013 Participants



Albalqa' Applied U.
Bauhaus U. Weimar
Beijing Inst. of Technology (2)
Beijing U. of Posts & Telecomm
Beijing U. of Technology
CWI
Chinese Academy of Sci.
Democritus U. Thrace
East China Normal U.
Georgetown U.
Harbin U. of Science & Technology
Indian Statistical Inst. (3)
IRIT
IIIT
Jiangsu U.
JHU HLTCOE
Kobe U.
LSIS/LIA
Microsoft Research
National U. Ireland Galway
Northeastern U.
Peking U.
Qatar Computing Research Inst.
Qatar U.
RMIT U.
Santa Clara U.
Stanford U. (2)
Technion
TU Delft
U. of Amsterdam
U. of Chinese Academy of Sciences
U. of Delaware (2)
U. of Florida
U. of Glasgow (2)
U. of Illinois, Urbana-Champaign
U. of Indonesia
U. of Lugano
U. of Massachusetts Amherst
U. of Michigan
U. of Montreal
U. of N. Carolina Chapel Hill
U. Nova de Lisboa
U. of Padova
U. of Pittsburgh
U. of Sao Paulo
U. of Stavanger
U. of Twente
U. of Waterloo (2)
U. of Wisconsin
Wuhan U.
York U.
Zhengzhou Information Technology Inst.



A big thank you to our assessors
(who don't actually get security vests)

Streaming Data Tasks

- Search within a time-ordered data stream
 - Temporal Summarization
 - widely-known, sudden-onset events
 - get reliable, timely updates of pertinent information
 - KBA
 - moderately-known, long duration entities
 - track changes of pre-specified attributes
 - Microblog
 - arbitrary topic of interest, X
 - "at time T, give me most relevant tweets about X"

KBA StreamCorpus

- Used in both TS and KBA tracks
- 17 months (11,948 hours) time span
 - October 2011-Feb 2013
 - >1 billion documents each with absolute time stamp that places it in the stream
- News, social (blog, forum,...), web (e.g., arxiv, linking events) content
- ~60% English [or language unknown]
- hosted by Amazon Public Dataset service

Temporal Summarization

- Goal: efficiently monitor the information associated with an event over time
 - detect sub-events with low latency
 - model information reliably despite dynamic, possibly conflicting, data streams
 - understand the sensitivity of text summarization algorithms and IE algorithms in online, sequential, dynamic settings
- Operationalized as two tasks in first year
 - Sequential Update Summarization
 - Value Tracking

Temporal Summarization

- 10 topics (events)
 - each has a single type taken from {accident, shooting, storm, earthquake, bombing}
 - each type has a set of attributes of interest (e.g., location, deaths, financial impact)
 - each has title, description (URL to Wikipedia entry), begin-end times, query

Topic 4

title: Wisconsin Sikh temple shooting

url: http://en.wikipedia.org/wiki/Wisconsin_Sikh_temple_shooting

begin: 1344180300

end: 1345044300

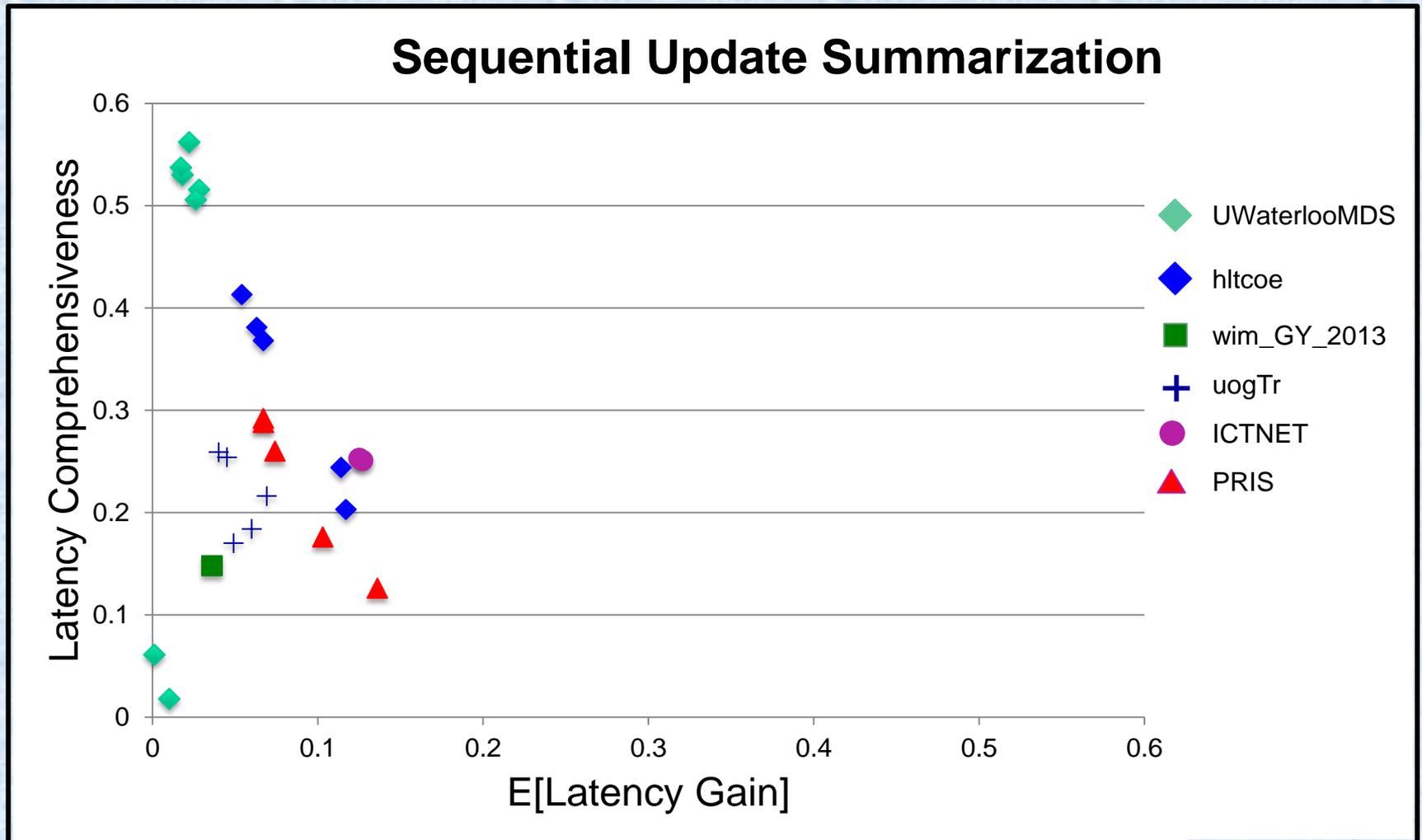
query: sikh temple shooting

type: shooting

Temporal Summarization

- Sequential Update Summarization task
 - system publishes a set of “updates” per topic
 - an update is a time-stamped extract of a sentence in the corpus
 - information content in a set of updates is compared to the human-produced gold standard information nuggets for that topic
 - evaluation metrics reward salience and comprehensiveness while penalizing verbosity, latency, irrelevance

Temporal Summarization



Temporal Summarization

- Value Tracking Task
 - for each topic-type-specific attribute, issue an update with an estimate of the attribute's value when the value changes
 - effectiveness generally not good
 - most runs concentrated on some subset of attributes (but metric defined over all)
 - metric also sensitive to the occasional very bad estimate, which systems made

Knowledge-Base Acceleration

- Entity-centric filtering
 - assist humans with KB curation task
 - i.e., keep entity profiles current
 - entity = object with strongly typed attributes
- 2013 tasks
 - Cumulative Citation Recommendation (CCR)
 - return documents that report a fact that would change the target's existing profile
 - Streaming Slot Filling (SSF)
 - extract the change itself:
 - both attribute type and new value of attribute

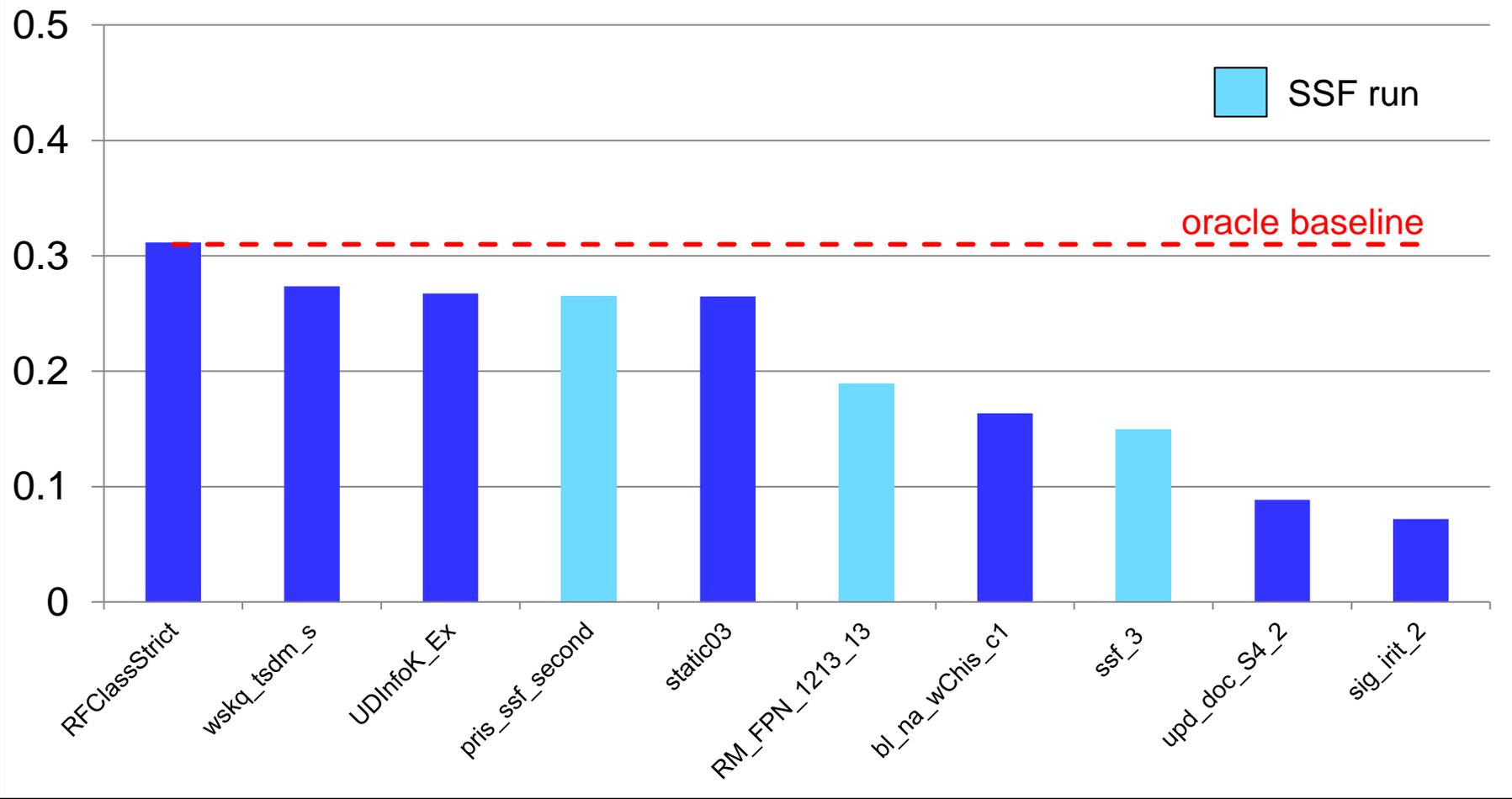
KBA

- 141 Target entities
 - 98 people, 19 organizations, 24 facilities
 - drawn from Wikipedia or Twitter
 - 14 inter-related communities
(e.g., Fargo, ND; Turing award winners)
- Systems return doc & confidence-score
 - confidence scores define retrieved sets for eval
- Evaluation
 - F, scaled utility on returned set
 - CCR: computed with respect to set of 'vital' documents
 - SSF: computed with respect to correct slot fills

KBA

Max over confidence level of average F, vital-only

Best run for top 10 groups



Microblog

- **Goal**
 - examine search tasks and evaluation methodologies for information seeking behaviors in microblogging environments
- **Started in 2011**
 - 2011 & 2012 used Tweets2011 collection
 - 2013 change to search as service model for document set access

Microblog

- Real-time ad hoc search task
 - real-time search: query issued at a particular time and topic is about something happening at that time
 - 59 new topics created by NIST assessors
 - [title, triggerTweet] pairs
 - triggerTweet defines the "time" of the query
 - triggerTweet may or may not be relevant to query
 - systems return score for all tweets issued prior to trigger Tweet's time
 - scoring: MAP, P(30), R-prec

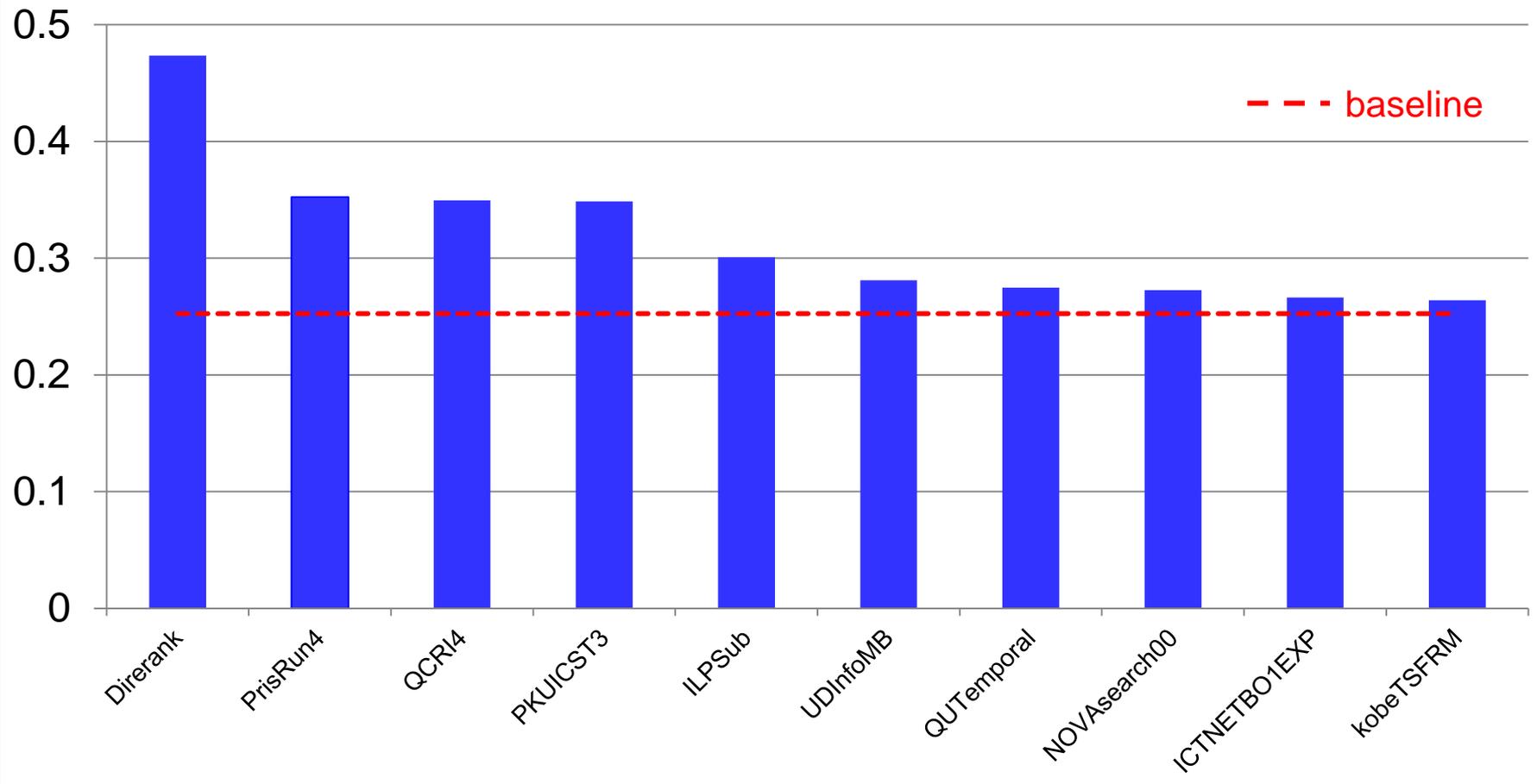
```
Query: water shortages
querytime: Fri Mar 29 18:56:02 +0000 2013
querytweettime: 317711766815653888
```

Microblog

- Search as Service model
 - motivation:
 - increase document set size by an order of magnitude over Tweets2011 (16mil->243mil) while complying with Twitter TOS
 - implementation:
 - centrally gather sample of tweets from Feb 1-Mar 31, 2013
 - provide access to set through Lucene API
 - API accepts query string and date, returns ranked list of matching tweets (plus metadata) up to specified date

Microblog

Best run by MAP for top 10 groups



ClueWeb12 Document Set

- Successor to ClueWeb09
 - ~733 million English web pages crawled by CMU between Feb 10—May 10, 2012
- Subset of collection (approx. 5% of the pages) designated as 'Category B'
- Freebase annotations for the collection are available courtesy of Google
- Used in remaining TREC 2013 tracks
 - sole document set for Session, Web, Crowdsourcing
 - part of collection for Contextual Suggestion, Federated Web Search

Contextual Suggestion

- “Entertain Me” app: suggest activities based on user’s prior history and current location
- Document set: open web or ClueWeb
- 562 profiles, 50 contexts
- Run: ranked list of up to 50 suggestions for each pair in cross-product of profiles, contexts

Contextual Suggestion

- Profile:
 - a set of judgment pairs, one pair for each of 50 example suggestions, from one person
 - example suggestions were activities in Philadelphia, PA defined by a URL with an associated short textual description
 - an activity was judged on a 5-point scale of interestingness based on the description and then based on the full site
 - a profile obtained from 500 Turkers and 62 members of the U. of Waterloo community

Contextual Suggestion

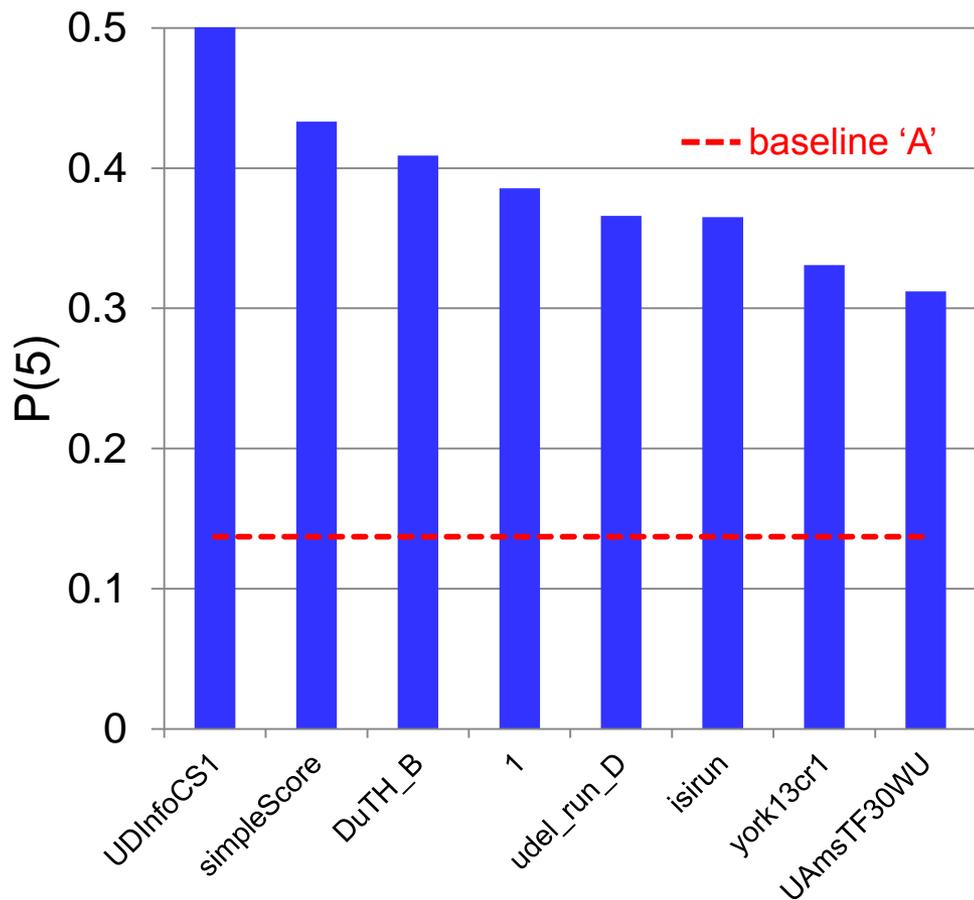
- Context
 - a randomly selected US city (excluding Phila.)
- Submitted suggestions
 - system-selected URL and description
 - ideally, description personalized for target profile

Contextual Suggestion

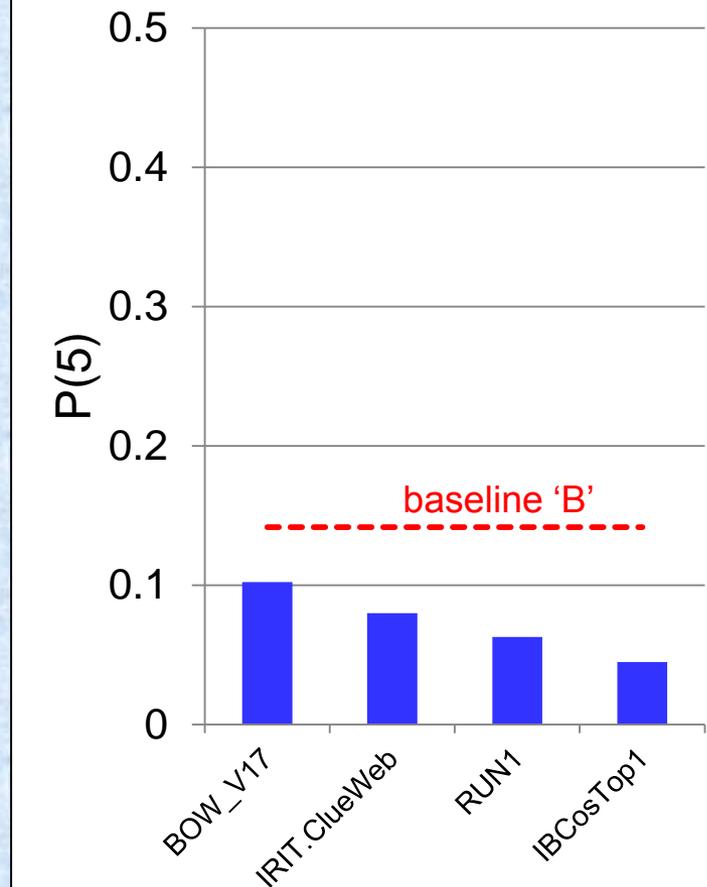
- Judging
 - separate judgments for profile match, geographical appropriateness
 - NIST assessors judged geo appropriateness
 - profile owner judged profile match and geo appropriateness
 - 223 profile-context pairs judged to depth 5
- Evaluation
 - P(5), MRR, Time-Biased Gain (TBG)
 - TBG measure penalizes actively negative suggestions and captures distinction between description and URL

Contextual Suggestion

Open Web



ClueWeb



Web

- Investigate Web retrieval technology
 - authentic web queries
 - (new) maximize effectiveness overall, without harming effectiveness for individual queries as compared to a quality baseline
- 2013 topics
 - total of 50 topics, half multi-faceted and half single-faceted
 - all topics developed from queries/query clusters observed in operational web engines' logs
 - participants receive simple query string only

Web

Faceted Topic

ham radio

how do you get a ham radio license?

1. <same as description>
2. What are the ham radio license classes?
3. How do you build a ham radio station?
4. Find information on ham radio antennas.
5. What are the ham radio call signs?
6. Find the web site of Ham Radio Outlet.

Single-facet Topics

i will survive

find the lyrics to the song "I Will Survive"

beef stroganoff recipe

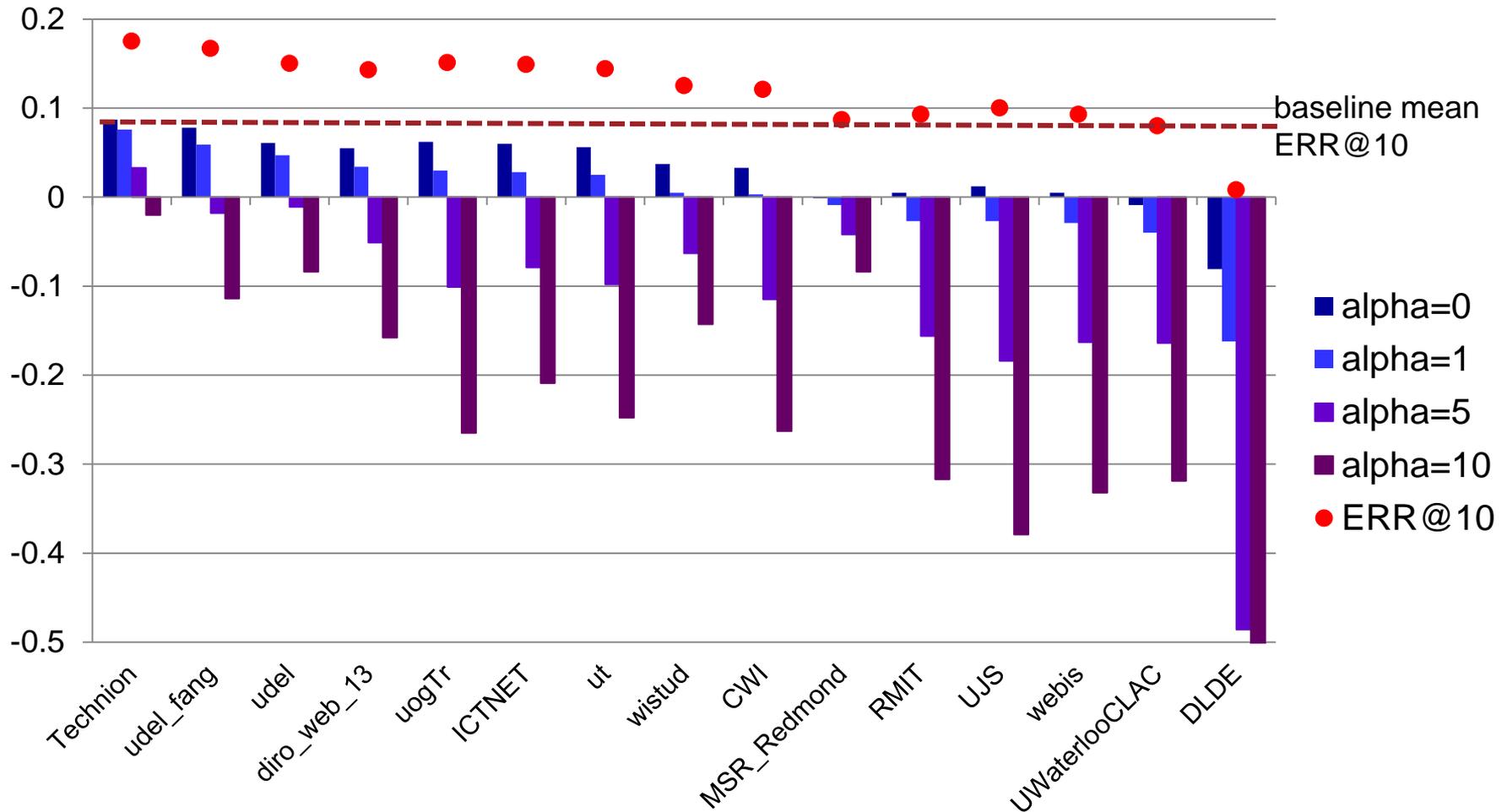
find complete (not partial) recipes for beef stroganoff

- Assessors judge pages with respect to each facet on 6-point scale
- Ad hoc search effectiveness measures: traditional, graded, diversity (e.g., MAP, nDCG@20, ERR-IA)
- Risk-sensitive task measure rewards high average effectiveness, and penalizes losses relative to baseline
 - α -parameter controls relative importance of mean effectiveness and risk penalty: $\alpha=0$ no penalty; larger α , more penalty

Web

Mean ERR@10 and Δ from baseline's ERR@10

best score across team's submissions



Crowdsourcing

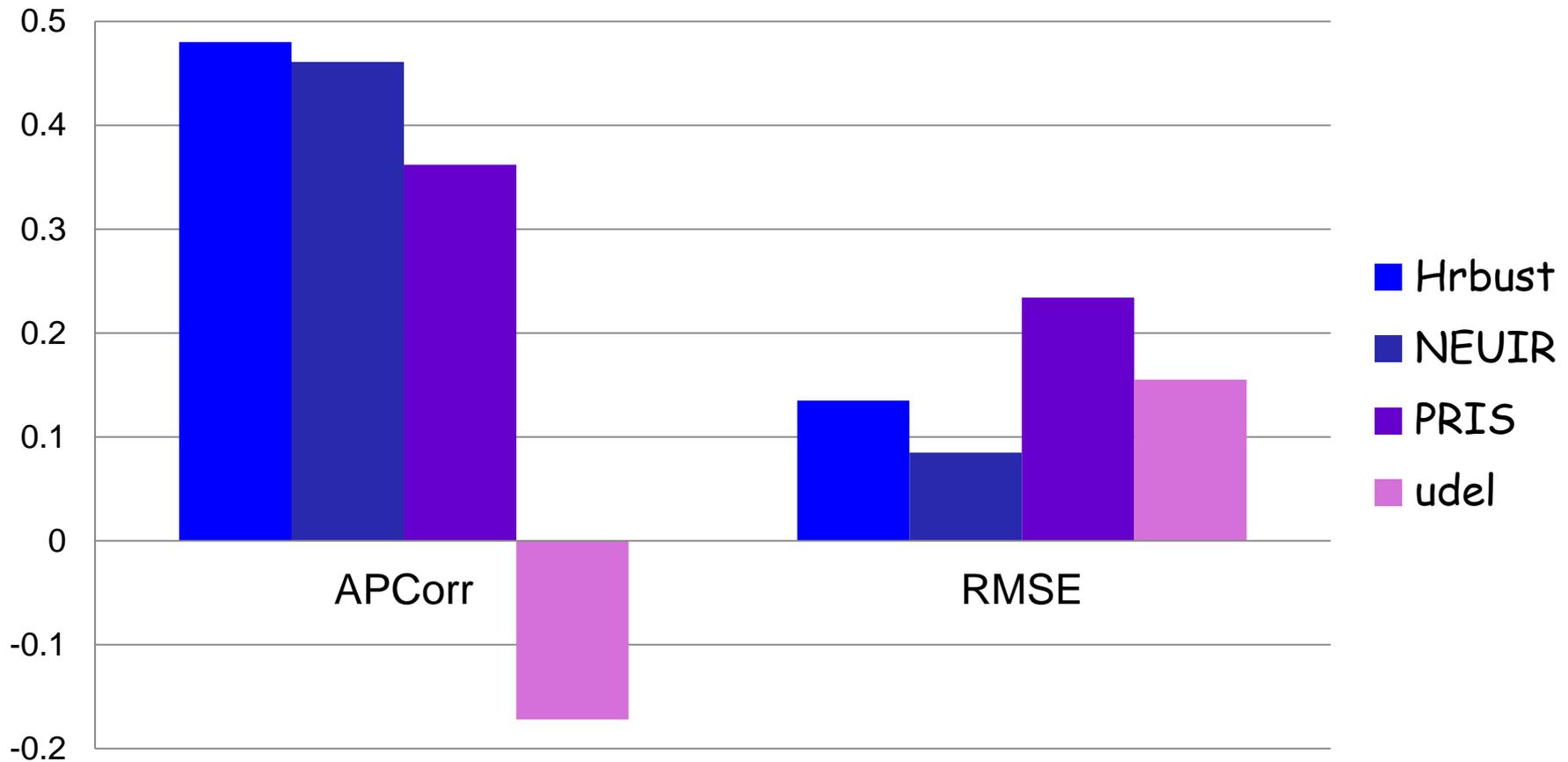
- A meta-track:
 - investigate best practices for using crowdsourcing to build IR evaluation resources...
 - ... though while crowdsourcing is the focus, actual goal is to produce judgments in a reliable, scalable manner by any combination of means
- 2013 task
 - do 2013 Web track judging
- Sponsors
 - thanks to Amazon and Crowd Computing Systems who offered track participants credits or discounted prices for track work

Crowdsourcing

- Web track judging
 - pools created for Web track for NIST assessors distributed to crowdsourcing participants, too
 - Crowd participants get judgments for those pools
 - first subtopic only for multi-faceted topics
 - subset of only 10 topics as "basic" version of task (~3.5k documents vs. ~20k documents for full 50 topic set)
 - quality of participant judgments evaluated in three ways, each using NIST judgments as gold standard
 - correlation of rankings when web track runs evaluated using NIST judgments & participant's judgments
 - RMSE of actual score values as computed by the two judgment sets
 - difference in labels themselves, as measured by GAP

Crowdsourcing

APCorr, RMSE computed using mean ERR@20 for 34 web track runs
(10 topics)

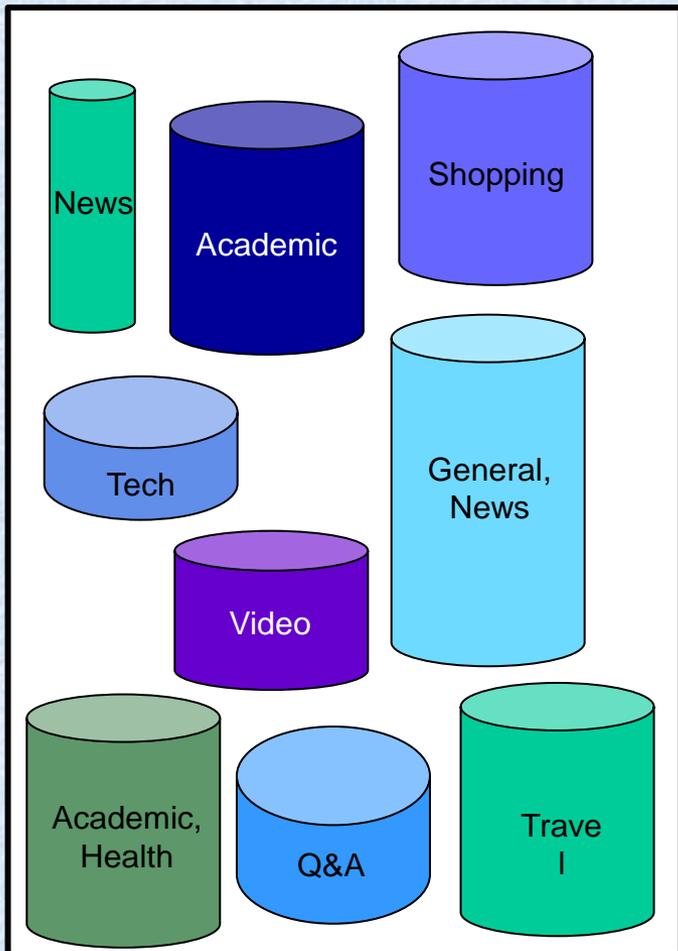


Federated Web Search

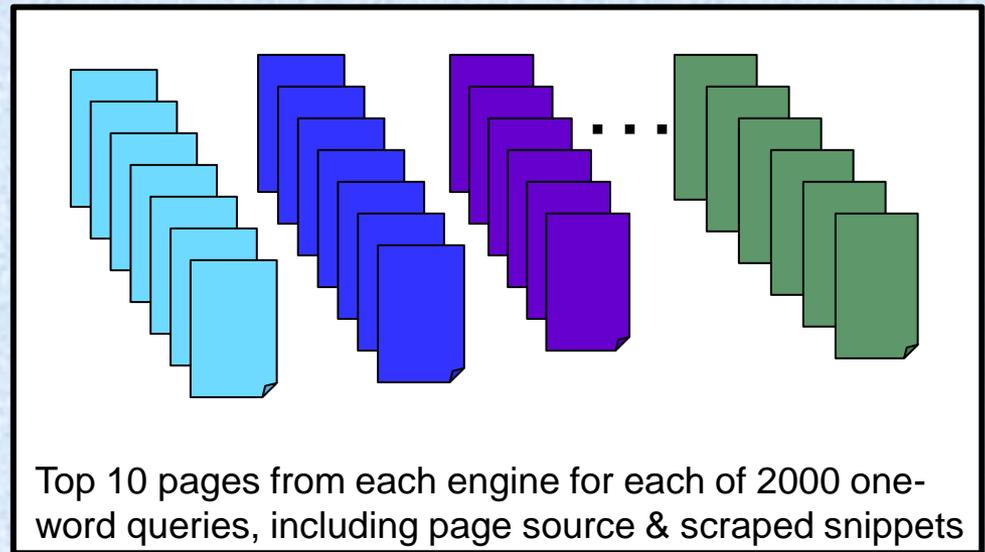
- New track for 2013
- Goal: promote research in federated search in realistic web setting
 - two tasks in initial year:
 - resource selection: pick engines to receive query
 - result merging: create document list from different engines' responses

Federated Web search

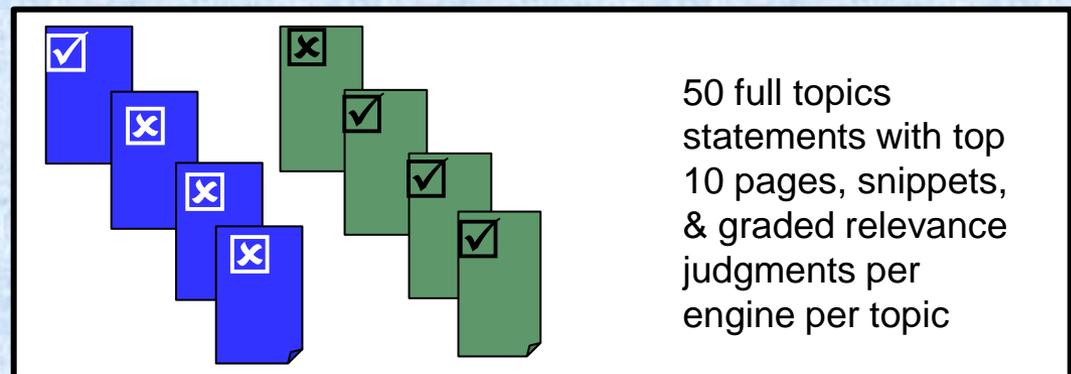
157 search engines in
24 categories



Sampled Collection



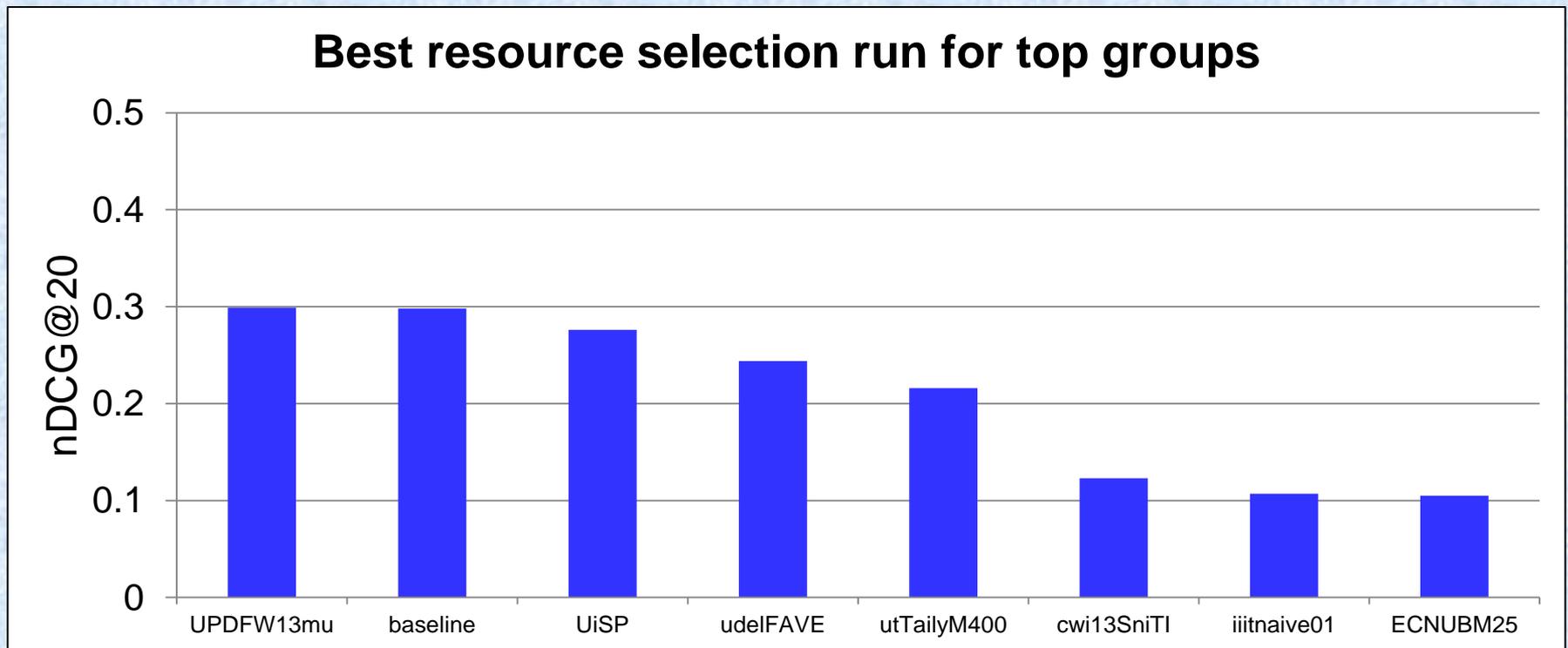
Test Collection



Federated Web Search

- Resource Selection

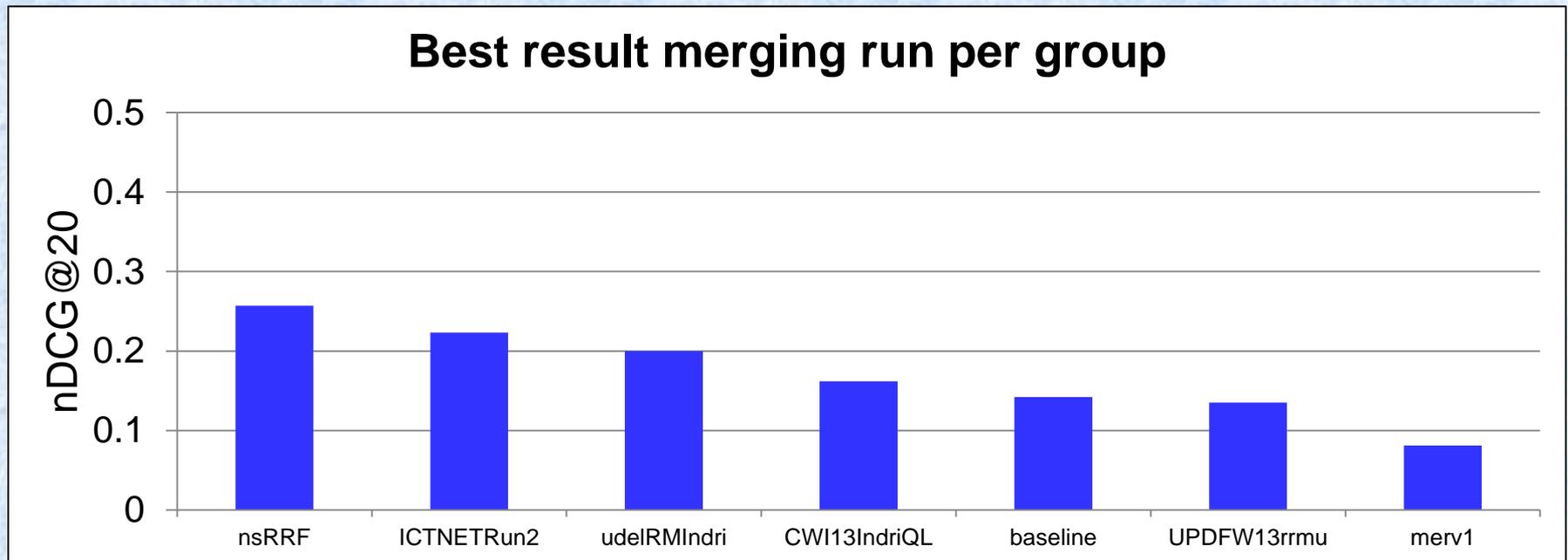
- rank 157 search engines per topic (having no access to test collection retrieval results)



Federated Web Search

- **Results Merging**

- produce a ranked list of page results per topic
- most submissions treated as a re-ranking problem over available results. More realistic federated search task is a significantly harder task.



Session

- **Goal**

study users' interaction over a set of related searches rather than single query

- **TREC 2013**

- best possible result list for final query in session
- single submission consists of 3 rankings (per session), one for each experimental condition

R1: result produced using final query text only

R2: result produced using any data in current session

R3: result produced using any data in all sessions

Session

- Topic set engineered along 2 dimensions
product {intellectual, factual} x goal quality {specific, amorphous}
 - Known-item search; Known-subject search;
Interpretive search; Exploratory search
- 61 topics total across the four types
- Humans searched for answers using instrumented search engine
 - resulted in 87 multiple-query sessions
 - additional 46 single-query session also released
 - session data includes queries, result lists, and clicks, all time-stamped from session start

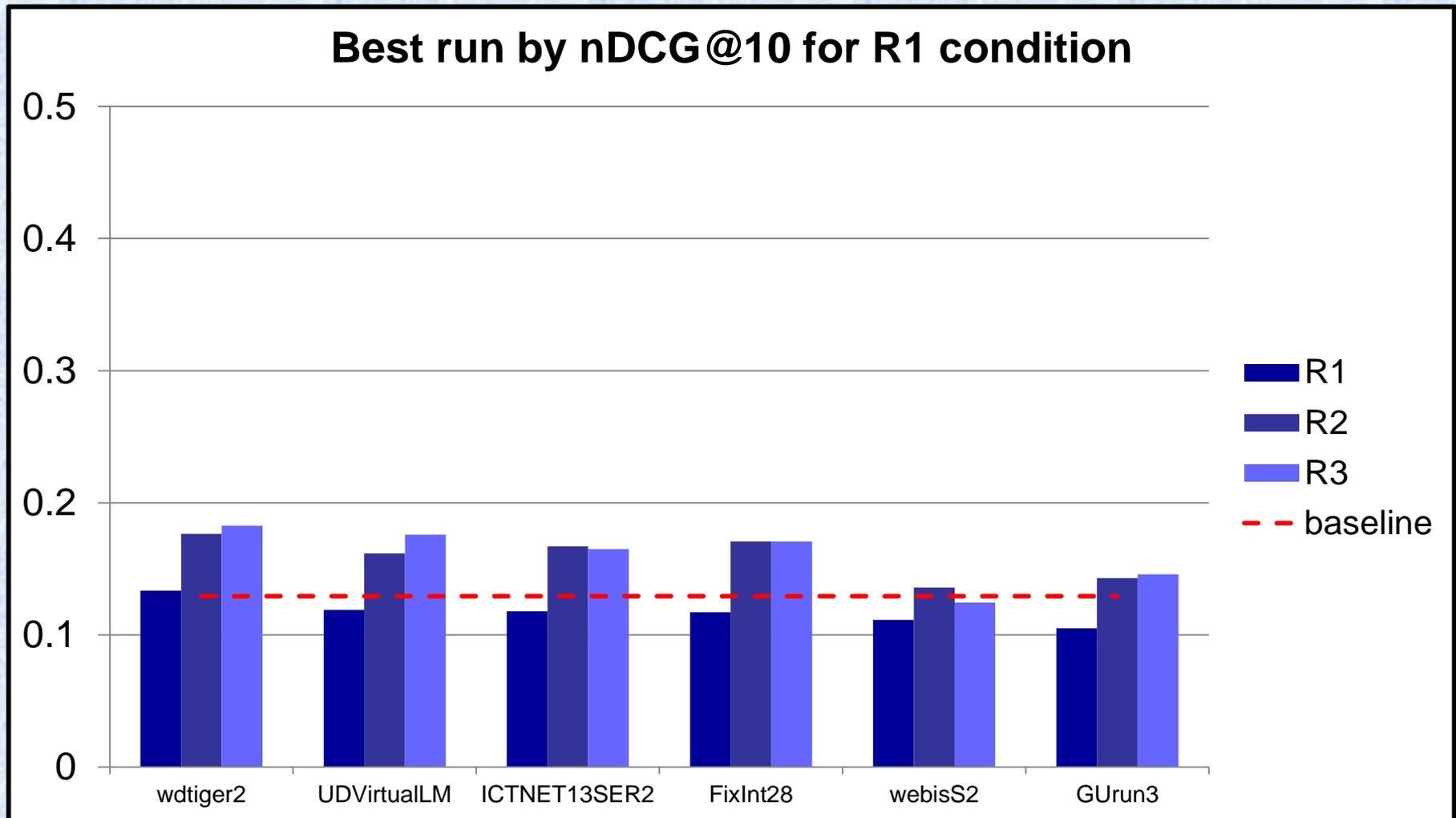
Session

- Evaluation

- judgment set for a given topic created from union of all documents encountered in session data (for all sessions associated with topic), plus top 10 docs from all ranked lists submitted for those sessions
- documents judged on 6-point scale on basis of topic as a whole

Session

Best run by nDCG@10 for R1 condition



TREC 2014

- Tracks
 - all tracks except Crowdsourcing continuing
 - new track on Clinical Decision Support
- TREC 2013 track planning sessions
 - 1.5 hours per track tomorrow (four-way parallel)
 - track coordinators attending 2013
 - you can help shape task(s); make your opinions known

Comic Sans
is never an
acceptable font.
Unless you are
an 8 year old
girl writing
a poem about
unicorns.

