Overview of TREC 2014

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

TREC 2014 Track Coordinators

Clinical Decision Support: Matthew Simpson, Ellen Voorhees, Bill Hersh

Contextual Suggestion: Adriel Dean-Hall, Charlie Clark, Jaap Kamps, Paul Thomas

Federated Web Search: Thomas Demeester, Djoerd Hiemstra, Dong Nguyen, Dolf Trieschnigg, Ke Zhou

Knowledge-Base Population: John Frank, Steven Bauer, Max Kleiman-Weiner, Dan Roberts

Microblog: Miles Efron, Jimmy Lin

Session: Ben Carterette, Evangelos Kanoulas, Paul Clough, Mark Hall

Temporal Summarization: Matthew Ekstrand-Abueg, Virgil Pavlu, Richard McCreadie, Fernando Diaz, Javad Aslam, Tetsuya Sakai

Web: Kevyn Collins-Thompson, Craig Macdonald, Fernando Diaz, Paul Bennett

TREC 2014 Program Committee

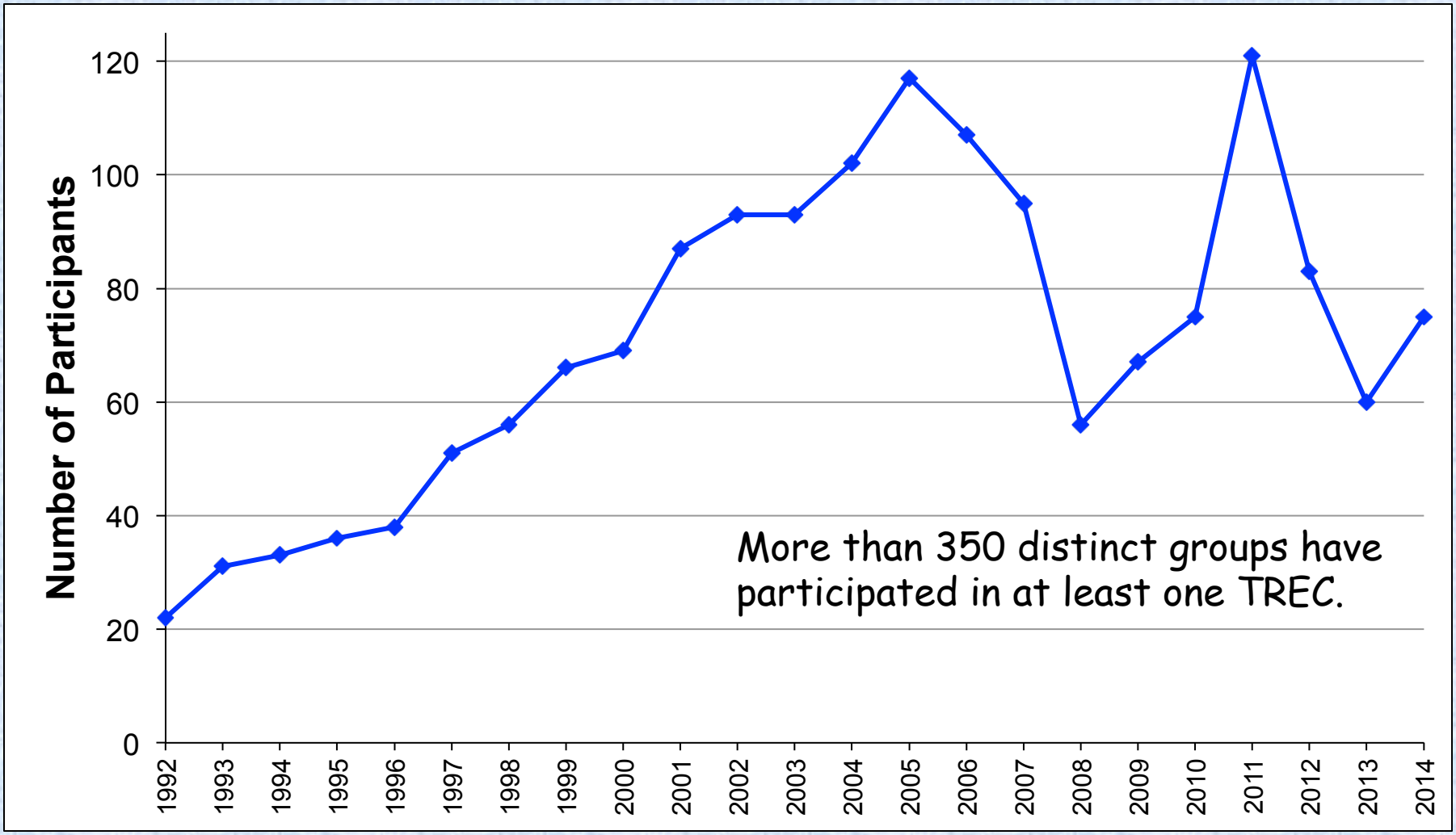
Ellen Voorhees, chair

James Allan
Chris Buckley
Ben Carterette
Gord Cormack
Sue Dumais
Donna Harman
Diane Kelly
David Lewis
Paul McNamee
Doug Oard
John Prager
Ian Soboroff
Arjen de Vries

75 TREC 2014 Participants

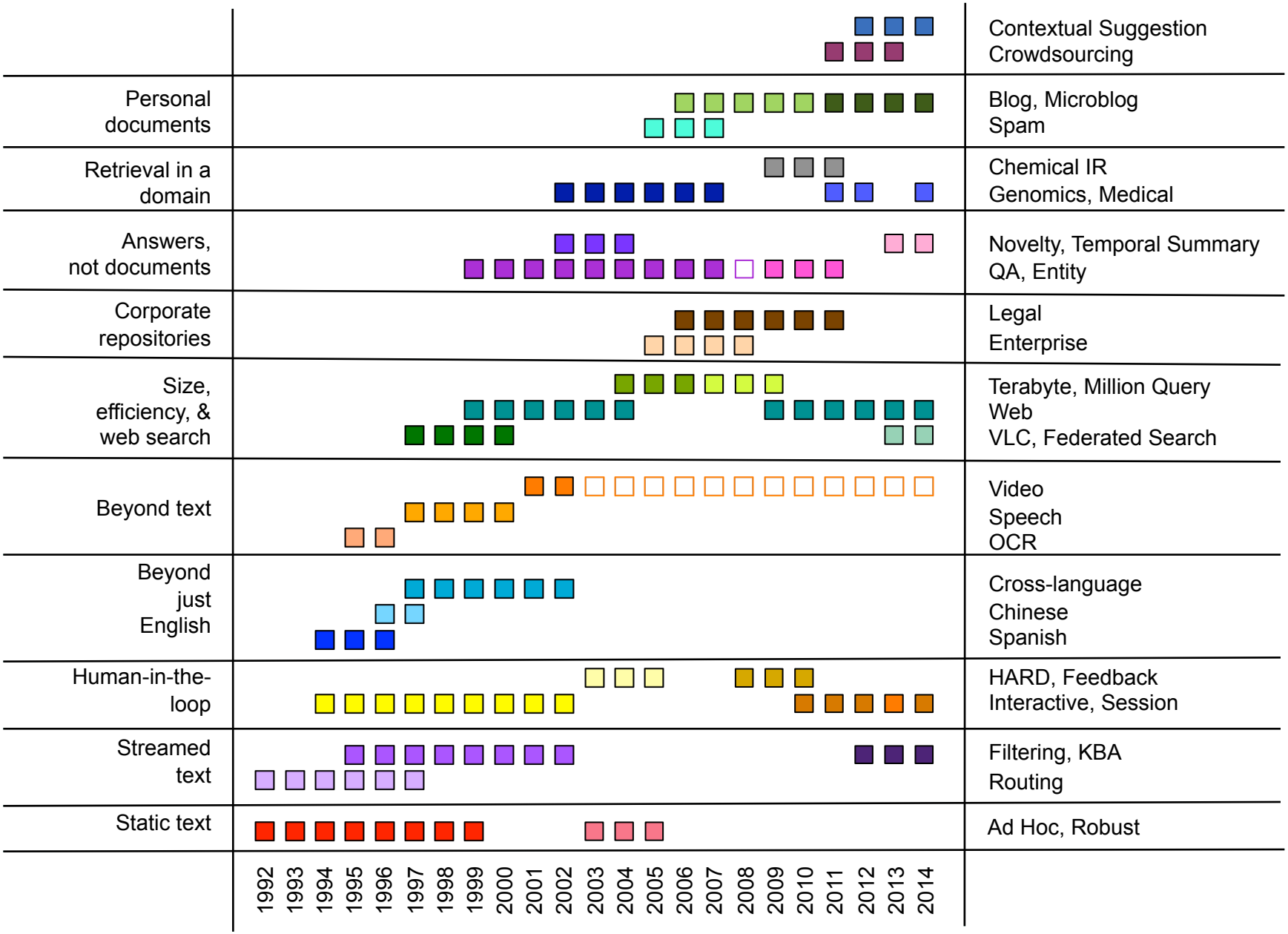
Atigeo
Bauhaus U. Weimar
Beijing Inst. of Technology
Beijing U. of Posts & Telecommunication (2)
Beijing U. of Technology
BiTeM_SIBtex, Geneva
Carnegie Mellon U. (2)
Chinese U. of Hong Kong
Chinese Academy of Sci.
Columbia U.
CRP Henri Tudor
CWI
Delft U. of Technology
Dhirubhai Ambani Inst.
Drexel U.
East China Normal U. (2)
Eindhoven U. of Tech.
Endicott College
Georgetown U. (2)
JHU HLT COE
Hebrew U. of Jerusalem
Hong Kong Polytechnic U.
Indian Inst. Tech, Varanasi
Inst. of Medical Info, NCKU
IRIT
Jiangxi U.
Korea Inst. of Sci & Tech
Kobe U.
Kware/LSIS
Leidos
LIMSI-CNRS
Merck KGaA
Microsoft Research
Oregon Health & Sci. U.
Peking U.
Phillips Research NA
Qatar Comp. Rsrch Inst
Qatar U.
Renmin U. of China
San Francisco State U.
Siena College
Santa Clara U.
Seoul Nat. U. Medical
South China U. of Tech.
Tianjin U. (2)
U. of Amsterdam
U. of Chinese Acad. Sci.
U. College London
U. of California, LA
U. of Delaware (2)
U. of Glasgow (2)
U. of Illinois
U. of Jaen
U. of Lugano
U. of Massachusetts
U. of Michigan
U. Nova de Lisboa
U. of Padova
U. of Pittsburgh
U. of Stavanger + Norwegian U. Sci & Tech
U. Texas at Dallas
U. of Twente
U. of Washington
U. of Waterloo
U. of Wisconsin + Hubei U. of Technology
Vienna U. of Technology
Wuhan U.
York U.

Participation in TREC



A big thank you to our assessors

TREC TRACKS


Contextual Suggestion
Crowdsourcing
Personal documents
Blog, Microblog
Spam
Retrieval in a domain
Chemical IR
Genomics, Medical
Answers, not documents
Novelty, Temporal Summary
QA, Entity
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
2014

Basics

- Generic tasks
  - ad hoc: known collection, unpredictable queries, response is a ranked list
  - filtering: known queries, document stream, response is a document set
- Measures
  - recall, precision are fundamental components
  - ranked list measures: MAP, nDCG@X, ERR
  - filtering measures: F, utility

Common Document Sets

- ClueWeb12 document set
  - ~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
- 2014 KBA Stream Corpus
  - 19 months (13,663 hours) Oct 2011-Apr 2013
  - ~1.2B documents each with absolute time stamp
  - news, social, web content
  - ~60% English, annotated with BBN's Serif tools
  - hosted by Amazon Public Dataset service

Clinical Decision Support

- Clinical decision support systems a piece of target Health IT infrastructure
  - aim to anticipate physicians' needs by linking health records to information needed for patient care
  - some of that info comes from biomedical literature
  - existing biomedical literature is immense, and its growth is accelerating, so it is difficult/impossible for clinicians to keep abreast

CDS Track Task

Given a case narrative, return biomedical articles that can be used to accomplish one of three generic clinical tasks:

- What is the diagnosis?
- What is the best treatment?
- What test should be run?

- Documents:
  - open access subset of PubMed Central, a database of freely-available full-text biomedical literature
  - contains 733,138 articles in NXML

CDS Track Task

- 30 topics
  - case narratives developed by NIH physicians plus label designating target clinical task
  - 10 topics for each clinical task type
  - each topic statement includes both a "description" and a shorter, more focused "summary"
  - case narratives used as an idealized medical record
- Judgments
  - judgment sets created using inferred measure sampling
  - 2 strata; ranks 1-20; 20% of 21-100
  - up to 5 runs per participant

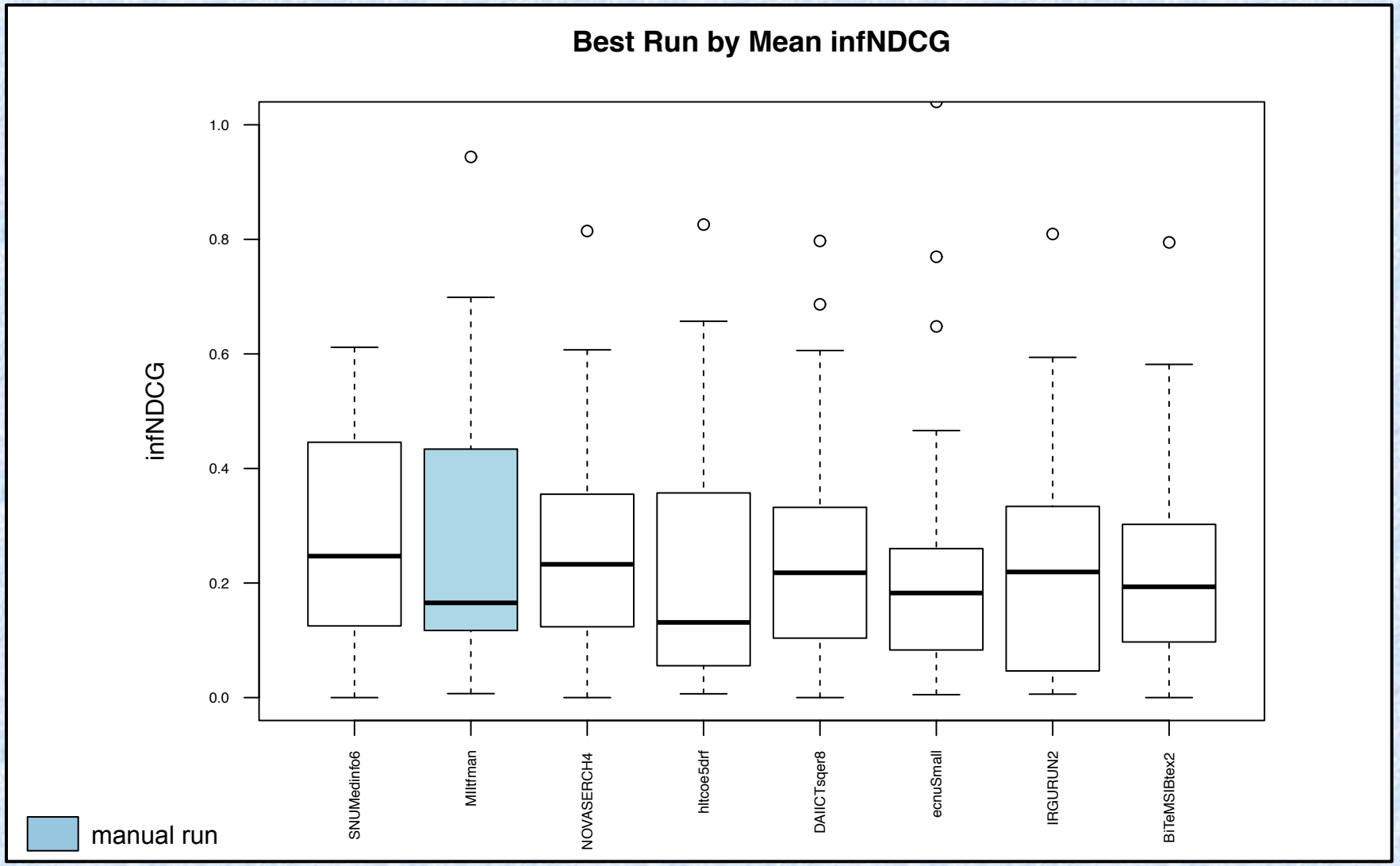
CDS Track Topics

<topic number="2" type="**diagnosis**">
Description: An 8-year-old male presents in March to the ER with fever up to 39 C, dyspnea and cough for 2 days. He has just returned from a 5 day vacation in Colorado. Parents report that prior to the onset of fever and cough, he had loose stools. He denies upper respiratory tract symptoms. On examination he is in respiratory distress and has bronchial respiratory sounds on the left. A chest x-ray shows bilateral lung infiltrates.
Summary:8-year-old boy with 2 days of loose stools, fever, and cough after returning from a trip to Colorado. Chest x-ray shows bilateral lung infiltrates.

<topic number="11" type="**test**">
Description: A 40-year-old woman with no past medical history presents to the ER with excruciating pain in her right arm that had started 1 hour prior to her admission. She denies trauma. On examination she is pale and in moderate discomfort, as well as tachypneic and tachycardic. Her body temperature is normal and her blood pressure is 80/60. Her right arm has no discoloration or movement limitation.
Summary: 40-year-old woman with severe right arm pain and hypotension. She has no history of trauma and right arm exam reveals no significant findings.

<topic number="29" type="**treatment**">
Description: A 51-year-old woman is seen in clinic for advice on osteoporosis. She has a past medical history of significant hypertension and diet-controlled diabetes mellitus. She currently smokes 1 pack of cigarettes per day. She was documented by previous LH and FSH levels to be in menopause within the last year. She is concerned about breaking her hip as she gets older and is seeking advice on osteoporosis prevention.
Summary: 51-year-old smoker with hypertension and diabetes, in menopause, needs recommendations for preventing osteoporosis.

Clinical Decision Support



Contextual Suggestion

- "Entertain Me" app: suggest activities based on user's prior history and current location
- Document set: open web or ClueWeb
- 183 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 100 example suggestions, from one person
  - example suggestions were activities in either Chicago, IL or Santa Fe, NM 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
  - profiles obtained from Mechanical Turk-ers

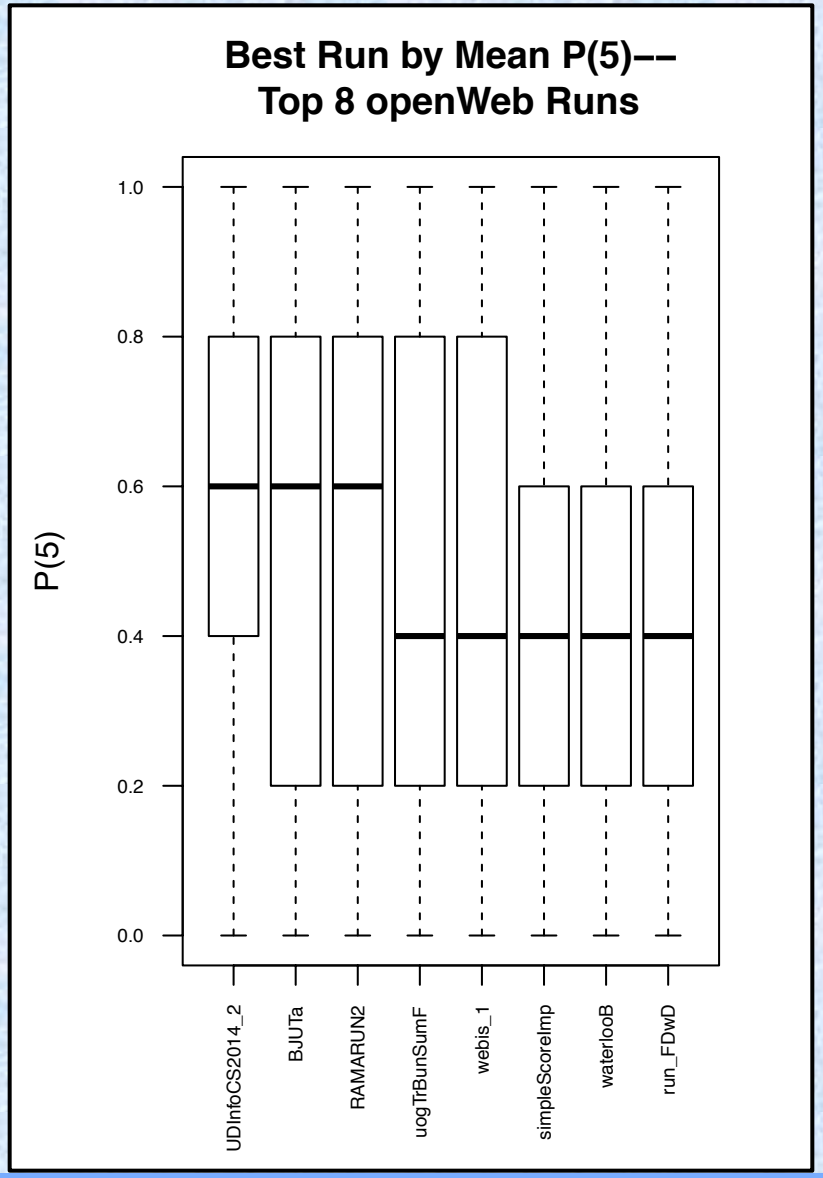
Contextual Suggestion

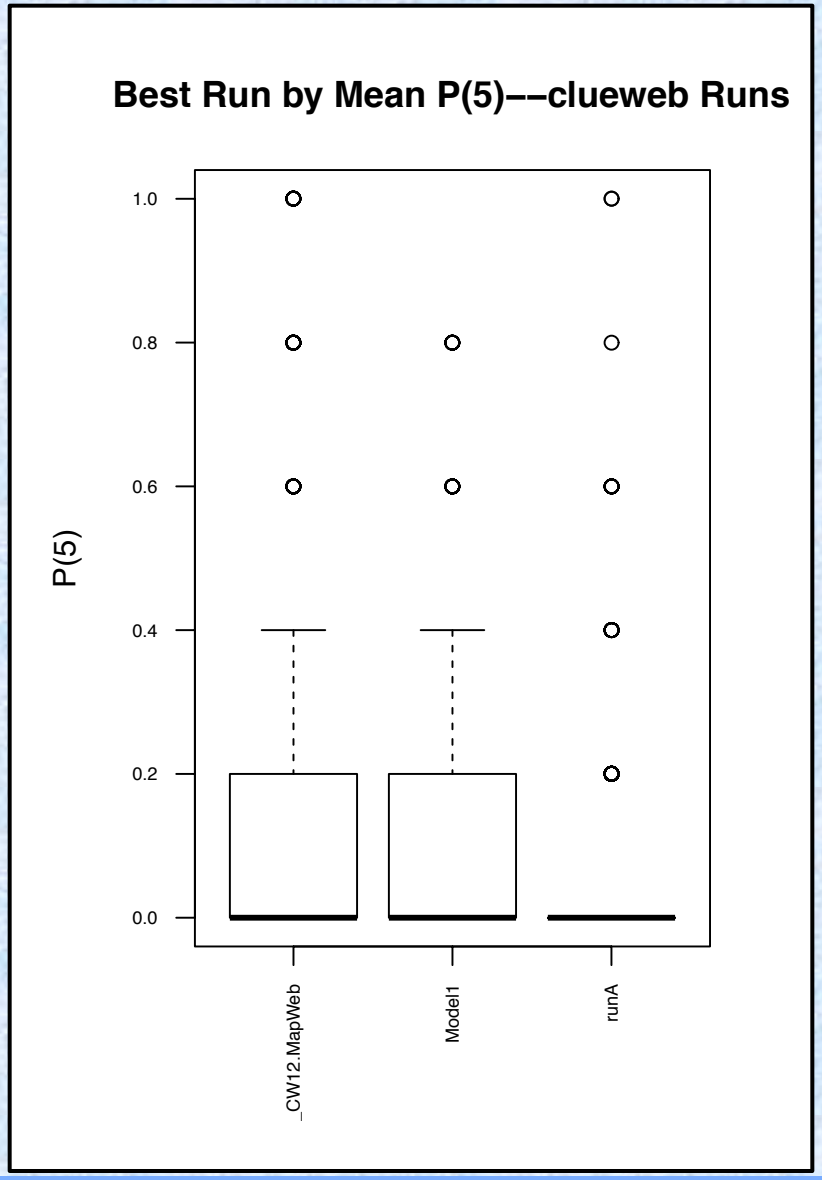
- Context
  - a randomly selected US city (excluding Phila.)
- Submitted suggestions
  - system-selected URL and description
  - ideally, description personalized for target profile

Contextual Suggestion

- Judging
  - separate judgments for profile match, geographical appropriateness
  - NIST assessors made geo judgments (48 contexts to depth 5)
  - profile owner judged profile match and geo for 299 profile-context pairs 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





Web

- Investigate Web retrieval technology
  - two tasks: ad hoc and risk minimization
  - diversity component to both tasks
  - risk-sensitive: maximize effectiveness overall w/out harming effectiveness for individual queries
- 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 sting only
- ClueWeb12 document set

Web

Faceted Topic

tooth abscess

What treatments are available for a tooth abscess?

1. <same as description>
2. What are the dangers/complications of leaving a tooth abscess untreated?
3. What are the concerns with extracting an abscessed tooth?
4. Which antibiotics are used to treat a tooth abscess?

Single-facet Topics

identifying spider bites

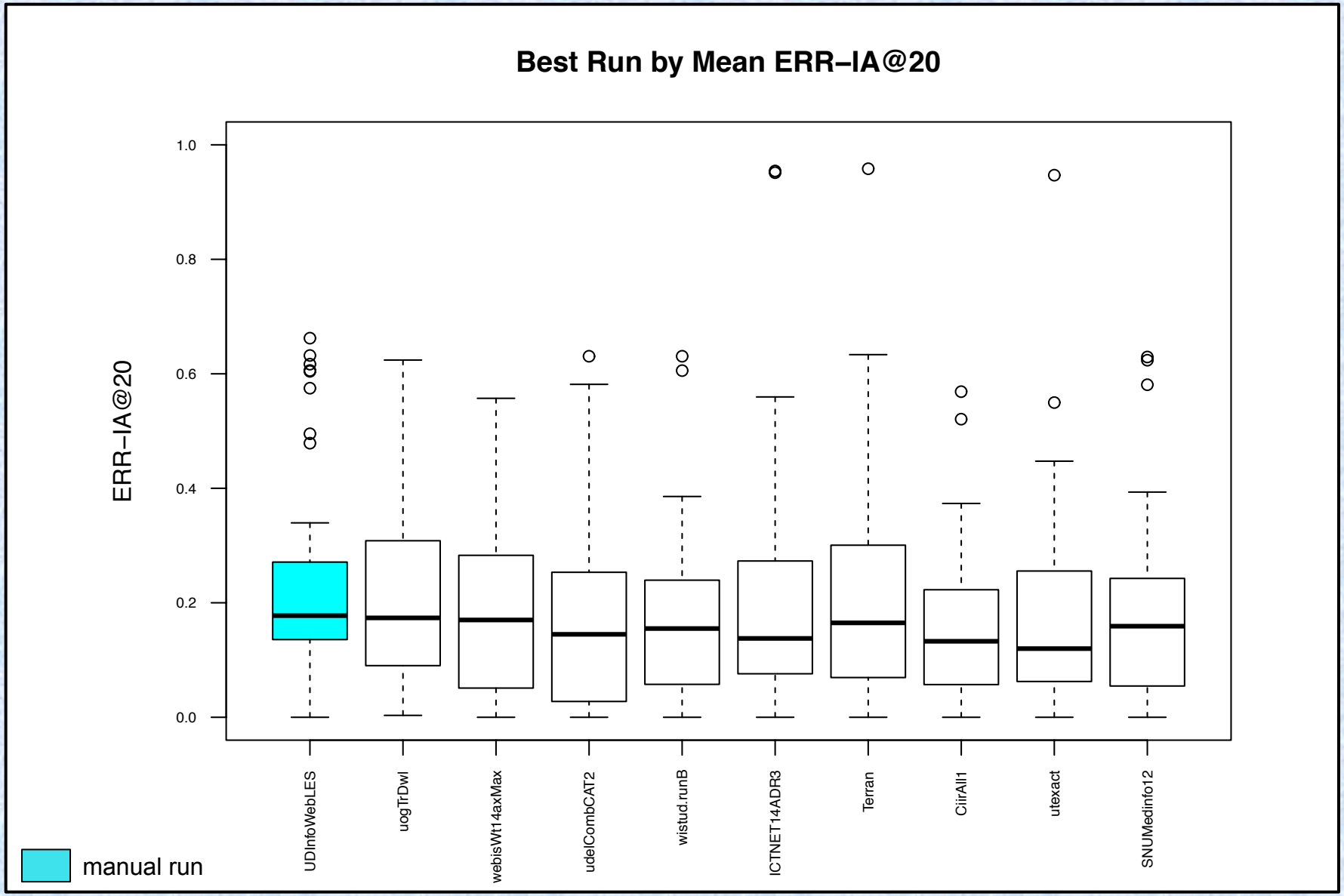
find data on how to identify spider bites

history of orcas island

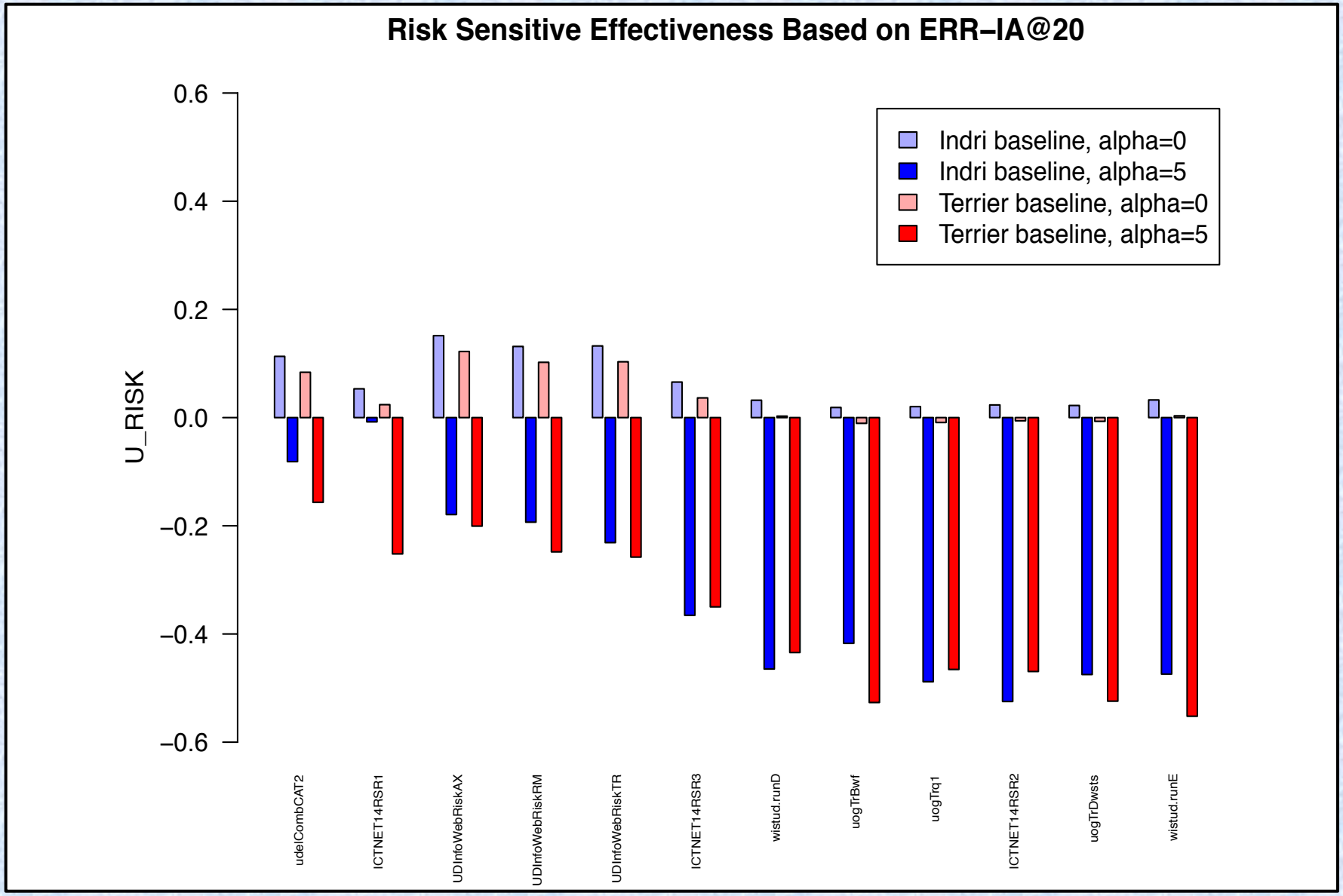
looking for any historical information related to Orcas Island: geographical, buildings, people, infrastructure, etc.

- Assessors judge pages with respect to each facet on 6-point scale
- Risk-sensitive task measure rewards high average effectiveness, and penalizes losses relative to a baseline
  - α-parameter controls relative importance of mean effectiveness and risk penalty: α=0 no penalty; larger α, more penalty
  - two baselines, Indri run and Terrier run

Web Ad Hoc Task



Web Risk-Sensitive Task

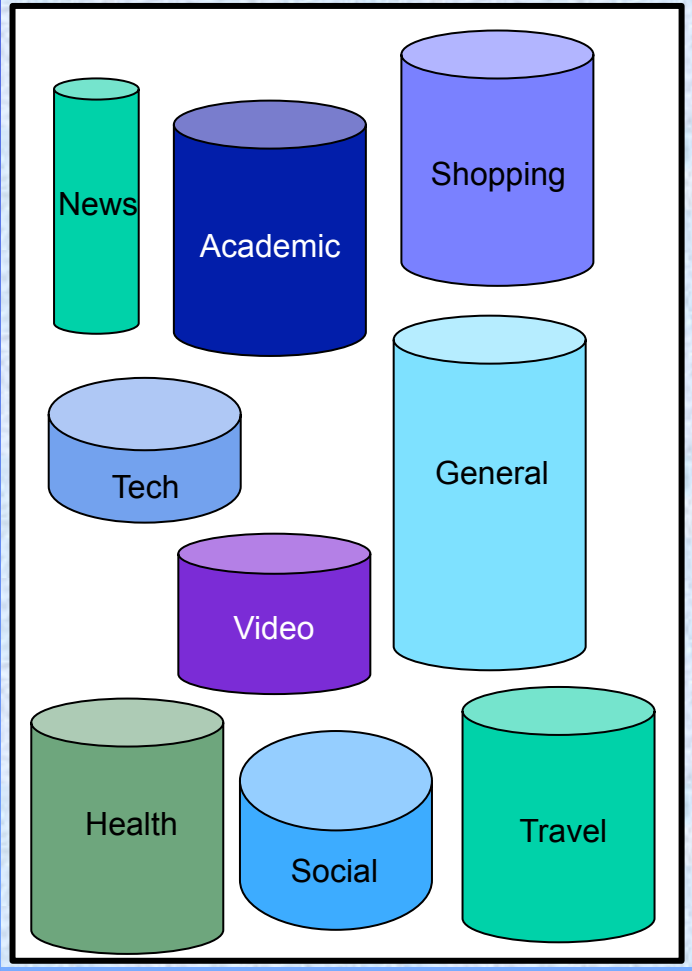


Federated Web Search

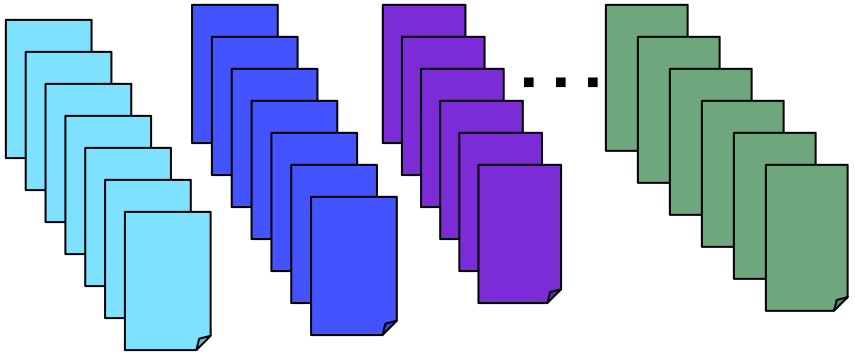
- Goal: promote research in federated search in realistic web setting
  - three tasks:
    - resource selection: pick engines to receive query (system result is ranked list of engines)
    - vertical selection: choose engine groups to receive query (categorization task: system result is a set of verticals)
    - result merging: create document list from different engines' responses (system result is a ranked list of web pages; pages restricted to top results of selected engines)

Federated Web Search

149 search engines in 24 verticals

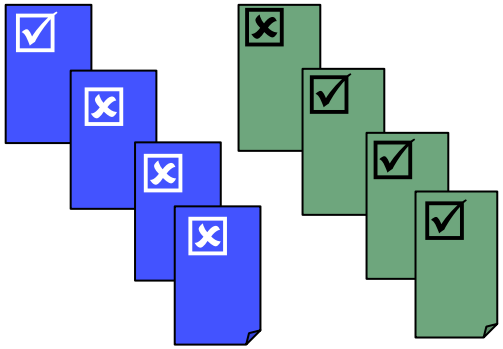


Sample Crawl



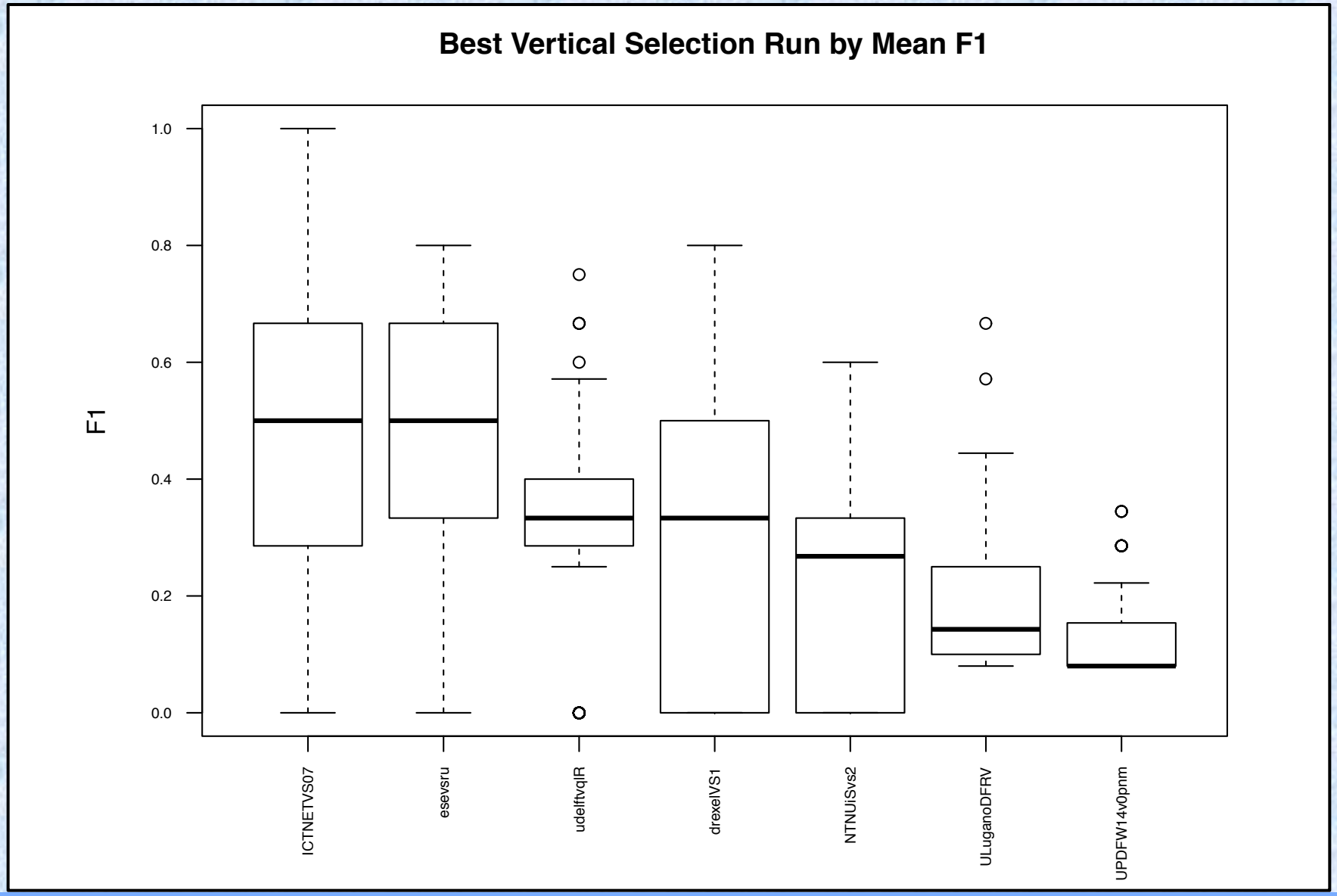
Top 10 pages from each engine for each of 4000 one-word queries, including page source & scraped snippets

Topic Crawl

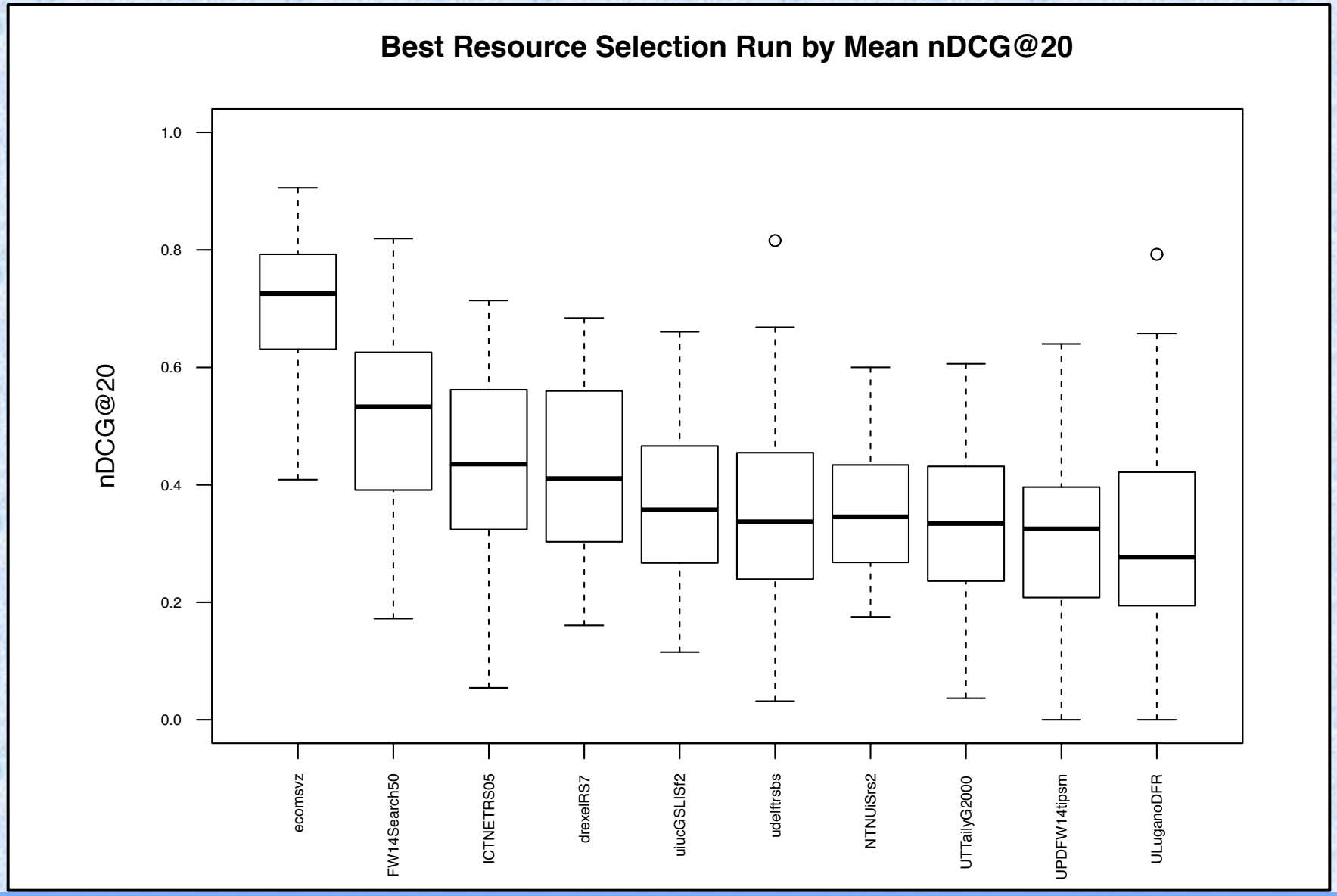


50 full topics statements with top 10 snippets. Graded relevance judgments per engine per topic

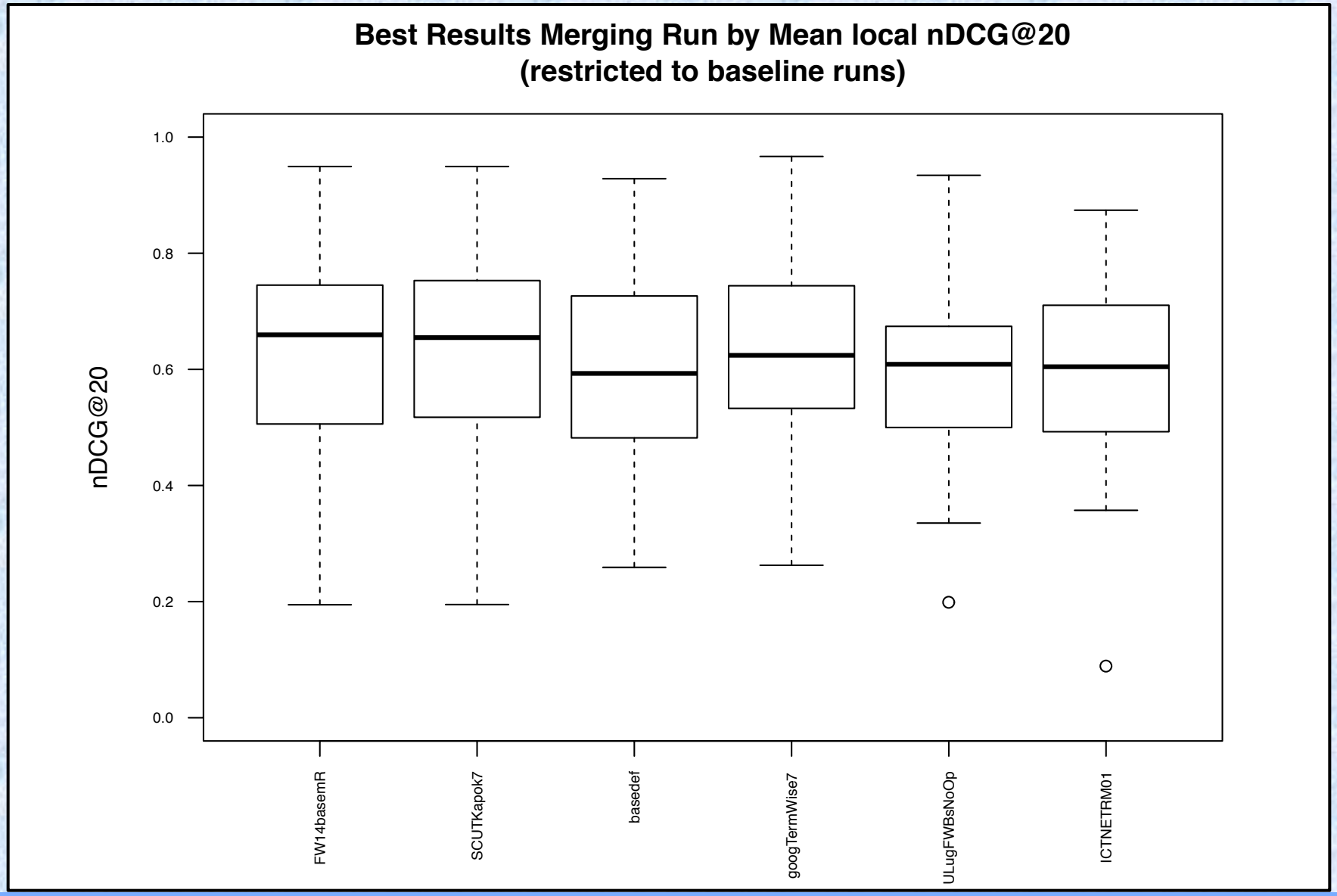
Federated: Vertical Selection Task



Federated: Resource Selection Task



Federated: Results Merging Task



Session

- Goal

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

- TREC 2014
  - best possible result list for final query in session
  - single submission consists of up to 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

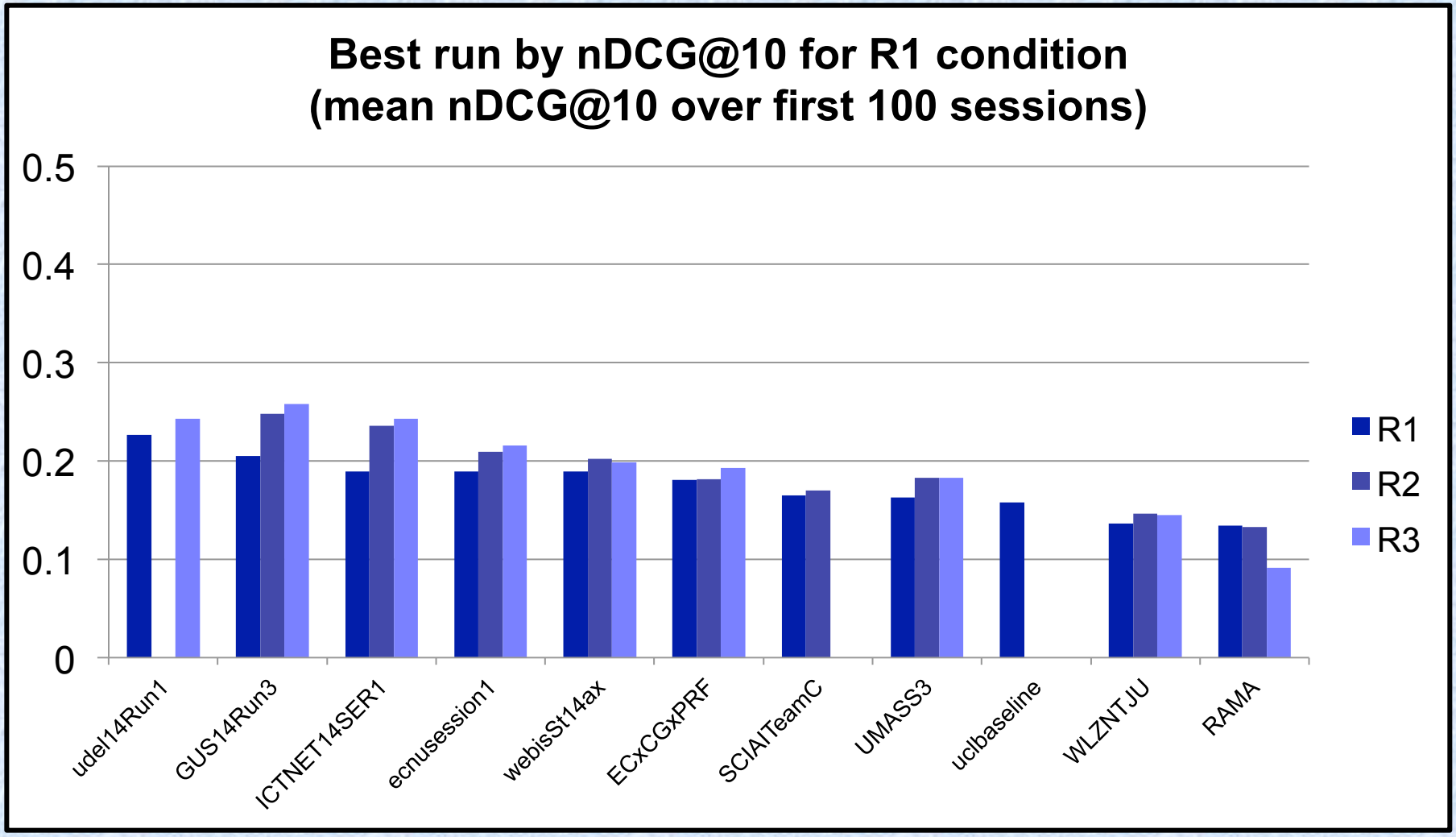
- 60 topics from previous years
  - judgment sets formed from sessions, so only 51 topics with judgments
- Mechanical Turk-ers searched for answers using instrumented search engine
  - resulted in 1,021 multiple-query sessions
    - additional 236 single-query session also released
  - session data includes queries, result lists, and clicks, all time-stamped from session start

```
<session num="10" starttime="0">
  <topic num="12">
    <desc>Your friend would like to quit smoking. You would like to provide him with relevant information about: the different ways to quit smoking, programs available to help quit smoking, benefits of quitting smoking, second effects of quitting smoking, using hypnosis to quit smoking, using the cold turkey method to quit smoking</desc>
  </topic>
  <interaction num="1" starttime="8.30123">
    <query>quit smoking</query>
    <results>
      <result rank="1">
        <url>http://quitsmoking.about.com</url>
        <clueweb12id>clueweb12-0005wb-77-27713</clueweb12id>
        <title>Quit Smoking | Quit Smoking Support | Smoking Cessation</title>
        <snippet>Quit Smoking | Quit Smoking Support | Smoking Cessation...</snippet>
      </result>
      ...
      <result rank="10">
        <url>http://www.heart.org/HEARTORG/GettingHealthy/QuitSmoking/Quit Smoking_UCM_001085
        <clueweb12id>clueweb12-0300tw-20-20611</clueweb12id>
        <title>Quit Smoking</title>
        <snippet>Quit Smoking ...0 Grams Trans Fat Oils and Fats Restaurant FAQs Other...</snippet>
      </result>
    </results>
    <clicked>
      <click num="1" starttime="12.984659" endtime="20.557844”> <rank>3</rank> </click>
      <click num="2" starttime="27.030967" endtime="55.220869”> <rank>1</rank> </click>
      <click num="3" starttime="55.220869" endtime="60.704926”> <rank>6</rank> </click>
      <click num="4" starttime="60.704926" endtime="69.165489”> <rank>5</rank> </click>
    </clicked>
  </interaction>
  <currentquery starttime="78.226578">
    <query>quit smoking cold turkey</query>
  </currentquery>
</session>
```

Session

- Runs
  - ranked lists over ClueWeb12 collection
- Evaluation
  - judgment set for a given topic created from union of all documents encountered in session data for all sessions associated with topic in the top 100 sessions, 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
(mean nDCG@10 over first 100 sessions)
0.5
0.4
0.3
0.2
0.1
0
udel14Run1
GUS14Run3
ICTNET14SER1
ecnusession1
webisSt14ax
ECxCGxPRF
SCIAITeamC
UMASS3
uclbaseline
WLZNTJU
RAMA
R1
R2
R3

Temporal Summarization

- Goal: efficiently monitor the information associated with an event over time
  - focus on widely-known, sudden-onset events
- Subtasks
  - detect sub-events with low latency
  - model information reliably despite dynamic, possibly conflicting, data streams (to detect novelty)

Temporal Summarization

- Subset of KBA Stream Corpus
- 15 topics (events)
  - each has a single type taken from {accident, bombing, hostage-taking, impact, protest, riot, shooting, storm}

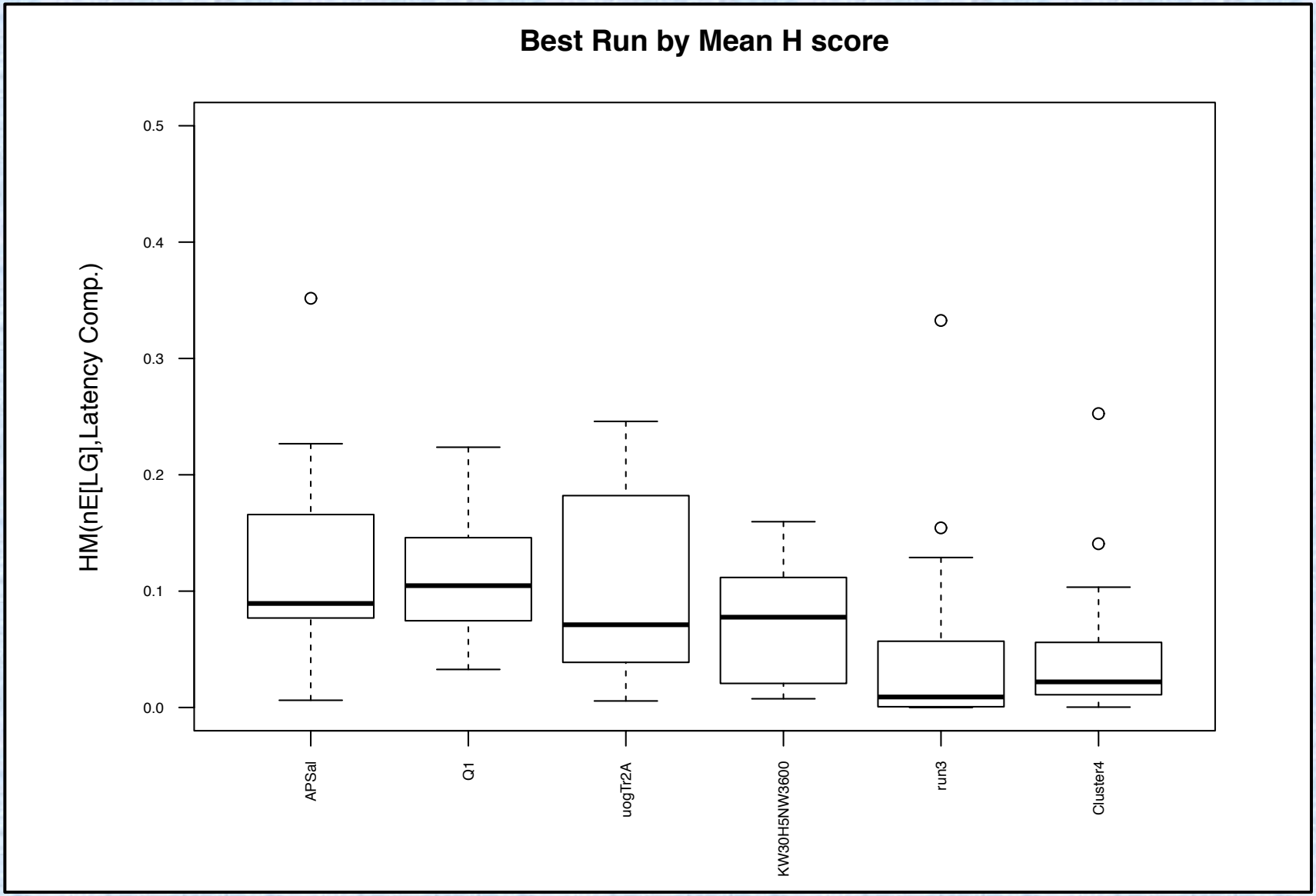
id: 15 **title:** Port Said Stadium riot
description: http://en.wikipedia.org/wiki/Port_Said_Stadium_riot
start: 1328103000 **end:** 1328967000
query: egyptian riots
type: riot

id: 21 **title:** Chelyabinsk meteor
description: http://en.wikipedia.org/wiki/Chelyabinsk_meteor
start: 1360898400 **end:** 1361762400
query: russia meteor
type: impact event

Temporal Summarization

- 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
    - normalized expected latency gain, latency comprehensiveness

Temporal Summarization



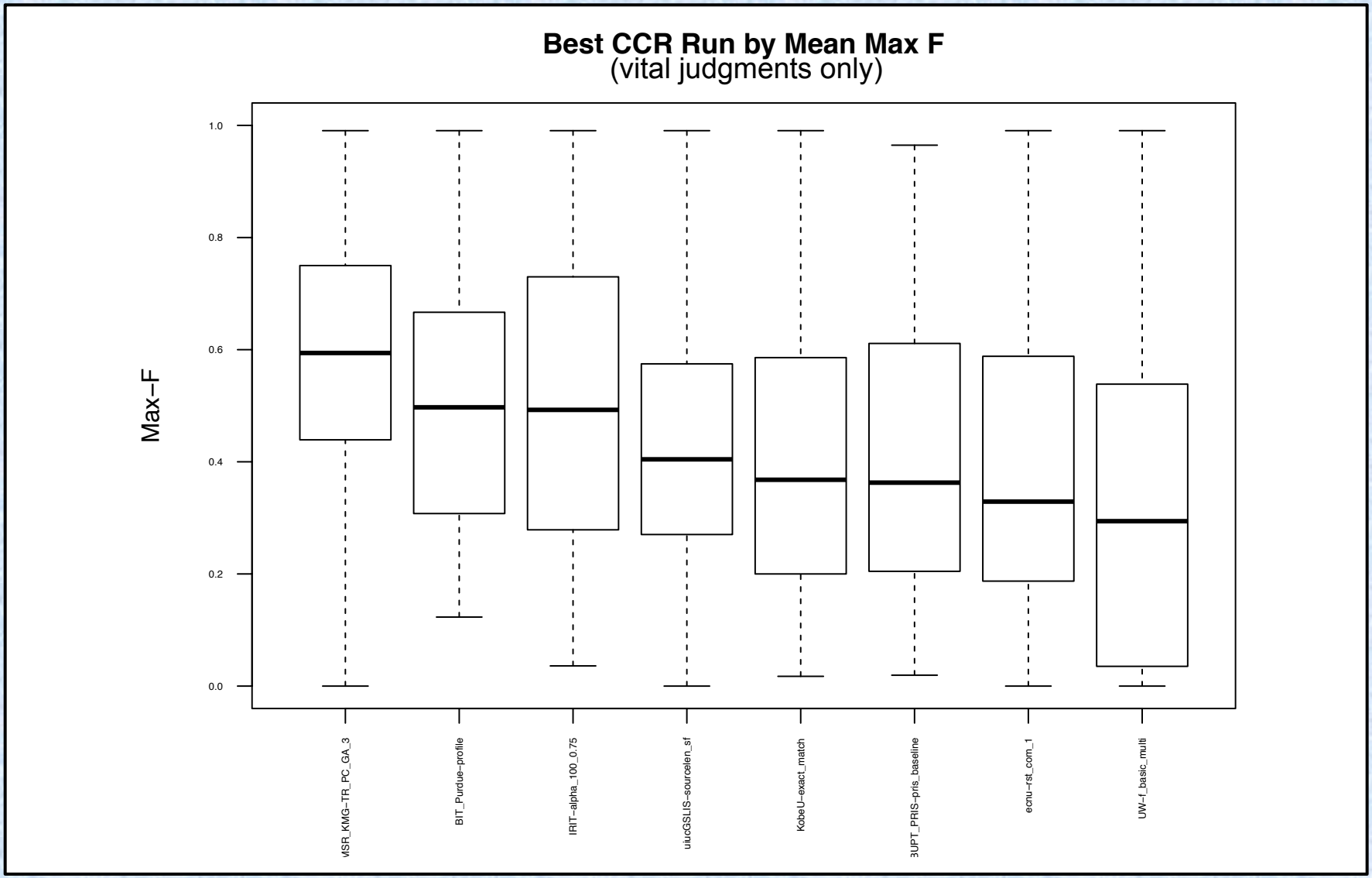
Knowledge-Base Acceleration

- Entity-centric filtering
  - assist humans with KB curation task
  - entity = object with strongly typed attributes
  - track changes of pre-specified attributes
- Tasks
  - Cumulative Citation Recommendation (CCR)
    - return documents that report a fact that would change the target's existing profile
  - Streaming Slot Filling (SSF)
    - find a fill for a slot for the target entity

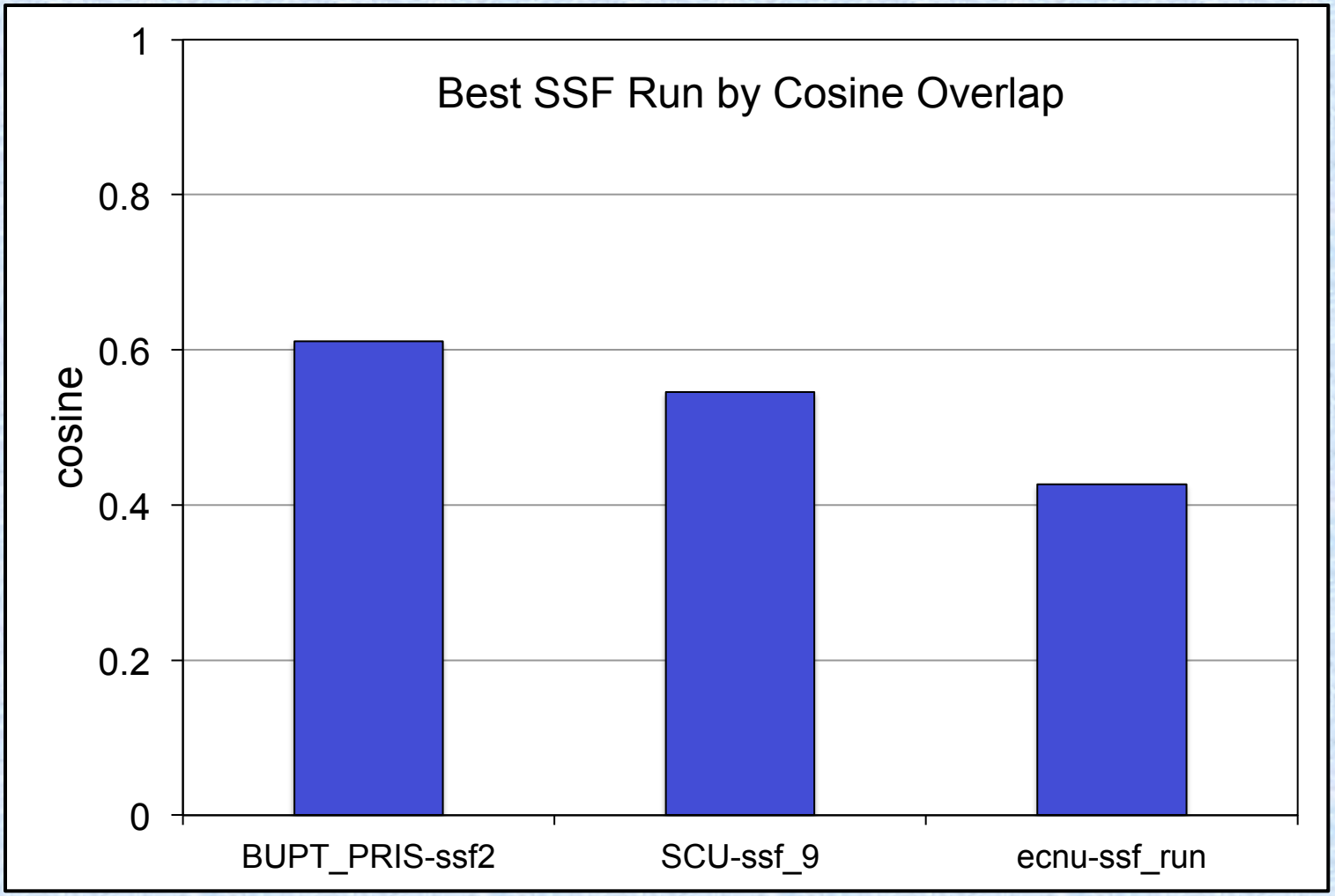
KBA

- 109 target entities
  - 86 people, 16 organizations, 7 facilities
  - all based between Seattle, WA & Vancouver, BC
  - most did not have Wikipedia entry
- Systems return doc & confidence-score
  - confidence scores define retrieved sets for eval
  - corpus is 2014 KBA Stream Corpus; first 20% of relevant docs is training corpus, rest test corpus
- Evaluation
  - F, scaled utility for CCR
  - vector overlap measures for SSF
    - word-vectors built from system and human slot fill sets

KBA CCR Task


Best CCR Run by Mean Max F
(vital judgments only)
Max−F
1.0
0.8
0.6
0.4
0.2
0.0
vSR_KMG−TR_PC_GA_3
BIT_Purdue−profile
IRIT−alpha_100_0.75
uiucGSLIS−sourcelen_sf
KobeU−exact_match
3UPT_PRIS−pris_baseline
ecnu−rst_com_1
UW−f_basic_multi

KBA SSF Task


Best SSF Run by Cosine Overlap
cosine
1
0.8
0.6
0.4
0.2
0
BUPT_PRIS-ssf2
SCU-ssf_9
ecnu-ssf_run

Microblog

- Goal
  - examine search tasks for information seeking behaviors in microblogging environments
- 2014 Tasks
  - ad hoc task
    - temporally anchored ad hoc search for arbitrary topic of interest, X:
    - "at time T, give me most relevant tweets about X"
  - Tweet Timeline Generation (TTG) task
    - categorization task
    - return one Tweet from each (imputed) category

Microblog

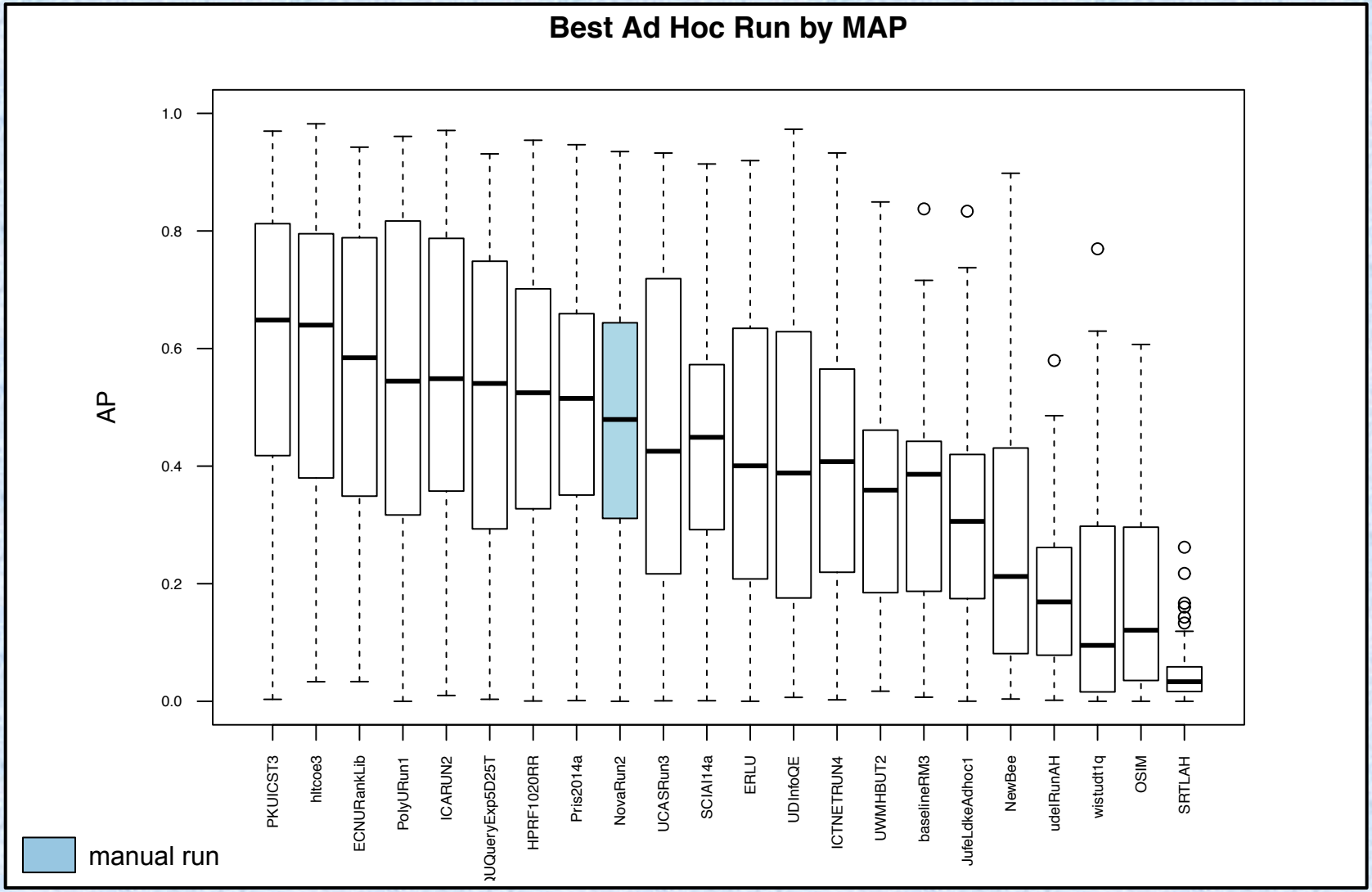
- Search as Service model
  - motivation:
    - larger document set than feasible with distribute-collection model while complying with Twitter TOS
  - implementation:
    - centrally gather sample of tweets from Feb 1-Mar 31, 2013
    - provide access to set through Lucene API
    - API accepts query string and date, returns ranked list of matching tweets (plus metadata) up to specified date

Microblog Ad Hoc

- Real-time search:
  - query issued at a particular time and topic is about something happening at that time
- 55 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 rank tweets issued prior to trigger Tweet's time

Query: muscle pain from statins
querytime: Sat Mar 23 18:21:09 EDT 2013
querytweettime: 315589058900418560

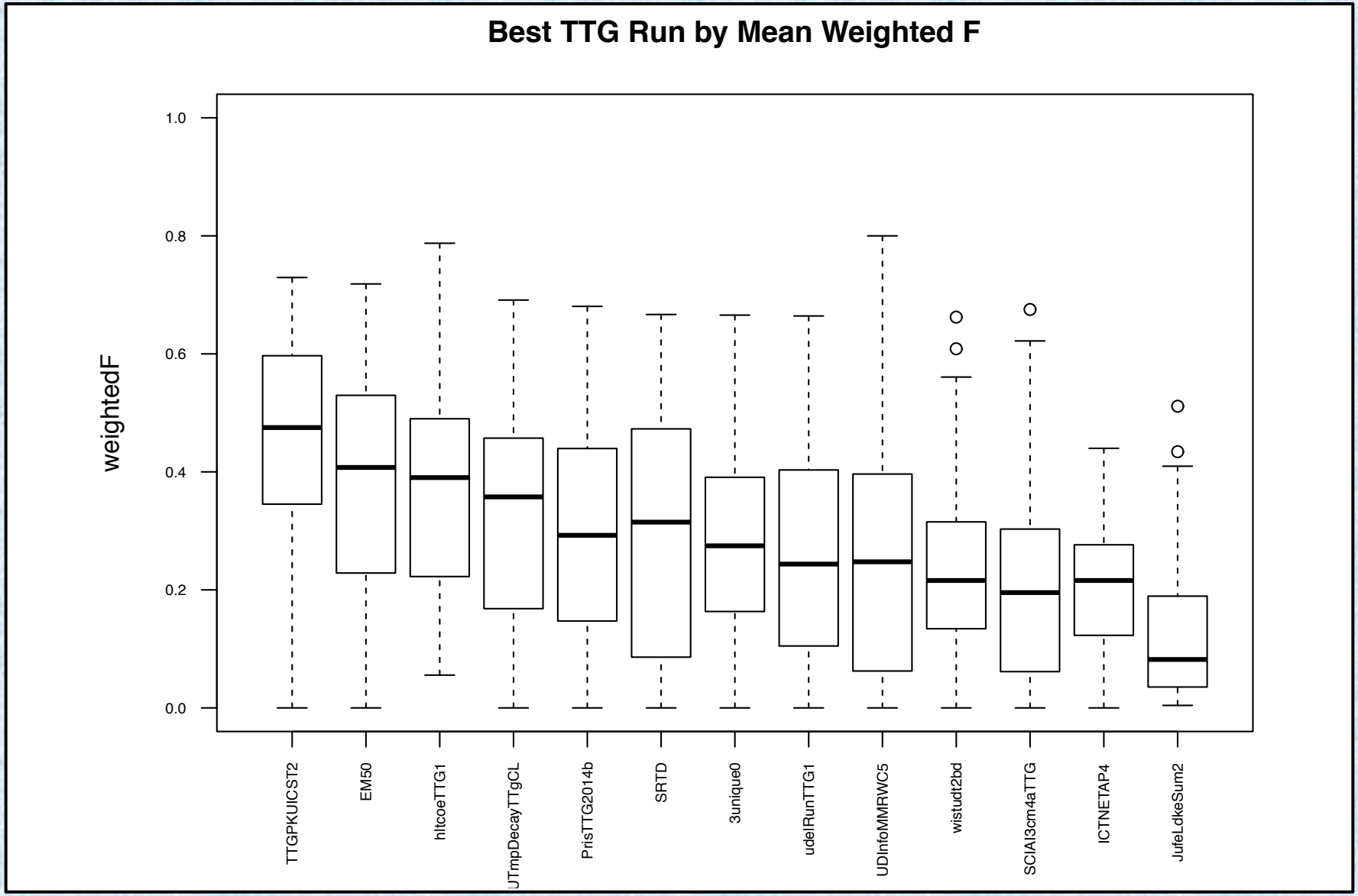
Microblog Ad Hoc Task



Microblog TTG Task

- For many topics, set of relevant tweets is highly redundant
  - retweets are not relevant by fiat, but many distinct tweets contain essentially the same info
  - TTG task focuses on retrieving minimal set of tweets that covers all relevant information
- Implementation:
  - track organizers cluster set of relevant tweets into equivalence classes
  - systems return a set of tweets that ideally contains exactly one tweet per class
  - count one tweet per class as relevant in retrieved set, and compute recall, precision, F

Microblog TTG Task



TREC 2015

- Tracks
  - CDS, Contextual Suggestion, Microblog, Temporal Summarization tracks continuing
  - new tracks covering: Dynamic Domains, Live Q&A, inducing Tasks, Total Recall
- TREC 2015 track planning sessions
  - 1.5 hours per track tomorrow (four-way parallel)
  - track coordinators attending 2014
  - you can help shape task(s); make your opinions known


EVERYTIME
YOU USE
THIS FONT
A DESIGNER
LOSES THEIR
WINGS.