Overview of TREC 2016

Ellen Voorhees

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

TREC 2016 Track Coordinators

Clinical Decision Support: Kirk Roberts, Dina Demner-Fushman, Bill Hersh, Ellen Voorhees

Contextual Suggestion: Seyyed Hadi Hashemi, Jaap Kamps, Julia Kiseleva, Charlie Clarke

Dynamic Domain: Grace Hui Yang, Ian Soboroff

Live QA: David Carmel, Dan Pelleg, Yuval Pinter, Eugene Agichtein, Donna Harman

OpenSearch: Krisztian Balog, Anne Schuth

Real-time Summarization: Jimmy Lin, Richard McCreadie, Adam Roegiest, Fernando Diaz

Tasks: Manish Verma, Evangelos Kanoulas, Emine Yilmaz, Rishabh Mehrotra, Ben Carterette, Nick Craswell, Peter Bailey

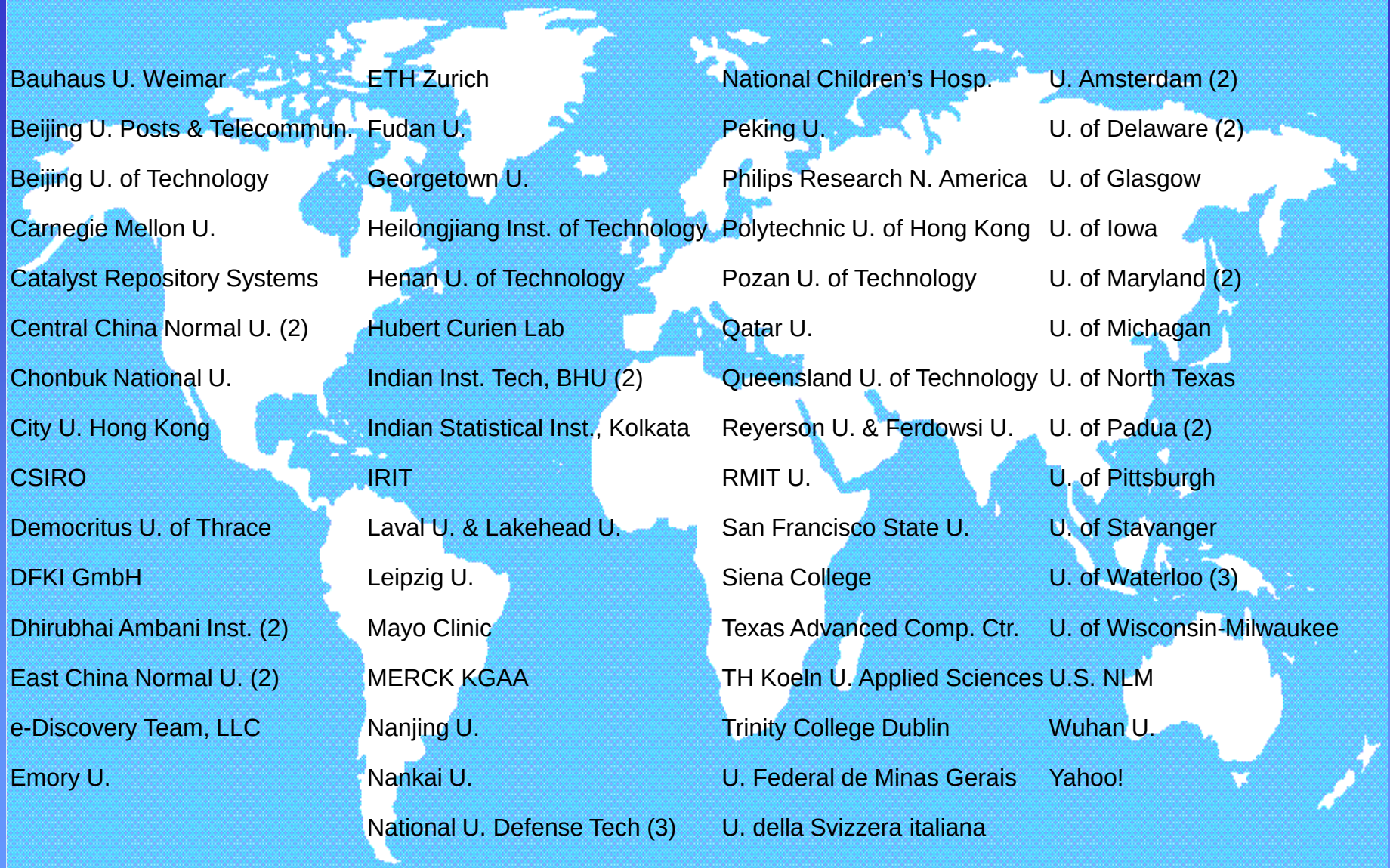
Total Recall:, Gord Cormack, Maura Grossman, Adam Roegiest Charlie Clarke

TREC 2016 Program Committee

Ellen Voorhees, chair

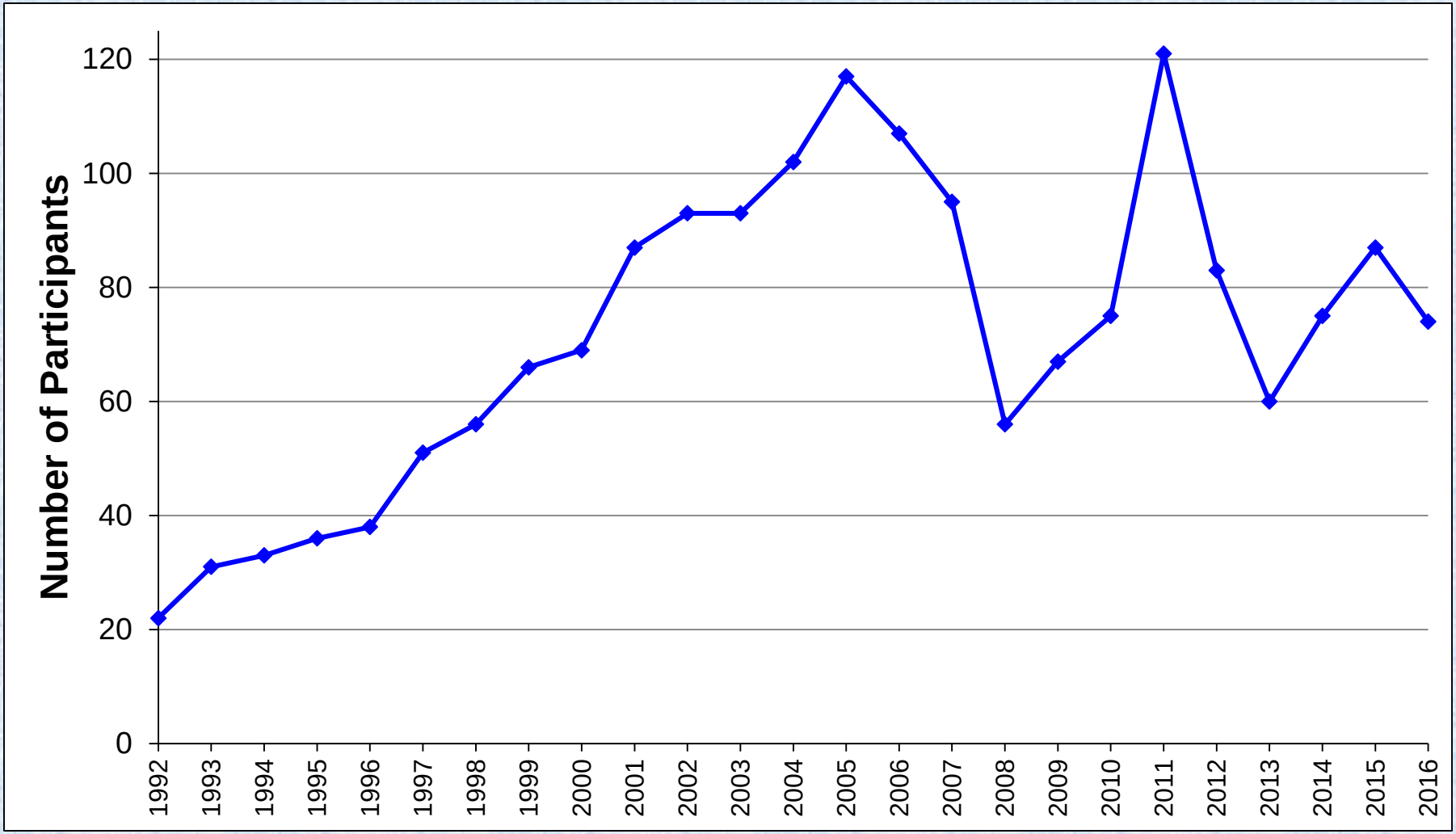
James Allan
Ben Carterette
Gord Cormack
Sue Dumais
Donna Harman
Diane Kelly
David Lewis
Paul McNamee
Doug Oard
John Prager
Ian Soboroff
Arjen de Vries

74 TREC 2016 Participants



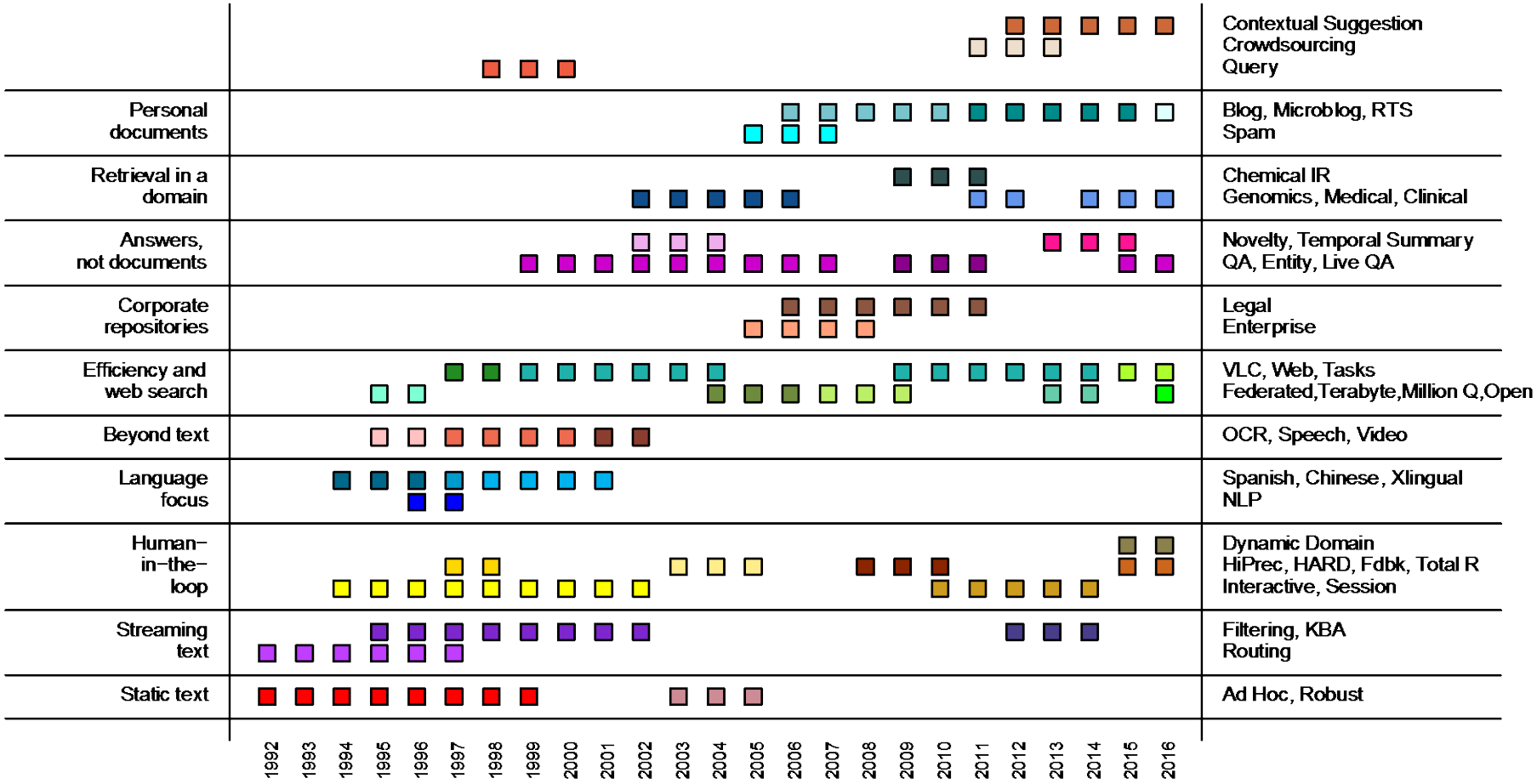
Bauhaus U. Weimar
Beijing U. Posts & Telecommun.
Beijing U. of Technology
Carnegie Mellon U.
Catalyst Repository Systems
Central China Normal U. (2)
Chonbuk National U.
City U. Hong Kong
CSIRO
Democritus U. of Thrace
DFKI GmbH
Dhirubhai Ambani Inst. (2)
East China Normal U. (2)
e-Discovery Team, LLC
Emory U.
ETH Zurich
Fudan U.
Georgetown U.
Heilongjiang Inst. of Technology
Henan U. of Technology
Hubert Curien Lab
Indian Inst. Tech, BHU (2)
Indian Statistical Inst., Kolkata
IRIT
Laval U. & Lakehead U.
Leipzig U.
Mayo Clinic
MERCK KGAA
Nanjing U.
Nankai U.
National U. Defense Tech (3)
National Children's Hosp.
Peking U.
Philips Research N. America
Polytechnic U. of Hong Kong
Pozan U. of Technology
Qatar U.
Queensland U. of Technology
Reyerson U. & Ferdowsi U.
RMIT U.
San Francisco State U.
Siena College
Texas Advanced Comp. Ctr.
TH Koeln U. Applied Sciences
Trinity College Dublin
U. Federal de Minas Gerais
U. della Svizzera italiana
U. Amsterdam (2)
U. of Delaware (2)
U. of Glasgow
U. of Iowa
U. of Maryland (2)
U. of Michagan
U. of North Texas
U. of Padua (2)
U. of Pittsburgh
U. of Stavanger
U. of Waterloo (3)
U. of Wisconsin-Milwaukee
U.S. NLM
Wuhan U.
Yahoo!

Number of Participants in TREC



A big thank you to our assessors

TREC Tracks


Contextual Suggestion
Crowdsourcing
Query
Personal documents
Blog, Microblog, RTS
Spam
Retrieval in a domain
Chemical IR
Genomics, Medical, Clinical
Answers, not documents
Novelty, Temporal Summary
QA, Entity, Live QA
Corporate repositories
Legal
Enterprise
Efficiency and web search
VLC, Web, Tasks
Federated,Terabyte,Million Q,Open
Beyond text
OCR, Speech, Video
Language focus
Spanish, Chinese, Xlingual
NLP
Human–in–the–loop
Dynamic Domain
HiPrec, HARD, Fdbk, Total R
Interactive, Session
Streaming 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
2015
2016

Basics

- Generic tasks
  - ad hoc: known collection, unpredictable queries, response is a ranked list
  - filtering: known queries, document stream, response is a document set,
  - question answering: unpredictable questions, response is an actual answer not a document
- Measures
  - recall, precision are fundamental components
  - ranked list measures: nDCG@X, IA-ERR, CubeTest
  - filtering measures: F, expected gain, latency

TREC 2016

- A year of consolidation
  - sophomore year for 4/8 tracks
  - OpenSearch track new
    - a "meta-track" focusing on new evaluation paradigm
  - continued high [engineering] barrier for participation
    - writing to track APIs to access track resources
    - hard time constraints for responses

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
- Implementation

  Given a case narrative, return biomedical articles that can be used to accomplish one of three generic clinical tasks:
  - What is the diagnosis? or What is the best treatment? or What test should be run?

CDS Track Task

- Documents:
  - new snapshot of the open-access subset of PubMed Central
  - contains ~ 1.25 million full-text articles
- 30 topics
  - new for 2016, based on nursing admission notes from existing medical record (MIMIC-II)
  - physicians created corresponding "description" and "summary" versions, as well as designated target clinical task
  - 10 topics for each clinical task type

CDS Track

- Judgments
  - judgment sets created using inferred measure sampling (2 strata; ranks 1-15; 20% of 16-100); main measure infNDCG
  - judgments made by physicians coordinated by OHSU
  - up to 5 runs per participant
  - all runs contributed to same set of pools

CDS Track Sample Topic

<topic number="20" type="**test**">

Note:

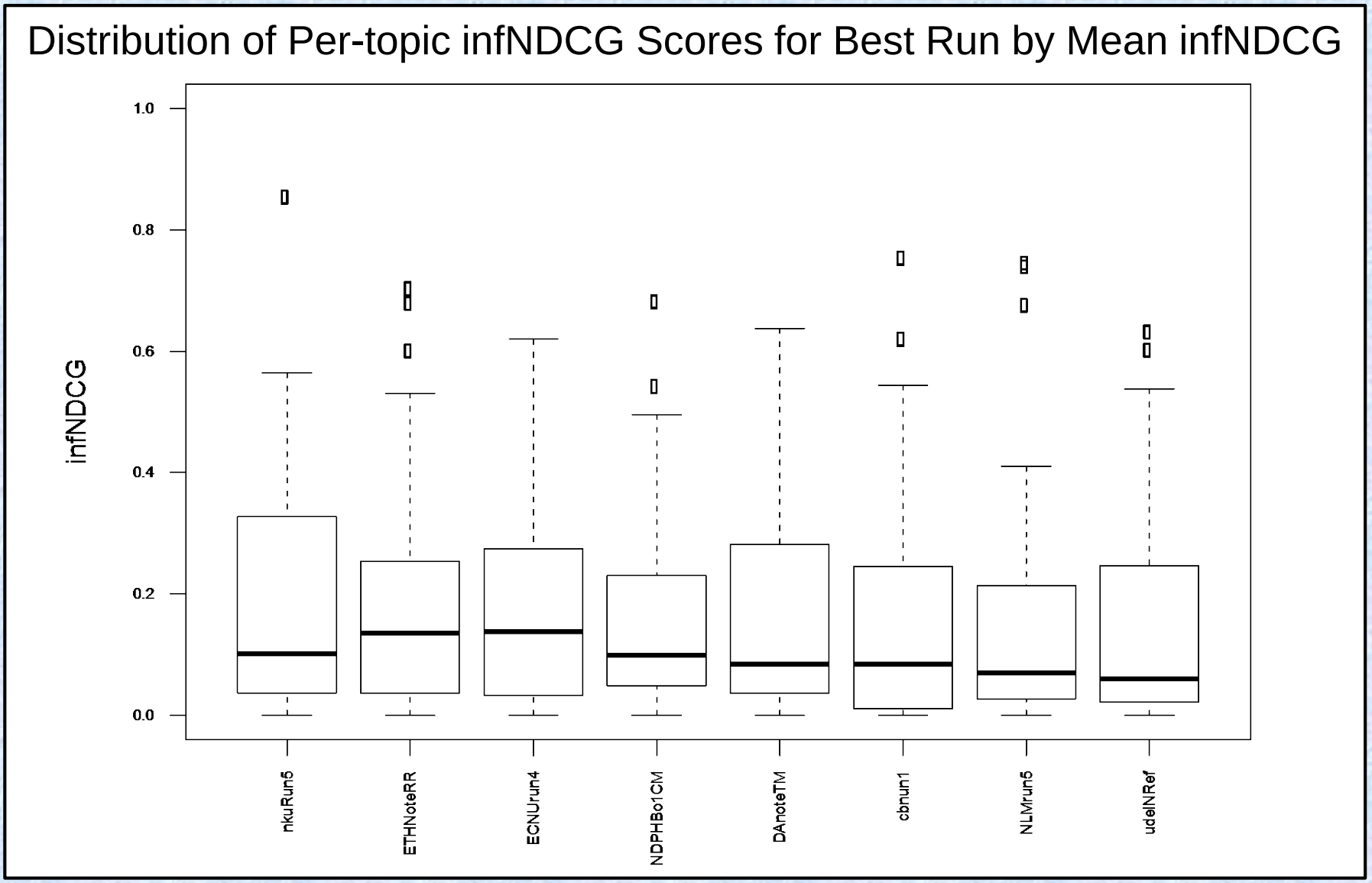
This is a 87 year old female NH resident with a history of chronic atrial fibrillation, hypertension and hypothyroidism who presents to the [**Hospital Unit Name 10**]. She had been in her usual state of health until 5 days ago when she suddenly began to have abdominal pain. Her abdominal pain was initially intermittent lasting for a few hours at at time. No clear correlation with food. Yesterday, she noticed that her pain was much more severe, [**3301-9-5**] in severity and more localized to the right. This was accompanied by nausea and vomitting. She vomitted twice, with clear liquid emesis and was sent to [**Hospital3 **]. At [**Hospital1 **], she was noted to have elevated amylase/lipase to 538 and 516 with elevated bili to 4.1 and AST/ALT to 198/115 and was given ciprofloxacin, flagyl and 500cc NS and was transferred to the [**Hospital1 1**] emergency department.

At [**Hospital1 1**] EDVS 97.9 HR 83 157/92 RR 18 97% RA.
Elderly F, oriented X 2, NAD, flat jvp, CTA decreased b/b, s1 s2
[**Last Name (un) **], decreased BS, + t at ruq, no edema

Description: A 87 year old female NH resident with a history of chronic atrial fibrillation, hypertension, and hypothyroidism who presents wit abdominal pain. She had been in her usual state of health until 5 days ago when she suddenly began to have abdominal pain. Her abdominal pain was initially intermittent lasting for a few hours at a time. No cleat correlation with food. Yesterday, she noticed that her pain was much more sever and much more localized to the right. This was accompanied by nausea and vomitting. She vomitted twice, with clear liquid emesis and was sent to a hospital. At the hospital, she was noted to have elevated amylase/lipase to 538 and 516 with elevated bili to 4.1 and AST/ ALT to 198/115 and was given ciprofloxacin, flagyl and 500cc NS and was transferred to the emergency department. At the emergency department her vital signs were TM 97.9 HR 83 BP 157/92 RR 18 sat 97% RA.

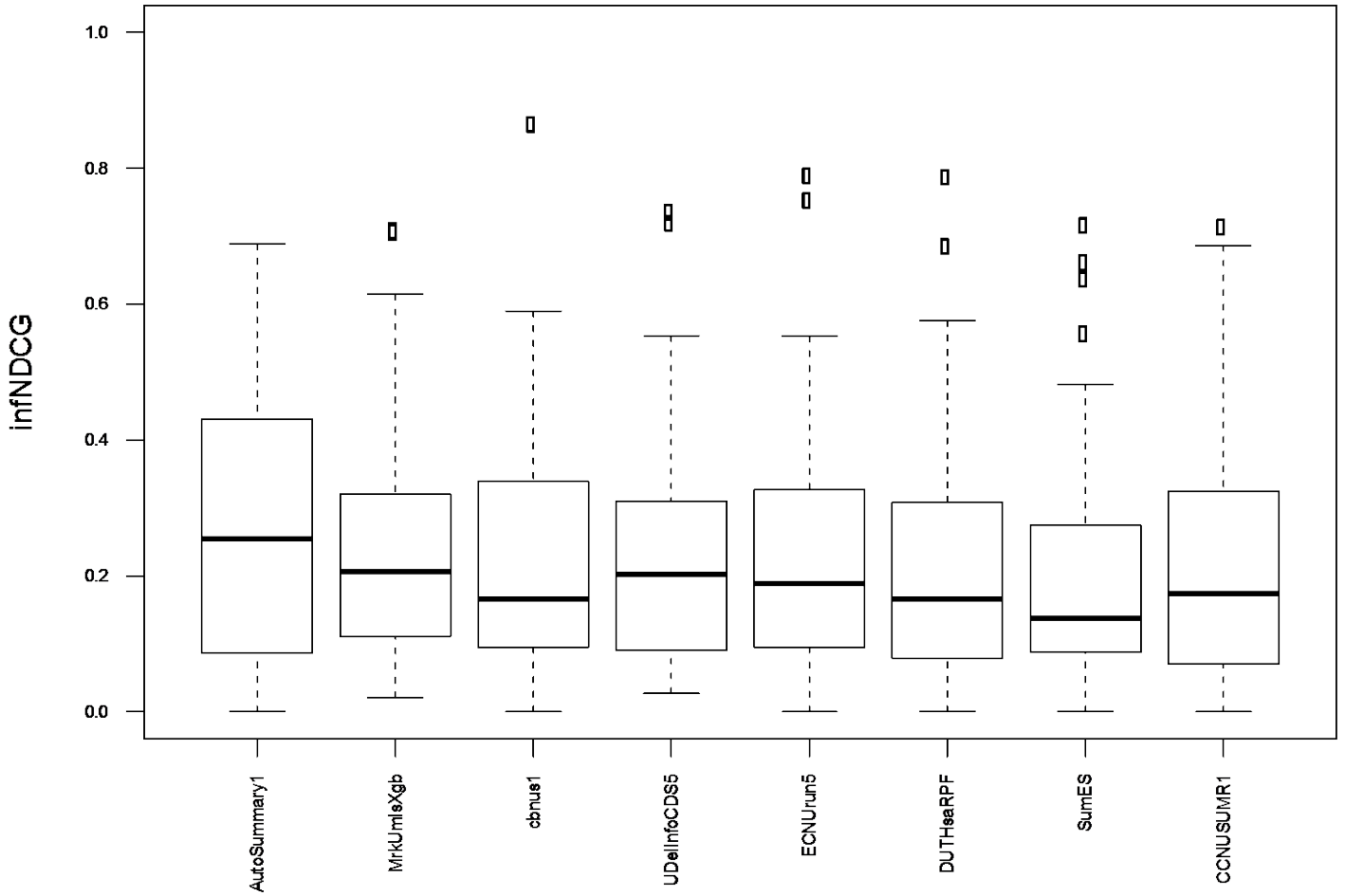
Summary: A 87 yo female reports several days abdominal pain, worse yesterday, severe and more localized to the right, accompanied by nausea and vomitting. Labs show elevated bilirubin, transaminitis, amylase and lipase.

CDS Results: Automatic, Note



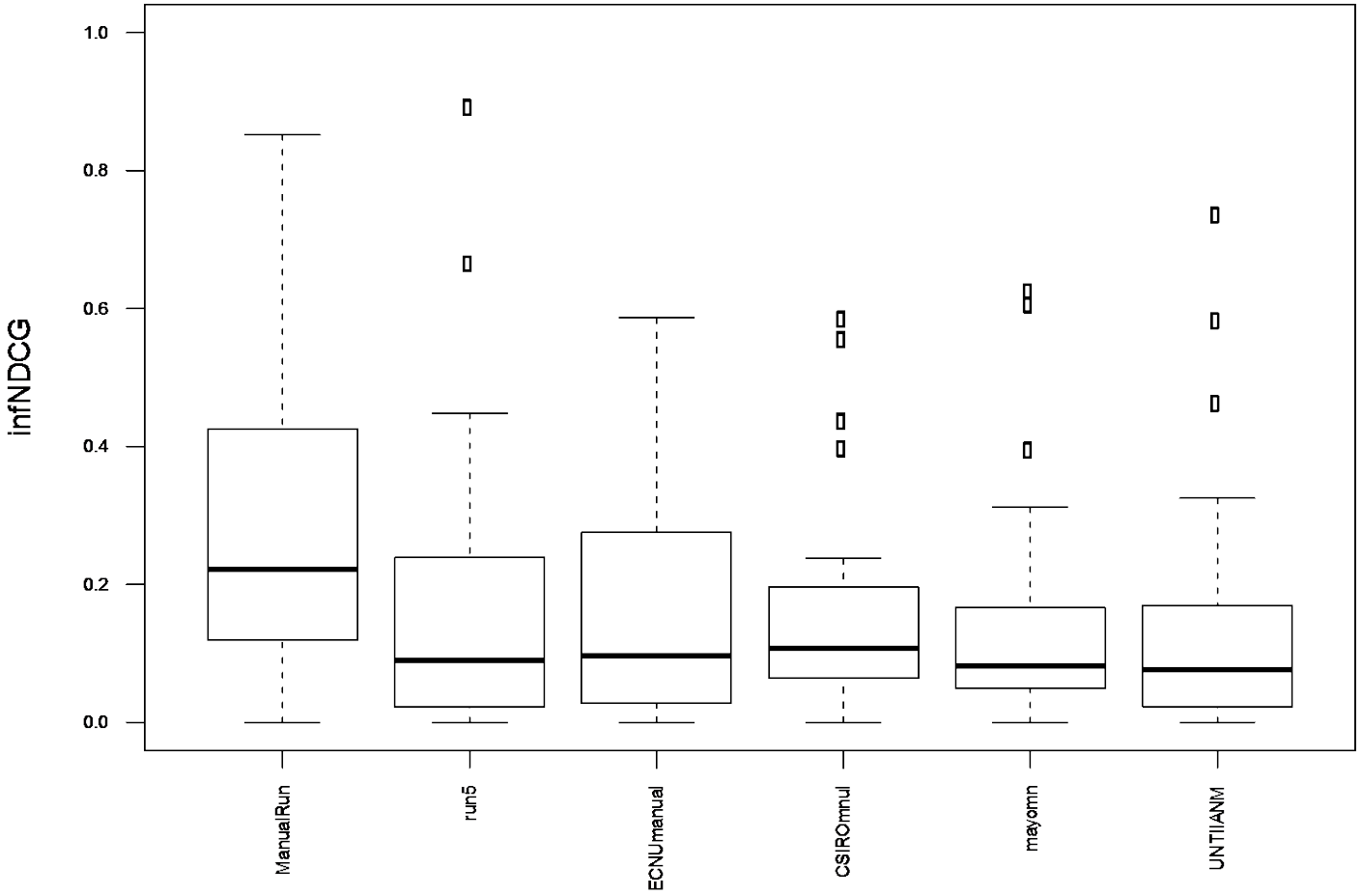
CDS Results: Auto, Summary

Distribution of Per-topic infNDCG Scores for Best Run by Mean infNDCG



CDS Results: Manual

Distribution of Per-topic infNDCG Scores for Best Run by Mean infNDCG



Dynamic Domain

- Goal
  - evaluate methods that support the entire information-seeking process for exploratory search in complex domains
  - systems must support dynamic nature of search in cost effective manner
- Implementation
  - interaction jig referred to as 'Simulated User'
  - participants submit 5-doc packets to Simulated User and get judgments for individual facets of the topic
    - each packet-submission with feedback is one iteration
  - system decides to stop when it thinks sufficient info for all facets has been retrieved

Dynamic Domain

- Domains
  - two domains with a total of 53 topics
    - Ebola: ~680,000 webpages/pdfs/tweets about Ebola outbreak in Africa in 2014-2015
    - Polar: ~1.7 million webpages/data files/images/code related to the polar sciences
- Topics
    - developed by assessors (Ebola) or USC (polar)
    - NIST assessors made judgments for docs found in multiple rounds of searching prior to topic release
    - assessors also created gold-standard set of facets for each topic based on these searches

Dynamic Domain Sample Topics

Polar

Topic: polar oceans freshwater sensitivity

How sensitive are the polar oceans to changes in freshwater input?

Subtopic 1: surface freshwater forcing
Subtopic 2: Arctic Freshwater Initiative
Subtopic 3: Freshwater Budget of the Canadian Archipelago
Subtopic 4: Freshwater Fluxes in the East Greenland Current
Subtopic 5: terrestrial and freshwater ecosystems

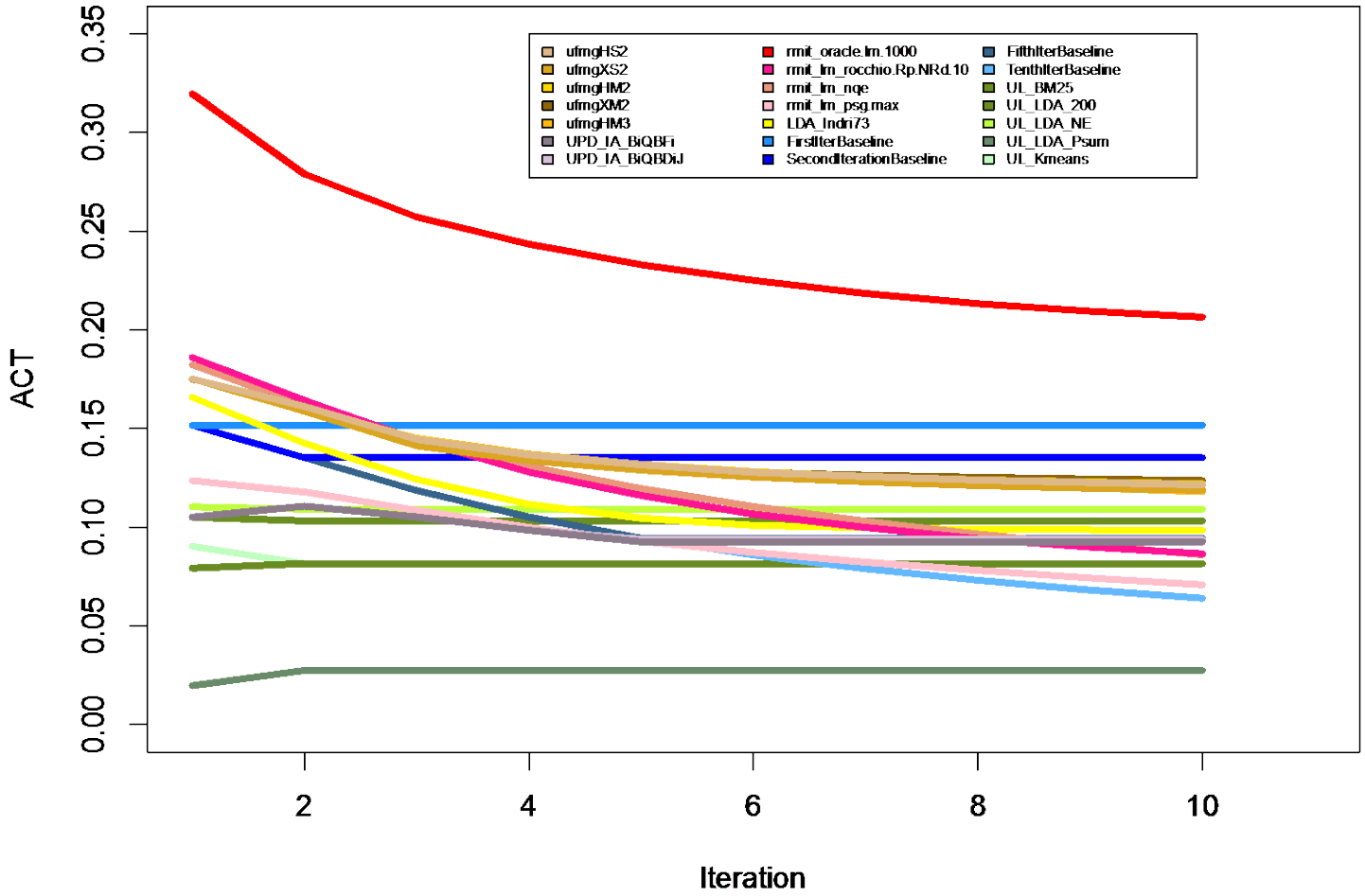
Ebola

Topic: Ebola Conspiracy Theories

Identify the conspiracies circulating about Ebola.

Subtopic 1: Efforts to Counter
Subtopic 2: Origins
Subtopic 3: The Claims

Dynamic Domain Results



Total Recall

- Goal
  - evaluate methods for achieving very high recall, including methods that use a human-in-the-loop
  - more emphasis on recall, less on different facets than Dynamic Domain track; both emphasize stopping criteria
- Implementation
  - participant system submits one doc at a time to a software jig; jig both records activity & responds to system with relevance judgment for that doc
  - participant decides when to terminate search; entire set of documents submitted to jig counts as retrieved set

Total Recall

At Home Collection

Jeb Bush email: 34 (new) topics against the email of Florida governor Jeb Bush

Sandbox Collections

Gov email: six topics against email of Rod Blagojevich and Patrick Quinn admins

Twitter: four topics against a collection of 800,000 tweets

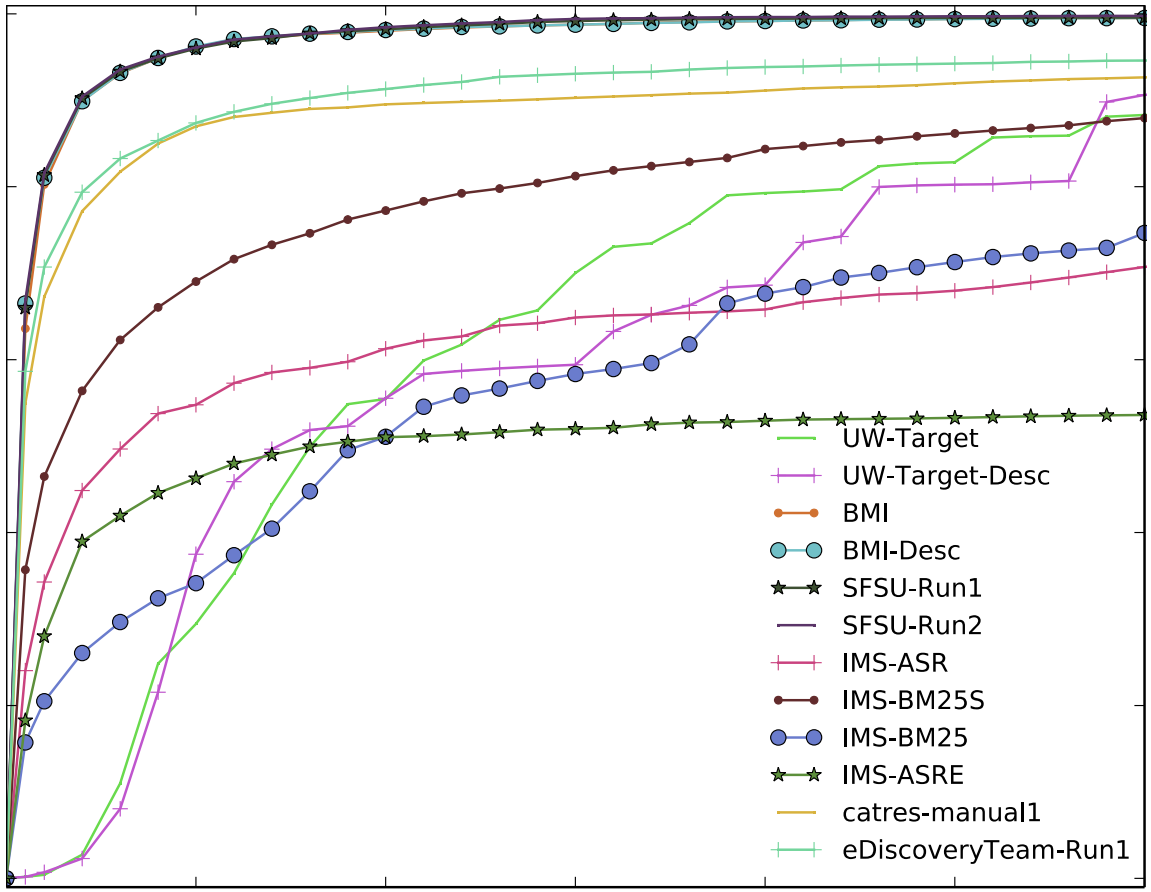
- Tasks
  - "**at home**": systems connect to jig over Internet; participant's machine contains document set and search runs there.
  - **sandbox**: participant's system sent as virtual machine that runs on isolated machine along with the jig. Participant never sees any documents, but gets counts of relevants returned as function of number documents submitted. Automatic only.

Total Recall

- At Home judgments:
  - NIST assessors judged multiple rounds of documents per topic
  - (small) sample of documents independently judged by multiple assessors
  - 3-way: not relevant, relevant, important
  - primary assessor also grouped relevant/important documents into clusters representing main subtopics (aspects) of topic
- Enables evaluation contrasts:
  - all relevant vs. important; assessor differences; coverage

Total Recall At Home Results

Average Gain Curve: All relevant primary assessor



OpenSearch

- "Live Labs" comes to TREC
  - provide access to real users doing their real searches at the actual time they search
  - at scale
- For TREC participants, an ad hoc search re-ranking task

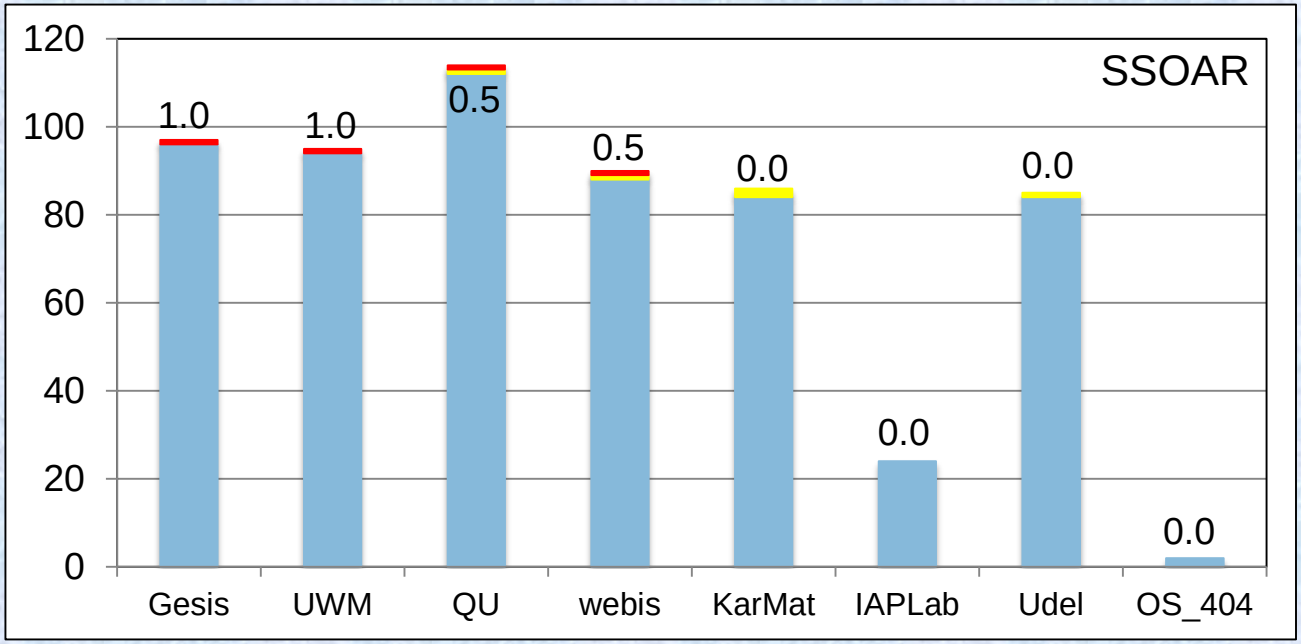
OpenSearch Protocol

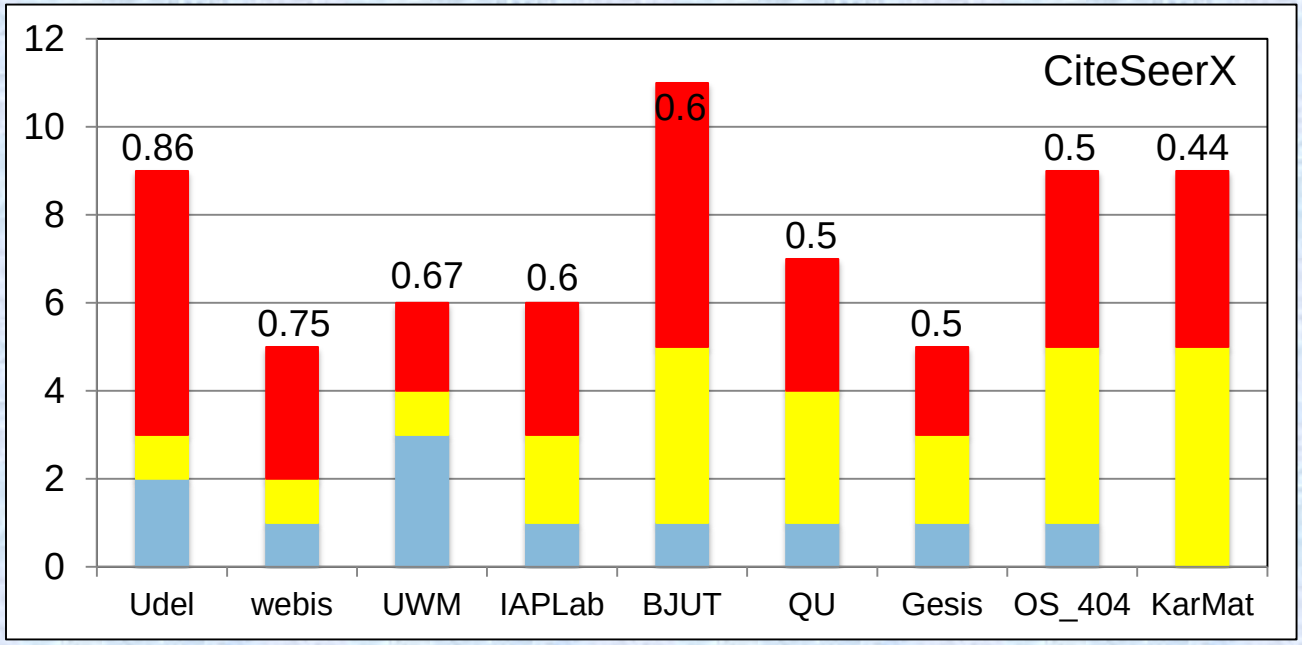
- Sites provide frequent queries and valid document sets
- Participants re-rank document sets for each query and upload new rankings
- Once a query in set is issued,
  - a participant is randomly selected and that participant's corresponding ranking is interleaved with site's native ranking
  - all user's interactions with ranked list recorded
  - participant's ranking declared to be a win, loss or tie with respect to native ranking

OpenSearch Results

Round 2

Outcome





Contextual Suggestion

- "Entertain Me" app: suggest activities based on user's prior history and target location
- Fifth year of track
  - this year consolidates work in 2015 focusing on creating a reusable test collection for task
  - suggestions required to come from track-created repository of activities
  - suggestions in profiles might be tagged features the profile owner finds attractive

Contextual Suggestion

- Terminology:
  - a profile represents the user
    - profile consists of a set of previously rated activities and possibly some demographic info
  - a system returns [a ranked list of] suggestions in response to a request
    - a request contains at least a profile and target location and possibly some other data (e.g., time)
    - a suggestion is an activity from the repository that is located in the target area

Contextual Suggestion Sample Request

location: Cape Coral, FL
group: Family **season**: Summer
trip_type: Holiday **duration**: Weekend trip
person:
gender: Male **age**: 23
preferences:
doc: 00674898-160 **rating**: 3
tags: Romantic, Seafood, Family Friendly
doc: 00247656-160 **rating**: 2
tags: Bar-hopping
doc: 00085961-160 **rating**: 3
tags: Gourmet Food
doc: 00086637-160 **rating**: 4
tags: Family Friendly, Local Food, Entertainment
doc: 00086298-160 **rating**: 0
doc: 00087389-160 **rating**: 3
tags: Shopping for Shoes, Family Friendly, Luxury Brand Shopping
doc: 00405444-152 **rating**: 3
tags: Art, Art Galleries, Family Friendly, Fine Art Museums

Contextual Suggestion

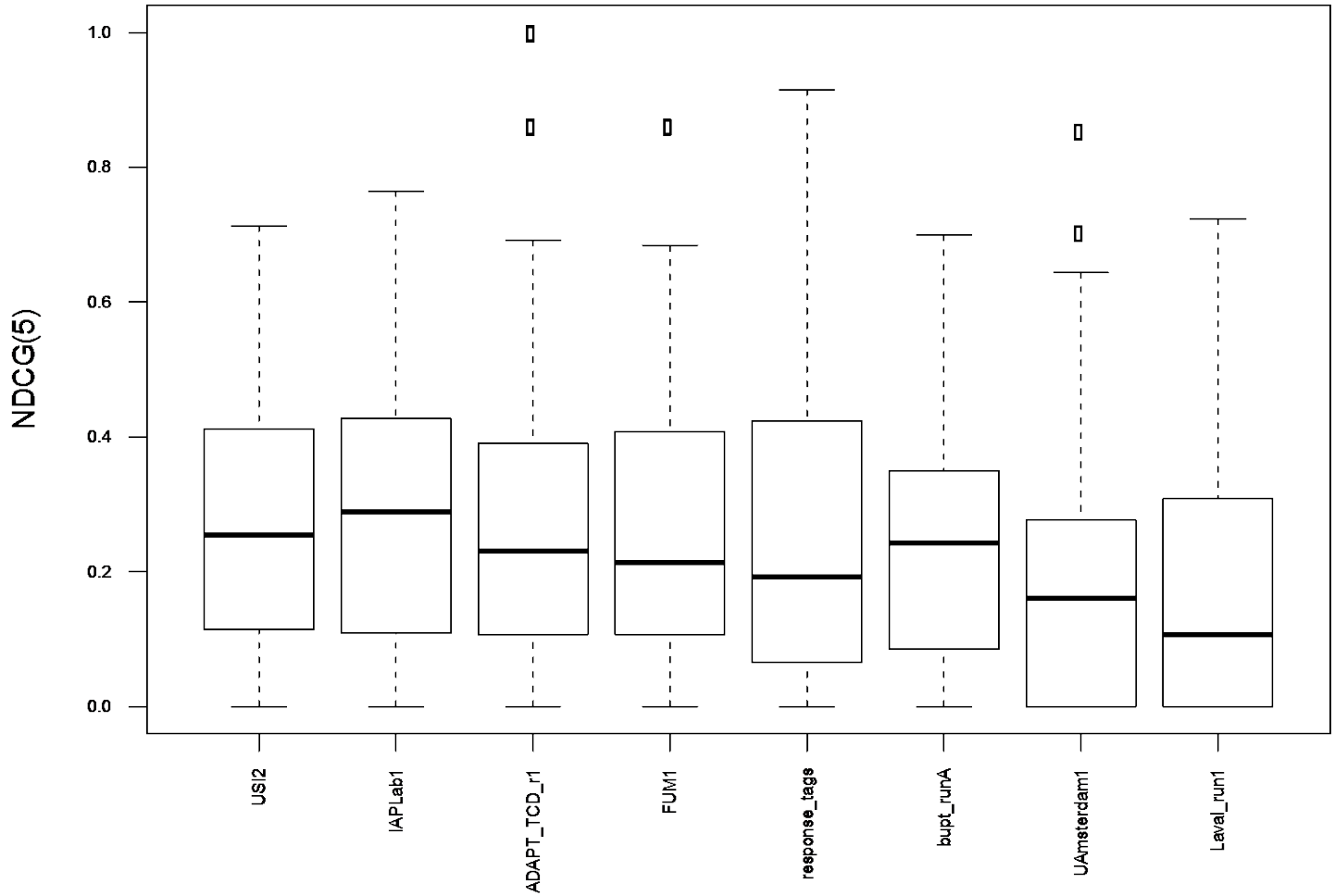
- Phase 1 task
  - crowd-sourced assessors rated attractions from a set of seed locations; these ratings formed initial profiles
  - participants returned suggestions for the cross product of profiles and set of different locations
  - suggestions from Phase 1 participants pooled and sent back to requestor for ratings and feature tags
  - evaluated a total of 61 requests in test set
  - ratings on 5-point scale Strongly Uninterested—Strongly Interested;
    - top 2 counted as 'relevant' for binary measures
    - NDCG computed using 3-rating as gain value

Contextual Suggestion

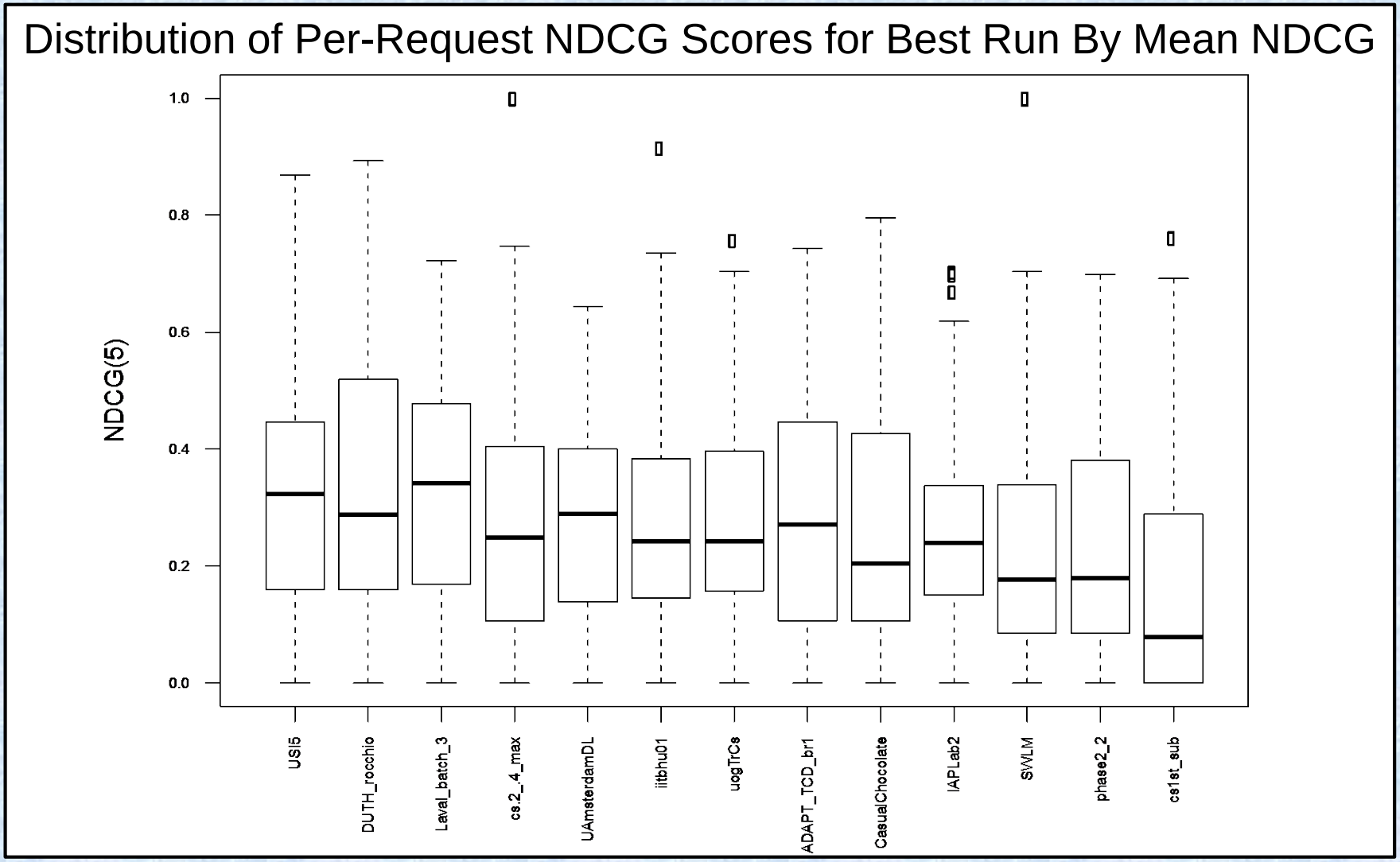
- Phase 2 task
  - essentially, a re-ranking task
  - a request contained the complete set of (unrated) suggestions from all Phase 1 participants for the request; Phase 2 task participants were required to return only suggestions from this set
  - 58 requests in evaluation set

Contextual Suggestion Phase 1 Results

Distribution of Per-Request NDCG Scores for Best Run By Mean NDCG



Contextual Suggestion Phase 2 Results



Tasks Track

- Goal
  - facilitate research on systems that are able to infer the underlying real-world task that motivates a query and then can retrieve documents useful for accomplishing all aspects of that real-world task
- Tasks
  - Task Understanding
    - return key phrases covering breadth of Task
  - Task Completion
    - return documents that are useful for whole Task
  - Web/ad hoc

Tasks Track

- ClueWeb12 document set
- 50 topics in test set
  - track organizers selected topics from logs and created the set of subtasks using their own resources plus participants' submissions
- Aspect-based judgments
  - depth 20 pools for phrases
  - depth 10 pools for documents (completion & ad hoc)
  - documents judged for both relevance and usefulness

Tasks Track Sample Topics

query: fake tan at home

You are trying to find out how to get a fake tan at home.

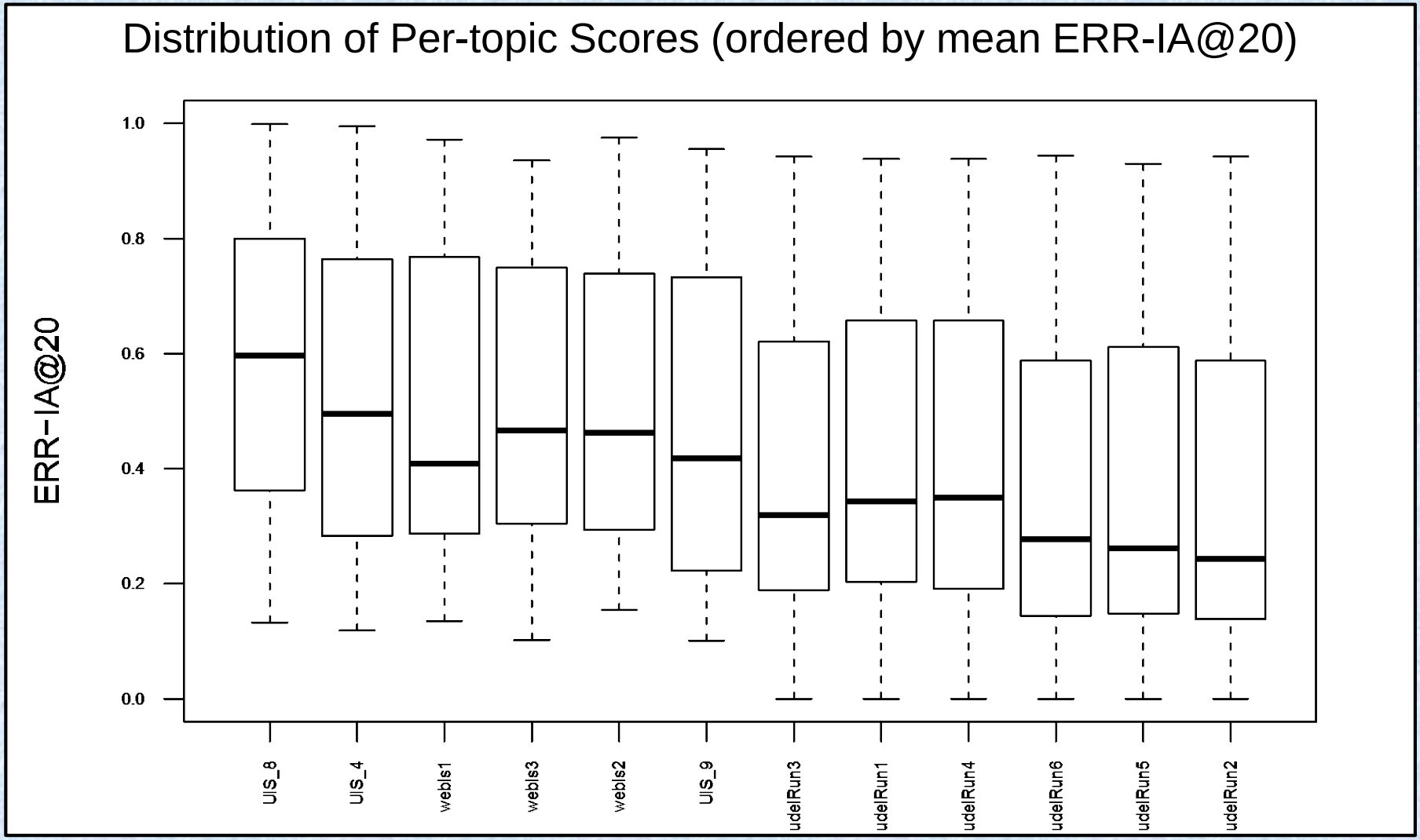
Subtask 1: Places to get a fake tan
Subtask 2: Determine fake tan for skin type
Subtask 3: Cost of getting a fake tan
Subtask 4: Products to get a fake tan
Subtask 5: Buy products to get a fake tan
Subtask 6: Fake tan DIY guide
Subtask 7: Get fake tan of some part of the body
Subtask 8: Precautions for getting a fake tan
Subtask 9: Pictures of fake tans

query: social media for learning

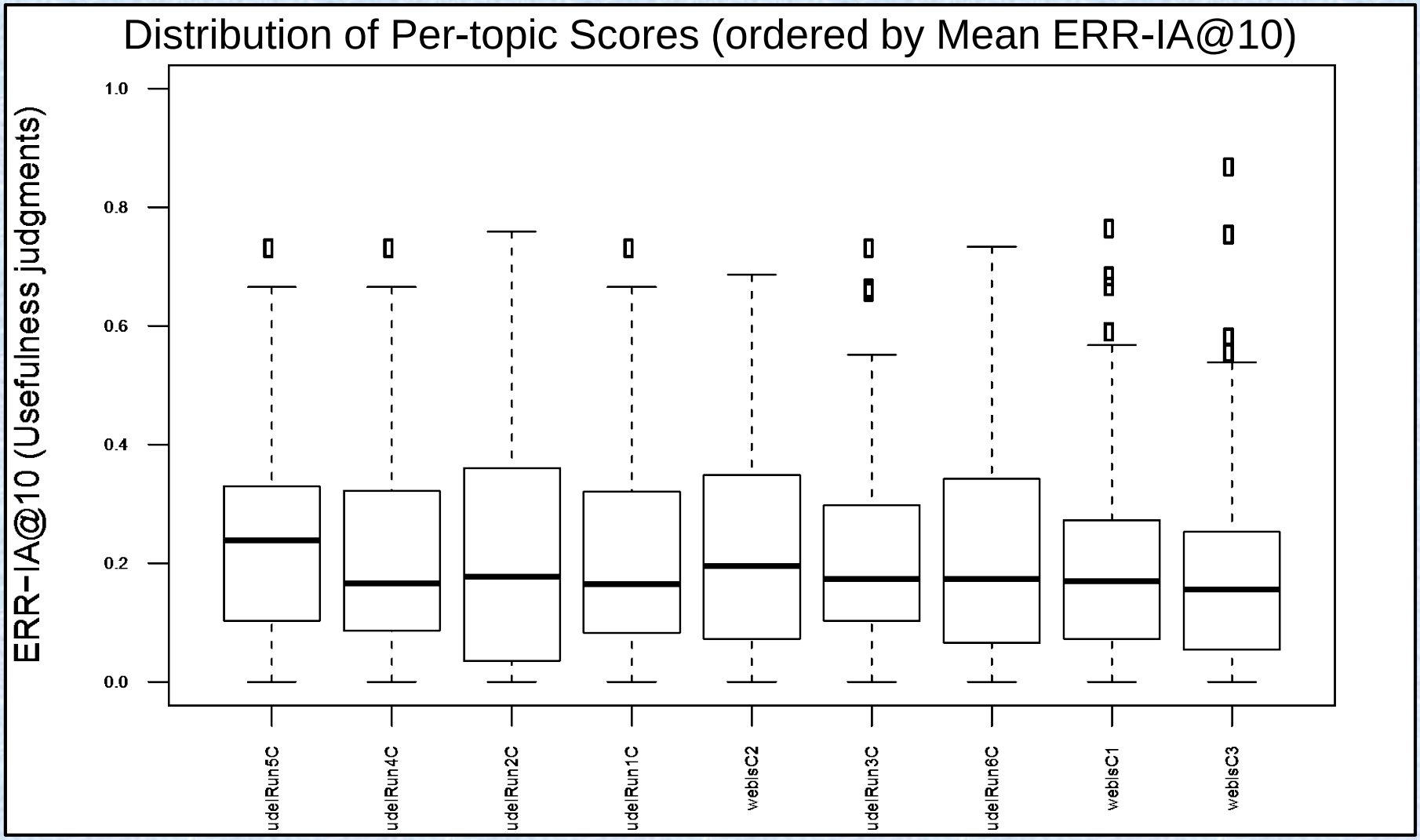
You want to learn different ways to use social media to enhance learning activities at school.

Subtask 1: How to use blogging for learning
Subtask 2: How to use collaborative calendaring
Subtask 3: How to use podcasting
Subtask 4: How to use social media for collaborative mindmapping
Subtask 5: How to use social media for sharing information
Subtask 6: How to use social media for presentation sharing
Subtask 7: How to use social media for collaborative working

Task Understanding Results



Task Completion Results



Live QA

- Goal
  - create systems that can generate answers in real time for real questions asked by real users
- Implementation
  - questions sampled from Yahoo Answers site
  - directed at participants' systems at the rate of about 1 per minute for 24 hours May 31-Jun 1
  - systems required to respond a question with a single [textual] answer in at most 1 minute; answers recorded by track server
  - at end of evaluation period, questions and responses sent to NIST for judgment

Live QA

- Questions
  - drawn from seven top-level Yahoo Answer categories, as self-labeled by asker
  - lightly filtered to remove objectionable material
  - final test set of 1015 questions
- Summarization pilot task:
  - systems return a concise re-statement of the question indicating its main focus

Live QA Sample Questions

Category: Health

Bat ran into me, should I be afraid of rabies?

Could I catch rabies from a bat which ran into me?

Category: Beauty & Style

Is waterproof mascara and eyeliner necessary for traveling in hot and humid areas (Thailand and Singapore)? If I just go with normal ones, will they not smudge or melt off?

Do I need to use waterproof mascara and eyeliner in hot, humid areas?

Category: Pets

One of my dogs left? I have two dogs and one just ran off and we can't find him. It's been three hours and the other one is really depressed she only moves to get water and is breathing heavily. What should I do if my dog does not come back.

What can I do if I can't find my lost dog?

Category: Sports

How does basketball finals work in the NBA? So I like all sorts of sports but my favorite is football (soccer). I was wondering why in basketball have like 4 games just to go to the final. I heard it's gonna be the Cavs vs Warriors?

How are NBA basketball finals structured?

Category: Arts & Humanities

Places to read about early human settlement/migrations in Modern Russia? I was recently listening to lectures by the great courses on big history by David Christain. At some point he brings up the fact that Humans started to settle in northern Ukraine/Russia in the sixth century. I am interested to read more on this settlement/migration. Do you have any suggested scholarly sources/suggestions?

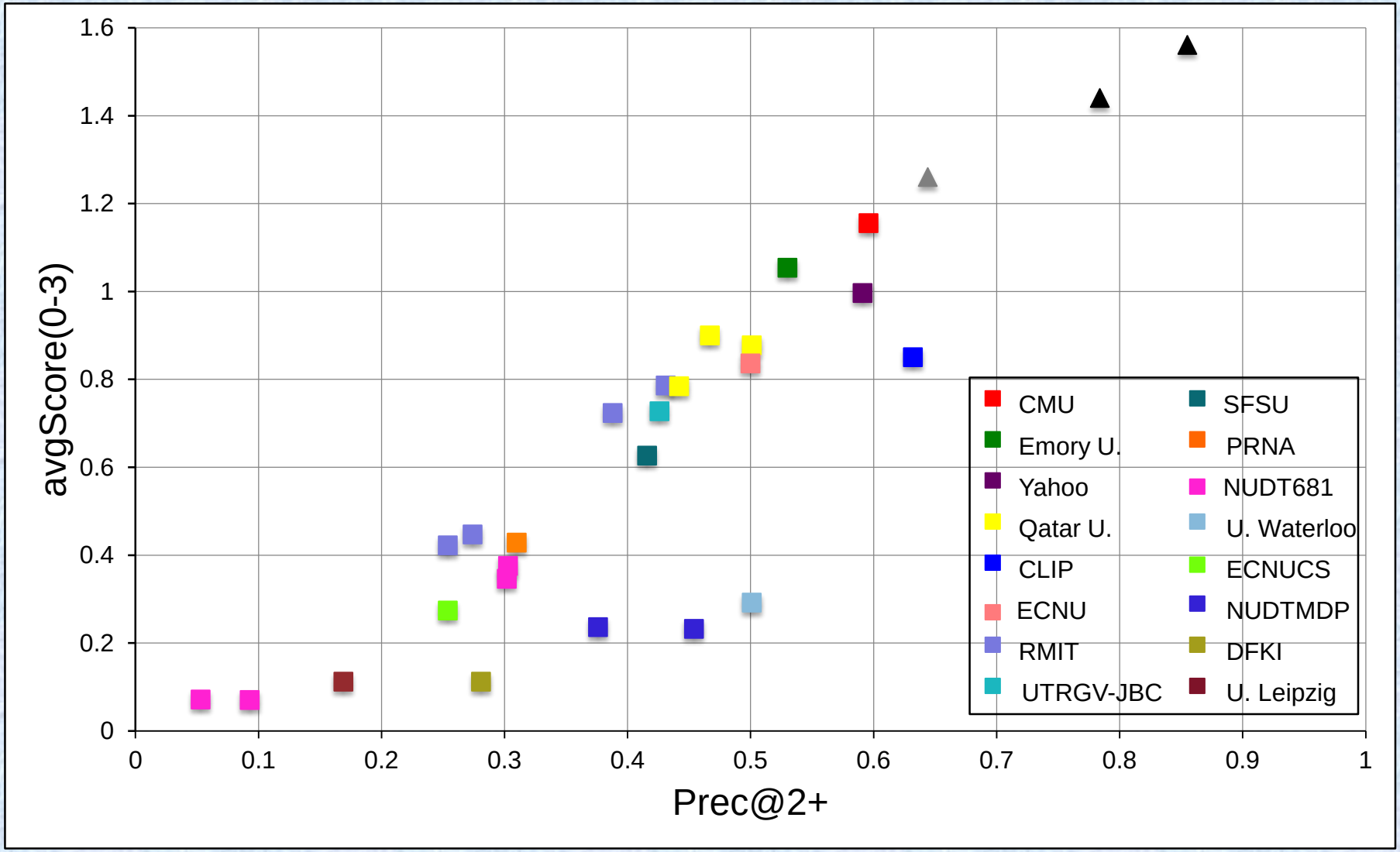
Where can I find out more about early settlements and migration in Russia?

Live QA

- 3 human baselines
  - best human answer on Yahoo! Answers site as voted by asker
  - fastest human answer on site
  - crowd-sourced answer within 1 min. response time
- Scoring
  - NIST assessors rated responses

    -2 Unreadable; 1 Poor; 2 Fair; 3 Good; 4 Excellent
  - runs' score a function of the rating assigned per q
    - avgScore(0-3): conflate all negative responses to 0 & subtract 1 from other ratings; take mean of ratings
    - prec@i+: number of q's with at rating of at least i divided by number of q's system responded to

Live QA Results



Real-time Summarization

- Goal
  - examine techniques for constructing real-time update summaries from social media streams in response to users' information needs
- Mash-up of TREC 2015 Microblog and Temporal Summarization tracks
  - (type of) filtering task over Tweet stream
    - Task A: deliver updates to mobile device
    - Task B: periodic digest of updates

Real-time Summarization

- Participant listens to Twitter public feed for evaluation period (~10 days in August)
- Pushes a tweet to the RTS server when it decides to retrieve a tweet for a profile
  - at most 10 tweets per day per profile
- Server records time of receipt
- A subset of the tweets pushed to crowd-sourced "mobile" assessors
    - assessor may or may not make judgment
    - for 2016, judgments not returned to participant
- Digest task results uploaded to NIST after eval period ended

Real-time Summarization

- Task A:
  - return at most 10 tweets/topic/day
  - lag between time tweet available and decision to return it to user should be minimized
  - scored using Expected Latency Gain (ELG)
- Task B:
  - return at most 100 [ranked] tweets/topic/day
  - decision period anytime within day is fine
  - scored using nDCG
- For both,
  - Automatic, Manual Preparation or Manual Interaction runs
  - manual clustering of relevant tweets define equivalence classes used for redundancy penalties in scoring
  - relevance judgment for unjudged tweets in equivalence class (eg, retweets) assigned as function of judged tweets in class

Real-time Summarization

- Profiles
  - combination of re-used TREC 2015 and newly developed profiles (total of 203)
  - describe prospective information need
- Judgments
  - mobile assessors judged tweets as relevant, redundant, not relevant
  - NIST assessors judged pools formed from both task A and task B runs
    - 3-way scale of not relevant, relevant, highly relevant
    - created clusters of semantically equivalent relevant tweets
  - 56 profiles with NIST judgments; 123 w/ mobile

Real-time Sample Topics

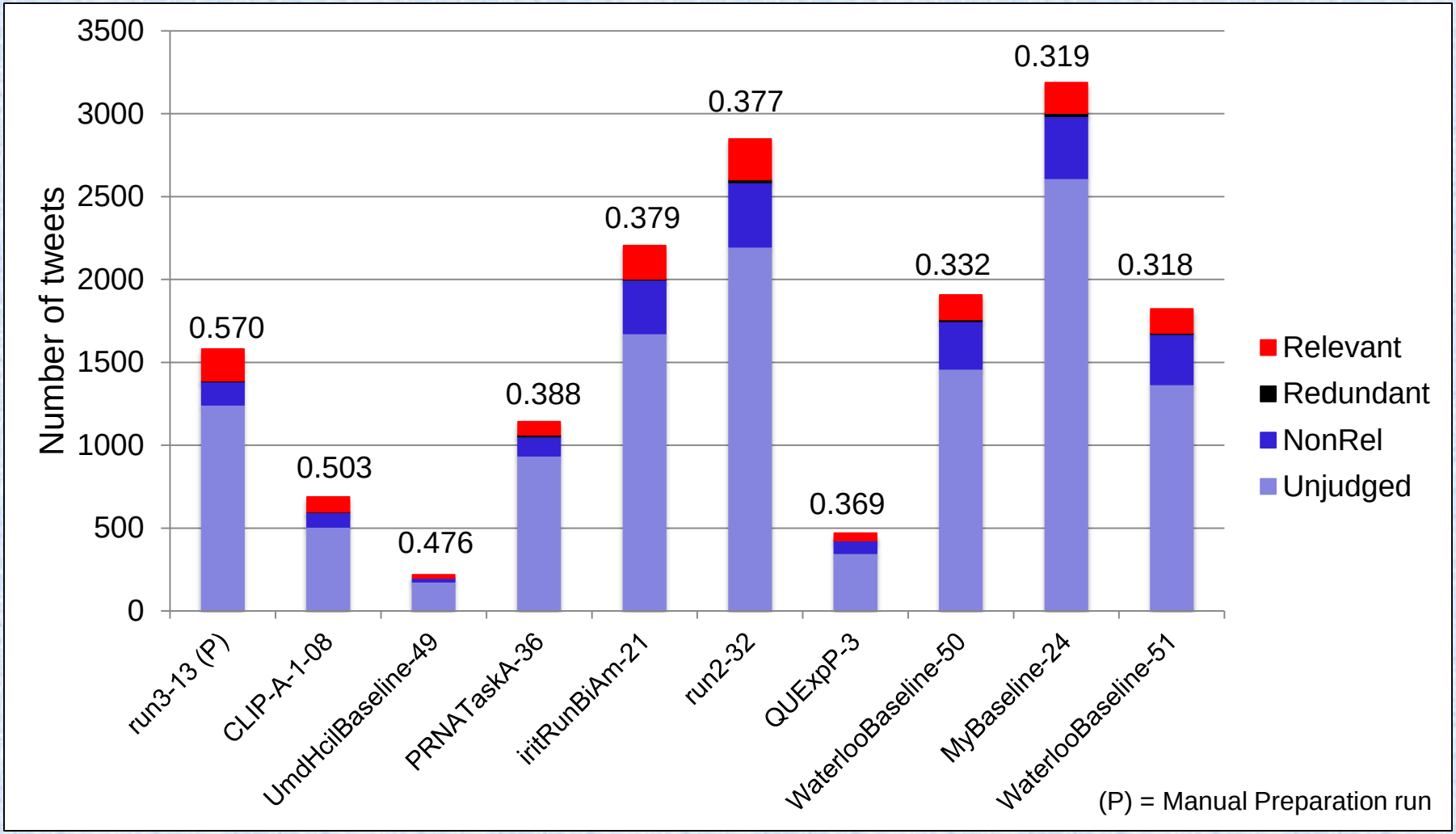
Title: Hershey, PA quilt show
Description: Find information on the quilt show being held in Hershey, PA
Narrative: The user is a beginning quilter who would like to attend her first quilt show. She has learned that a major quilt show will happen in Hershey, PA, and wants to see Tweets about the show, including such things as announcement of classes, teachers or vendors attending the show; prize-winning quilts; comments on logistics, travel information, and lodging; opinions about the quality of the show.

Title: FIFA corruption investigation
Description: Find information related to the ongoing investigation of FIFA officials for corruption.
Narrative: The user is a soccer fan who is interested in the current status of the ongoing investigation by various governments of corruption and bribery by officials of FIFA (Federation Internationale de Football Association). This includes tweets giving information on various investigations and possible rebidding of the 2018 and 2022 World Cup games.

Title: Mount Rushmore
Description: Find tweets about people's reactions to and experiences when visiting Mount Rushmore.
Narrative: The user is considering a trip to South Dakota to see Mount Rushmore. She would like to see what reaction other tourists have had to the site as well as any traveling tips and advice to make the trip more enjoyable.

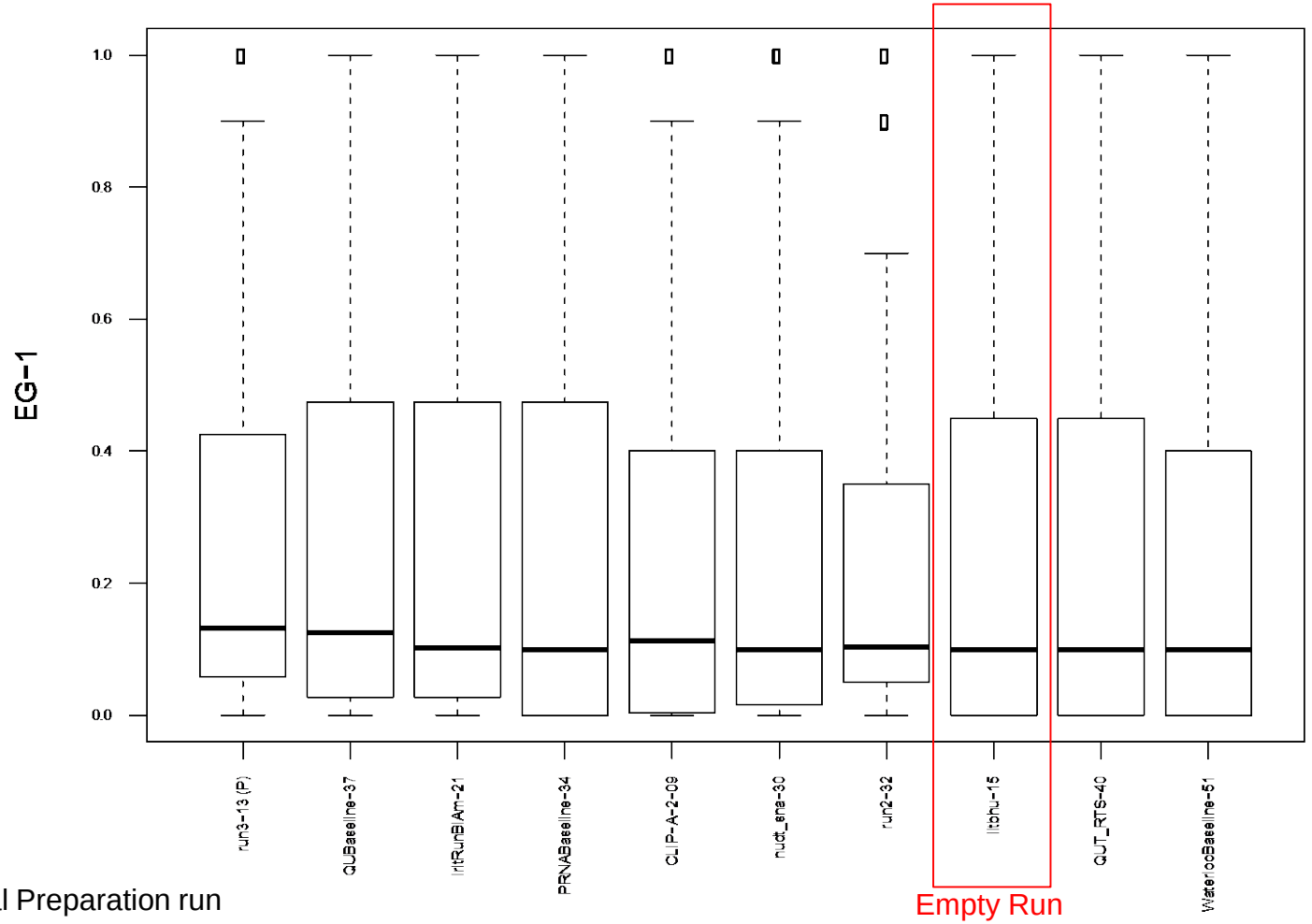
Top Task A Runs: Mobile Judgments

Ordered by Strict Precision



Top Task A Runs: NIST Judgments

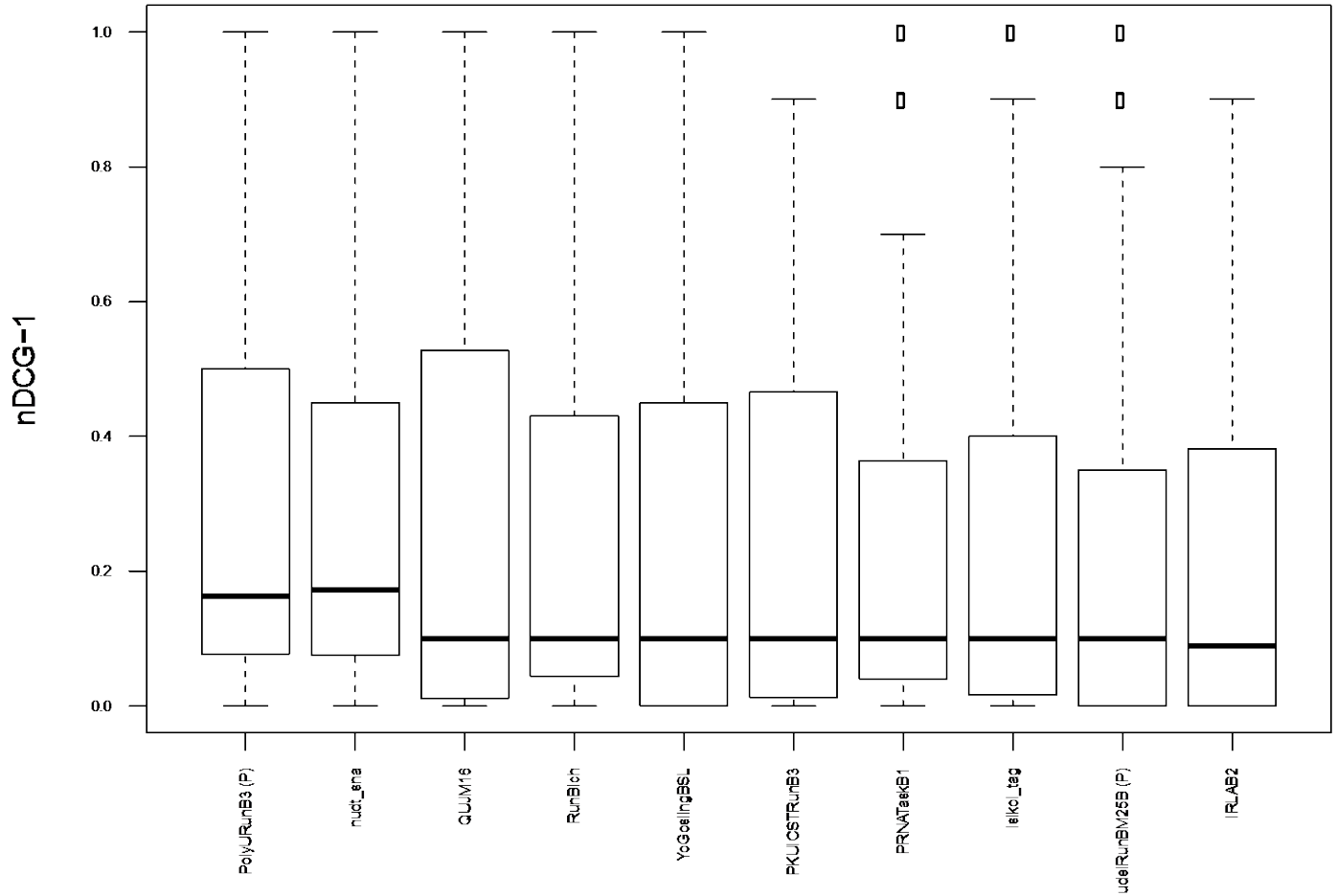
Distribution of Per-topic EG-1 Scores for Best Run by Mean EG-1



(P) = Manual Preparation run

Top Task B Runs

Distribution of Per-topic nDCG-1 Scores for Best Run by Mean nDCG-1



(P) = Manual Preparation run

TREC 2017

- Tracks
  - Dynamic Domain, Live QA, OpenSearch, and Real-time Summarization, and Tasks tracks continuing
  - CDS→ Precision Medicine
  - new track: Complex Answer Retrieval
- TREC 2017 track planning sessions
  - 1.5 hours per track tomorrow (3- or 4-way parallel)
  - track coordinators attending 2016
  - you can help shape task(s); make your opinions known


Things I Hate
1. T-Shirts
2. Lists
3. Comic Sans
4. Irony