Overview of TREC 2015

Ellen Voorhees

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

TREC 2015 Track Coordinators

Clinical Decision Support: Kirk Roberts, Ellen Voorhees, Bill Hersh

Contextual Suggestion: Adriel Dean-Hall, Charlie Clark, Jaap Kamps,
Julia Kiseleva

Dynamic Domain: Grace Hui Yang, John Frank, Ian Soboroff

Live QA: Eugene Agichtein, David Carmel, Donna Harman

Microblog: Miles Efron, Jimmy Lin

Tasks: Ben Carterette, Nick Craswell, Evangelos Kanoulas,
Manisha Verma, Emine Yilmaz

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

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

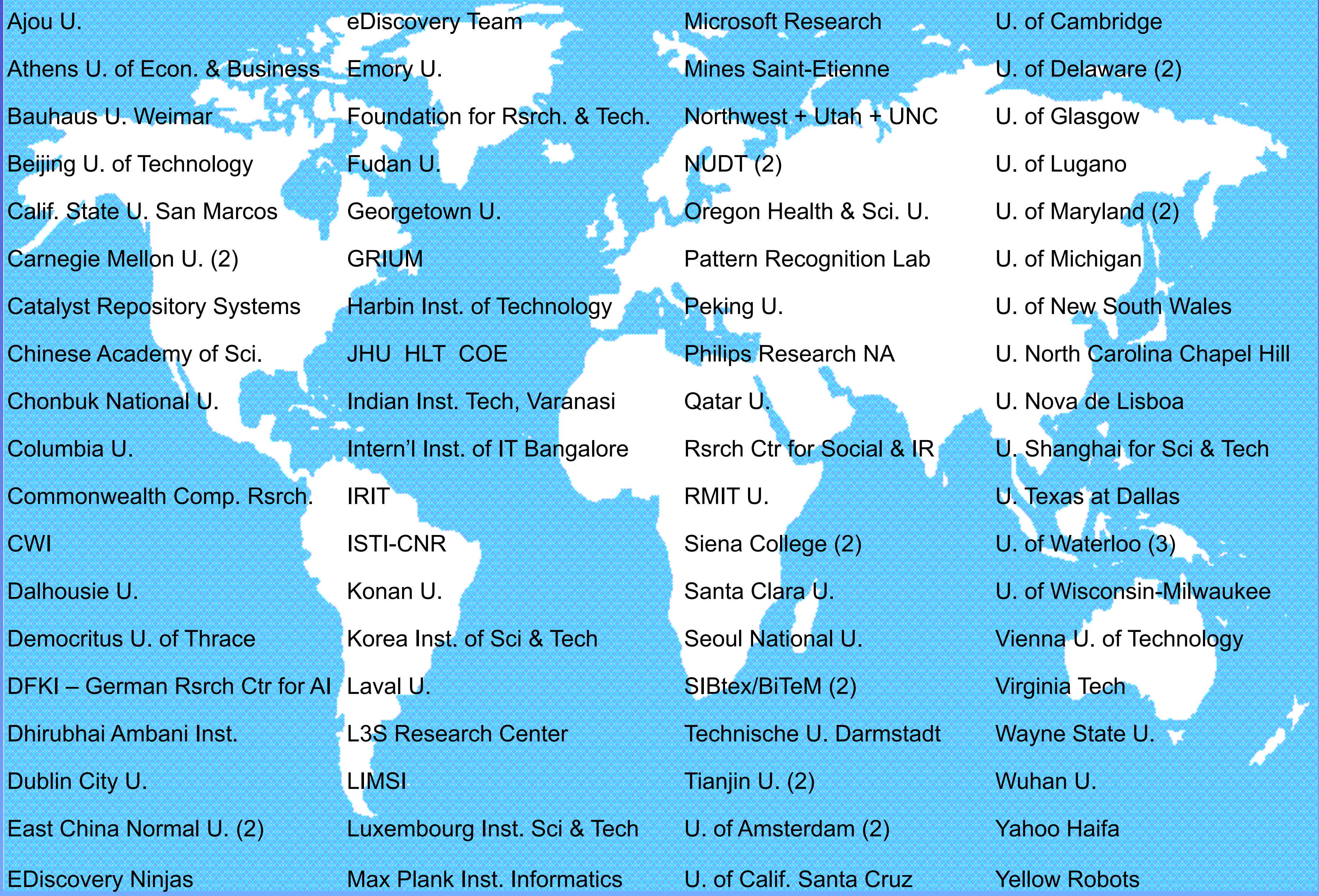
TREC 2015 Program Committee

Ellen Voorhees, chair

James Allan
Chris Buckley
Ben Carterette
Gord Cormack
Sue Dumais
Donna Harman
Diane Kelly

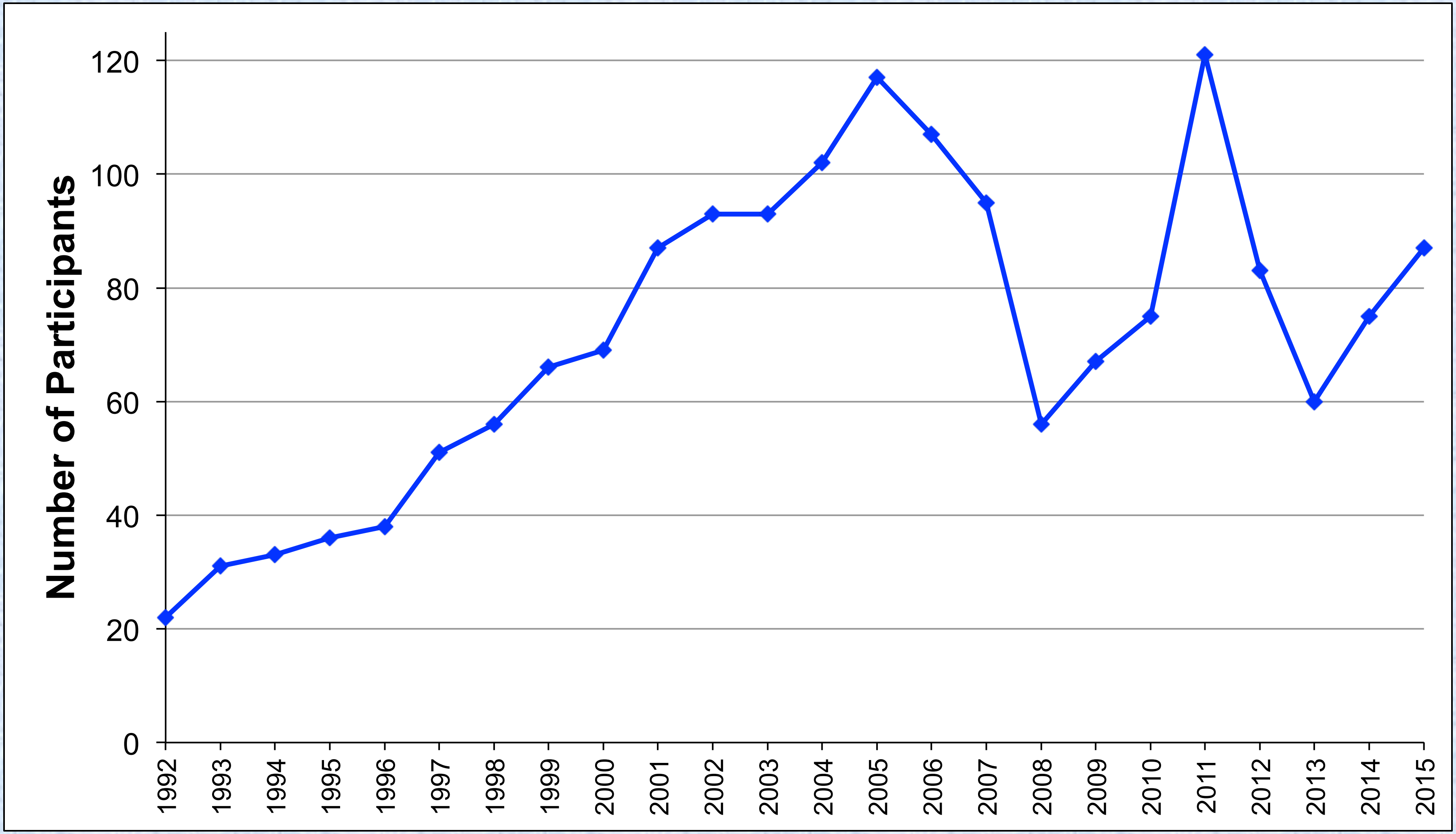
David Lewis
Paul McNamee
Doug Oard
John Prager
Ian Soboroff
Arjen de Vries

87 TREC 2015 Participants



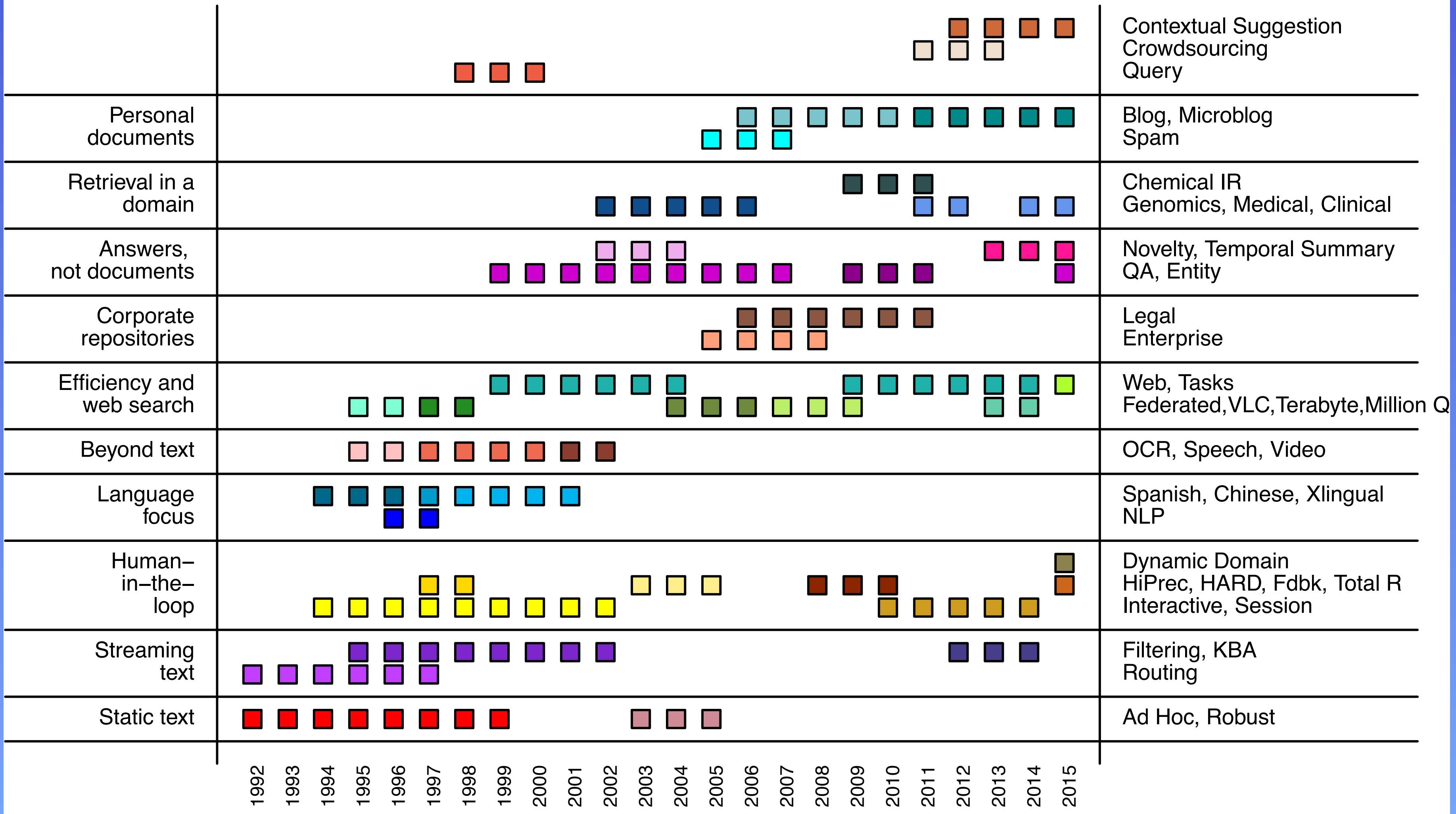
- Ajou U.
- Athens U. of Econ. & Business
- Bauhaus U. Weimar
- Beijing U. of Technology
- Calif. State U. San Marcos
- Carnegie Mellon U. (2)
- Catalyst Repository Systems
- Chinese Academy of Sci.
- Chonbuk National U.
- Columbia U.
- Commonwealth Comp. Rsrch.
- CWI
- Dalhousie U.
- Democritus U. of Thrace
- DFKI – German Rsrch Ctr for AI
- Dhirubhai Ambani Inst.
- Dublin City U.
- East China Normal U. (2)
- EDiscovery Ninjas
- eDiscovery Team
- Emory U.
- Foundation for Rsrch. & Tech.
- Fudan U.
- Georgetown U.
- GRIUM
- Harbin Inst. of Technology
- JHU HLT COE
- Indian Inst. Tech, Varanasi
- Intern'l Inst. of IT Bangalore
- IRIT
- ISTI-CNR
- Konan U.
- Korea Inst. of Sci & Tech
- Laval U.
- L3S Research Center
- LIMSI
- Luxembourg Inst. Sci & Tech
- Max Plank Inst. Informatics
- Microsoft Research
- Mines Saint-Etienne
- Northwest + Utah + UNC
- NUDT (2)
- Oregon Health & Sci. U.
- Pattern Recognition Lab
- Peking U.
- Philips Research NA
- Qatar U.
- Rsrch Ctr for Social & IR
- RMIT U.
- Siena College (2)
- Santa Clara U.
- Seoul National U.
- SIBtex/BiTeM (2)
- Technische U. Darmstadt
- Tianjin U. (2)
- U. of Amsterdam (2)
- U. of Calif. Santa Cruz
- U. of Cambridge
- U. of Delaware (2)
- U. of Glasgow
- U. of Lugano
- U. of Maryland (2)
- U. of Michigan
- U. of New South Wales
- U. North Carolina Chapel Hill
- U. Nova de Lisboa
- U. Shanghai for Sci & Tech
- U. Texas at Dallas
- U. of Waterloo (3)
- U. of Wisconsin-Milwaukee
- Vienna U. of Technology
- Virginia Tech
- Wayne State U.
- Wuhan U.
- Yahoo Haifa
- Yellow Robots

Number of Participants in TREC



A big thank you to our assessors

TREC Tracks


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

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@X
  - filtering measures: F, expected latency gain (ELG)

TREC 2015

- High barrier for participation
  - first year for 4/8 tracks
    - "Tasks" track has non-standard task
  - engineering challenges
    - Live QA, Contextual Suggestion Live task
    - live tweet monitoring in Microblog track
    - write to jig API in Dynamic Domain, Total Recall
- Emphasis on time
    - filtering tasks with latency penalties
    - live tasks have performance demands

New Feature inTREC 2015

- Added `Open Runs'
  - to increase repeatability/reproducibility of IR experiments, encouraged participants to package system that produced a submission into a github repository
  - URL of that github object provided at submission time and included in run description

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 in late August
  - 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 eight top-level Yahoo Answer categories, as self-labeled by asker
  - lightly filtered to remove objectionable material
  - final test set of 1087 questions
- 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 Sample Questions

Category: Health

*Have I stopped growing yet? Ok so I'm 14 years old and I think I stopped growing im 5'4 rn I got a deep voice at age 10 and now I am growing a beard and ***. My mom is 5'2 and my dad is 5'10. Any help?*

Category: Beauty & Style

Workout fast? So, I'm going on holidays in a week and really want to get toner. I have a bikini body guide and I was wondering if I did a week in one day every day this week will it be as effect as doing it for 7 weeks??

Category: Computers & Internet

My laptop can support 1080p, so how come when I watch a video on Youtube it's usually on 480p?

Category: Pets

My 105 lb. Lab mix ate part of a box of raisin bran. She is acting normal. No vomitting etc. Should I be worried?

Category: Home & Garden

Is it safe to use diluted clorox to get stains off tea cups?

Category: Sports

Which is worse? Gambling or cheating with PEDs on the game of baseball? Why?

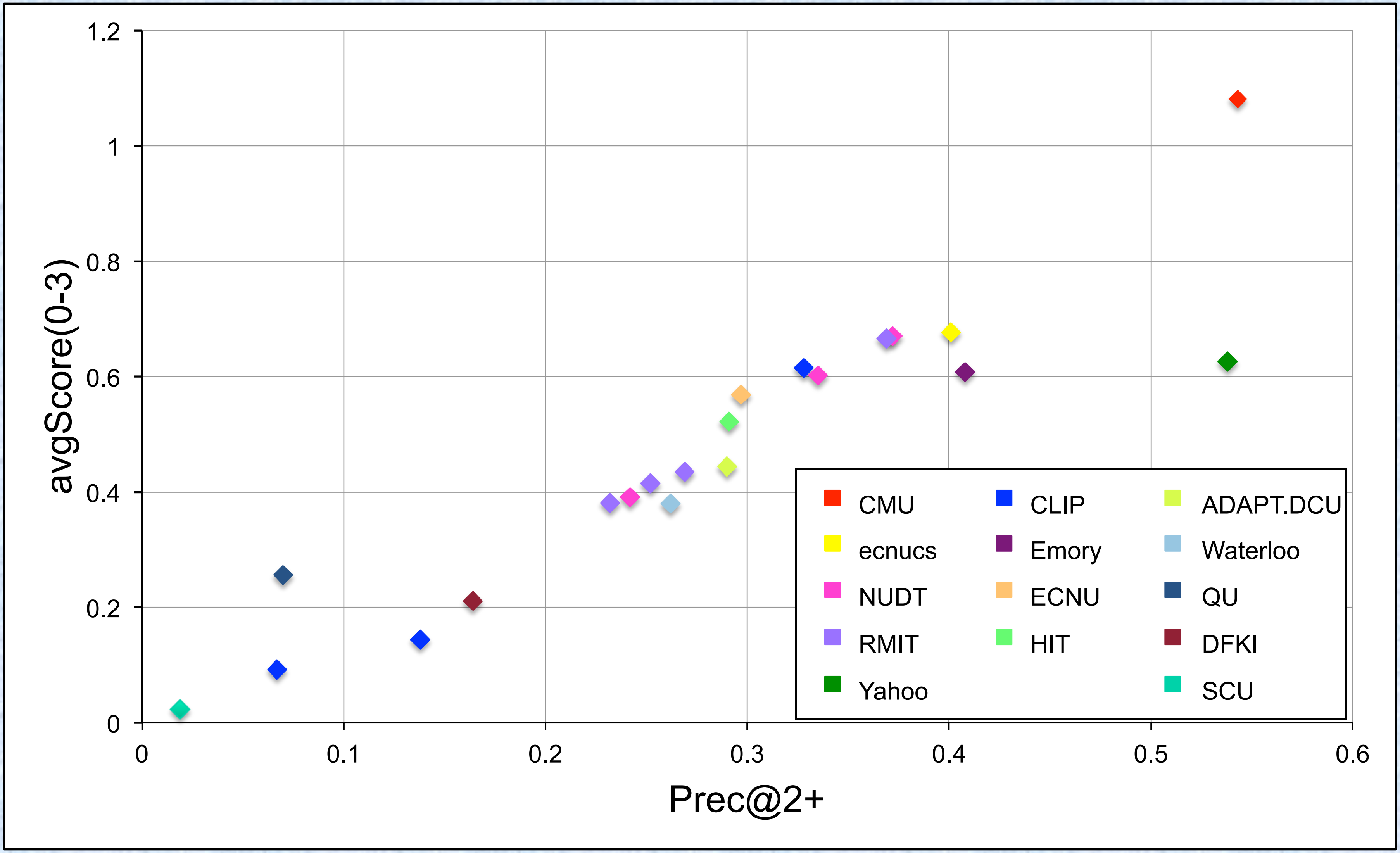
Category: Arts & Humanities

Was the Victorian bustle designed to conceal a women's bottom, or create an exaggerated illusion that highlighted it?

Category: Travel

What is the best convenient way to go to Kerala?

Live QA Results



Contextual Suggestion

- "Entertain Me" app: suggest activities based on user's prior history and target location
- Fourth edition of track, with major rework this year
  - new live task introduced
  - 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

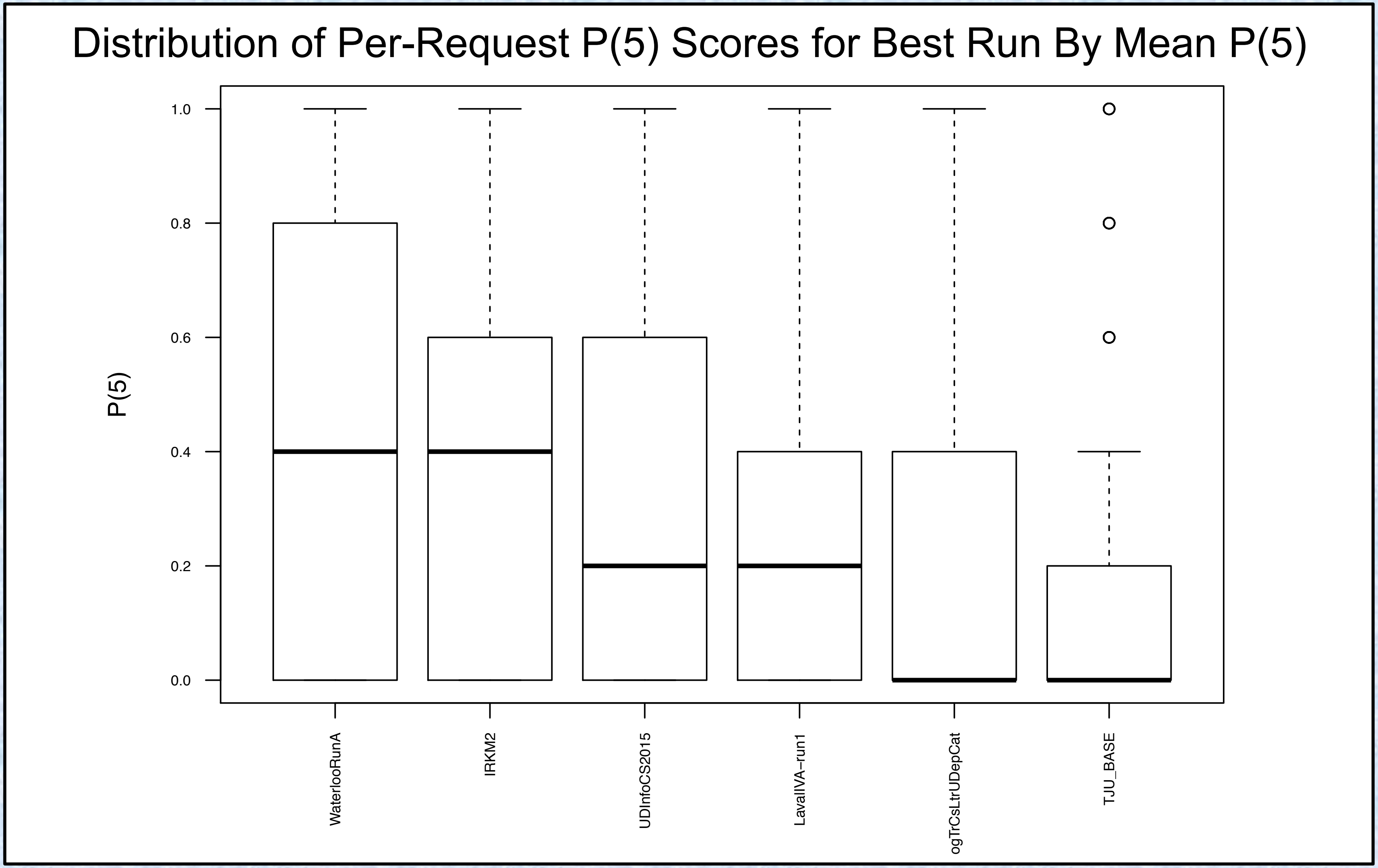
doc: 00405444-152 **rating**: 3

tags: Art, Art Galleries, Family Friendly, Fine Art Museums

Contextual Suggestion

- Live task
  - 3 week evaluation period in late July
  - systems received requests and responded with their suggestions
  - suggestions from all Live participants pooled and sent back to requestor (Mechanical Turk-er) for ratings and feature tags
  - requestor might issue new request; all previously rated suggestions from this requestor included in new request
  - total of 380 requests in test set
  - ratings on 5-point scale Strongly Uninterested—Strongly Interested; top 2 counted as 'relevant'

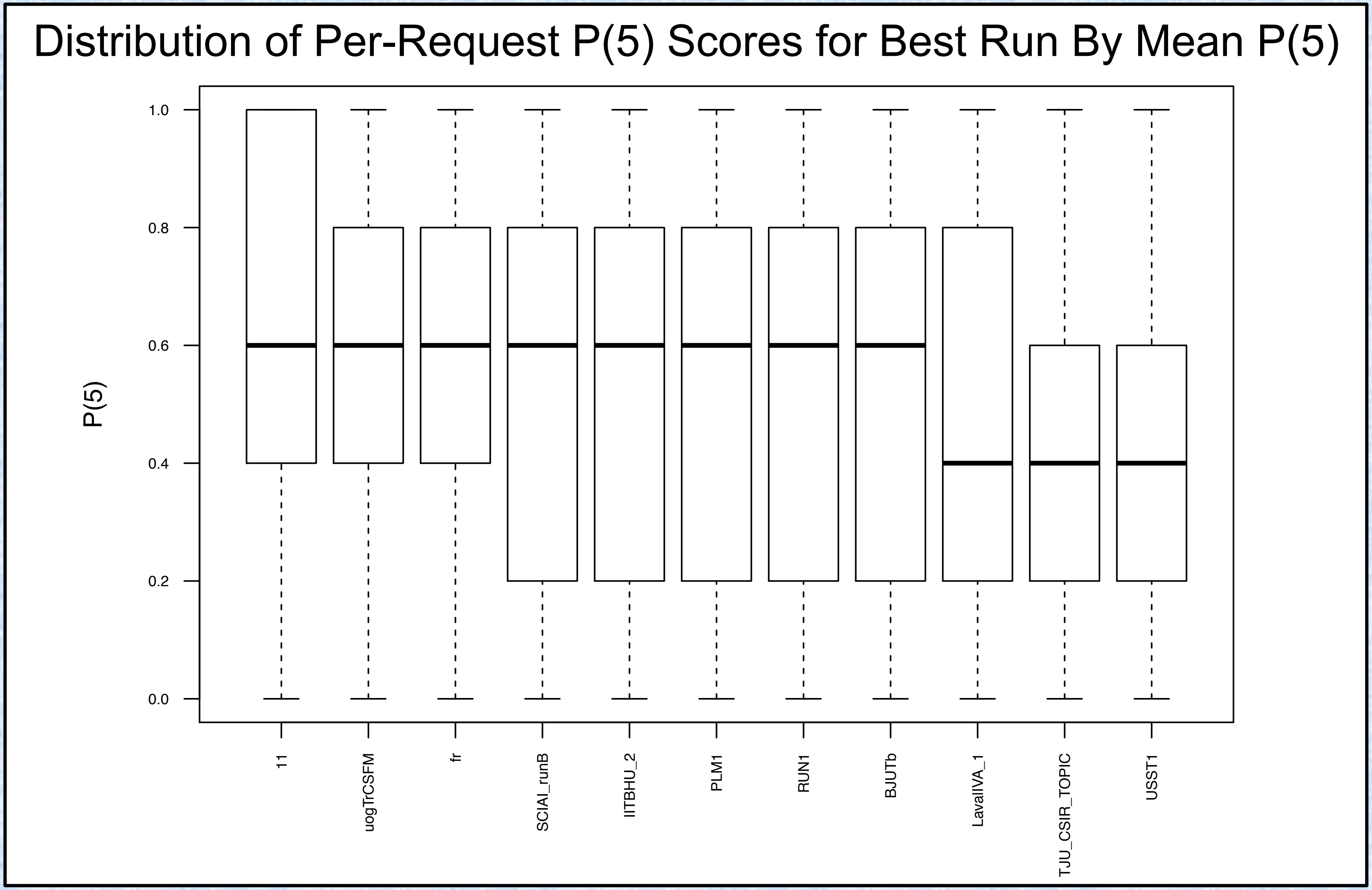
Contextual Suggestion Live Results



Contextual Suggestion

- Batch task
  - test set consisted of 211-request subset of requests used in Live task
  - a Batch task request also contained the complete set of (unrated) suggestions from all live task participants for the request; Batch task participants were required to return only suggestions from this set

Contextual Suggestion Batch Results



Total Recall

- Goal
  - evaluate methods for achieving very high recall, including methods that use a human-in-the-loop
  - as such, a successor to the interactive track, but with a focus on recall rather than precision
- Implementation
  - participant system submits a document at a time to a software jig; jig both records activity & responds to system with relevance judgment for doc
  - participant decides when to terminate search; entire set of documents submitted to jig counts as retrieved set

Total Recall

At Home Collections

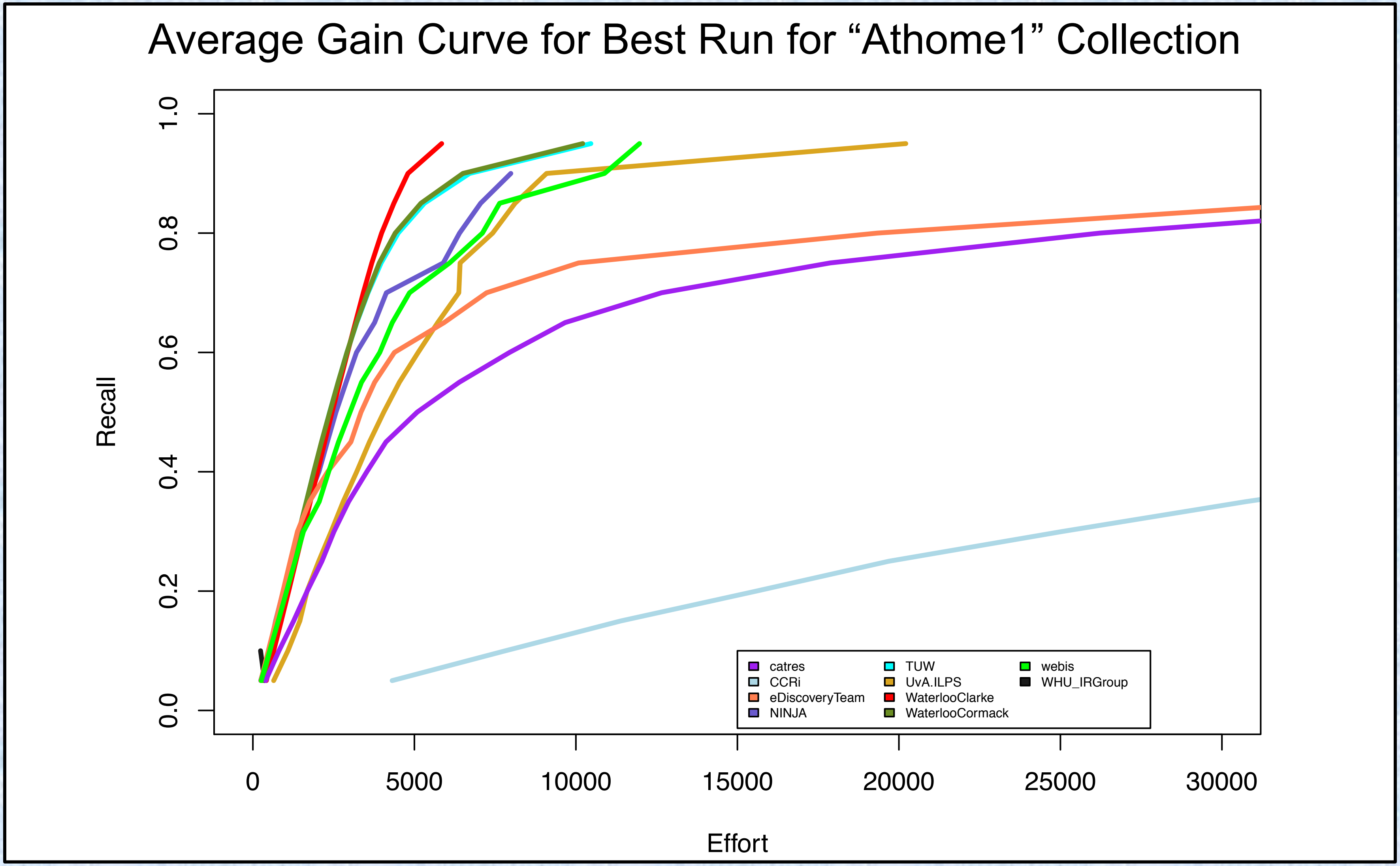
Jeb Bush email: ten topics against the email of Florida governor Jeb Bush
Illicit Goods: ten topics from the Dynamic Domain track's Illicit Goods domain
Local Politics: ten topics from the Dynamic Domain track's Local Politics domain

Sandbox Collections

Kaine email: four topics corresponding to archivists' categories against the email of Virginia Governor Tim Kaine
MIMIC II: nineteen topics corresponding to top-level ICD-9 codes against text-based fields of clinical records

- Tasks
  - "**at home**": systems connect to jig over Internet; participant's machine contains document set and search runs there. "Limited" participation allowed: first At Home collection only
  - **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 Results



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
  - similar jig as in Total Recall track's At Home task; jig referred to as Simulated User
  - participants submit docs to Simulated User and get judgments for individual facets of the topic
  - system decides to stop when it thinks sufficient info for all facets has been retrieved

Dynamic Domain

- Domains
    - three domains with a total of 118 topics
    - Illicit Goods, Ebola, Local Politics
- Topics
    - developed by NIST assessors who 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
    - [but, goal of good coverage not met. This led to two tasks: main task and 'Judged-only' task where participants could search/submit judged docs only]

Dynamic Domain Sample Topics

Illicit Goods

Topic: Silk Road 1 marketplace shut down

Discuss Ross Ulbricht's underground black market, the Silk Road

Subtopic 1: Silk road founder found guilty
Subtopic 2: What was Silk Road
Subtopic 3: How it was taken down
Subtopic 4: Alternatives to Silk Road
Subtopic 5: About Ross Ulbricht, the alleged founder

Ebola

Topic: Hand washing importance

Find information on hand washing to prevent the spread of Ebola.

Subtopic 1: training
Subtopic 2: proper technique
Subtopic 3: recommendations
Subtopic 4: Nigerian campaign
Subtopic 5: Sierra Leone campaign
Subtopic 6: disease transmission

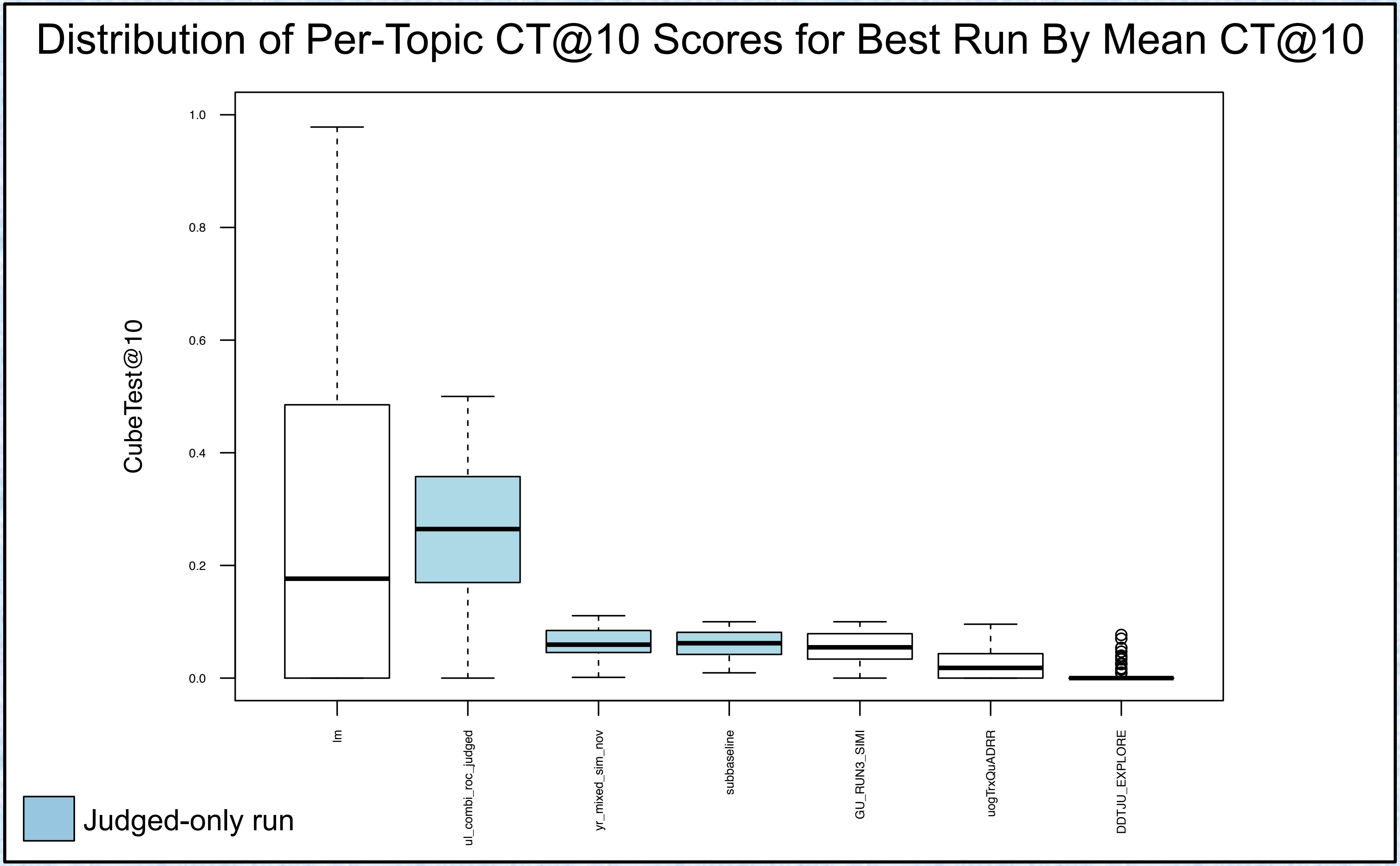
Local Politics

Topic: Washington liquor sale privatization

Find info on Washington state's initiative 1183 to privatize state liquor stores.

Subtopic 1: costco backing of initiative
Subtopic 2: privatization of liquor stores (WA)
Subtopic 3: revenue effect of liquor sale privatization

Dynamic Domain Results



Microblog

- Goal
    - examine search tasks for information seeking behaviors in microblogging environments

- 2015 track significantly revamped
    - filtering task using live Tweet stream
        - Task A: deliver updates to mobile device
        - Task B: periodic digest of updates
    - participants required to listen to stream for entire evaluation period (~10 days in late July)
    - uploaded final sets of retrieved Tweets to NIST at conclusion of evaluation period

Microblog

- Topics
  - 225 in test set; 51 in scoring set
  - syntactically the same as traditional topic statements, but describe <u>prospective</u> information need rather than retrospective
  - developed by NIST assessors in June; they constructed topics that they projected might get tweets in late July
  - same assessor as created topic judged it
    - 3-way scale of not relevant, relevant, highly relevant

Microblog 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.

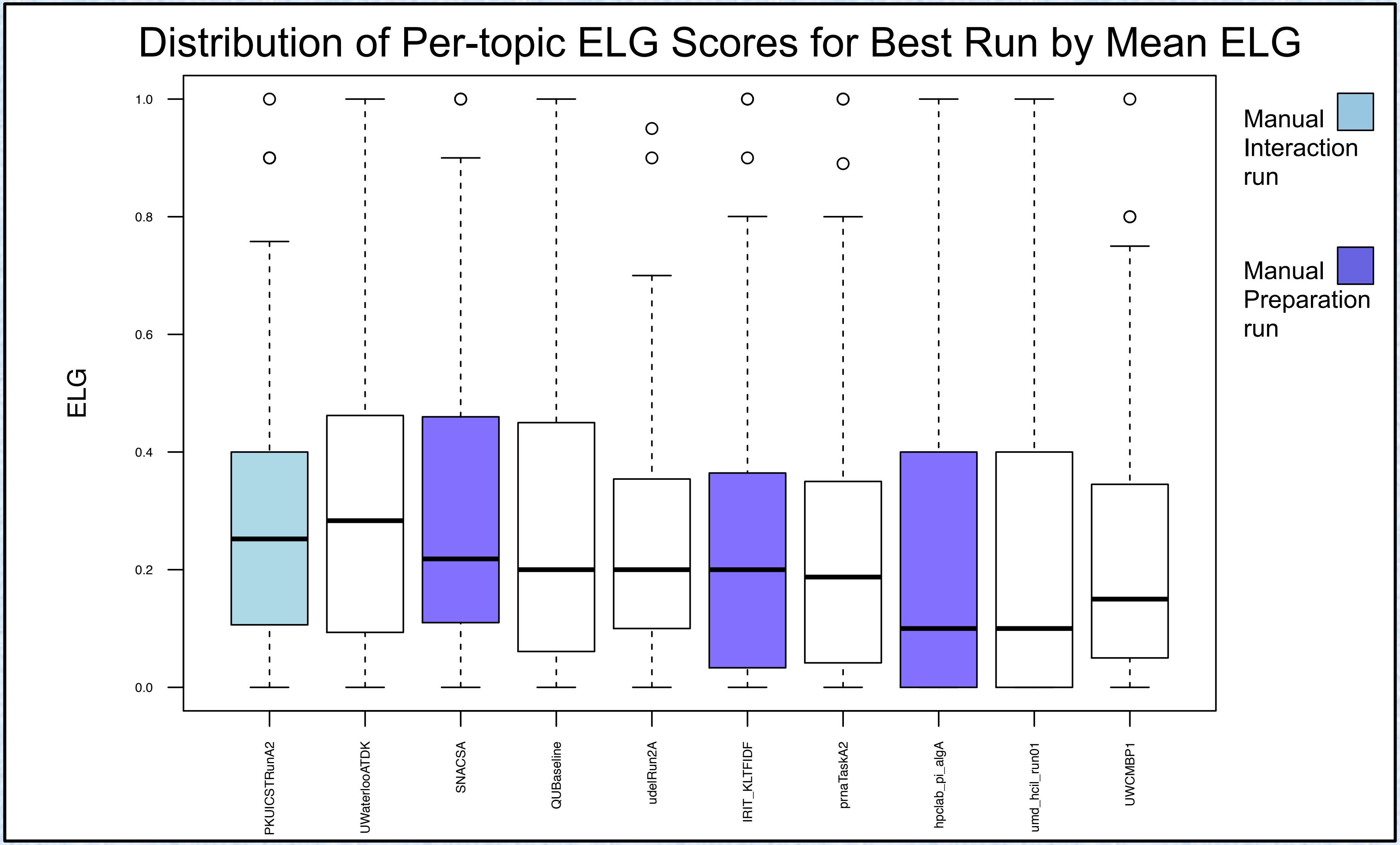
Title: FIFA corruption investigation
Description: Find information related to the ongoing investigation of FIFA officials for corruption.
Narrative: The user is a soccer fan who is interested in the current status of the ongoing investigation by various governments of corruption and bribery by officials 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.

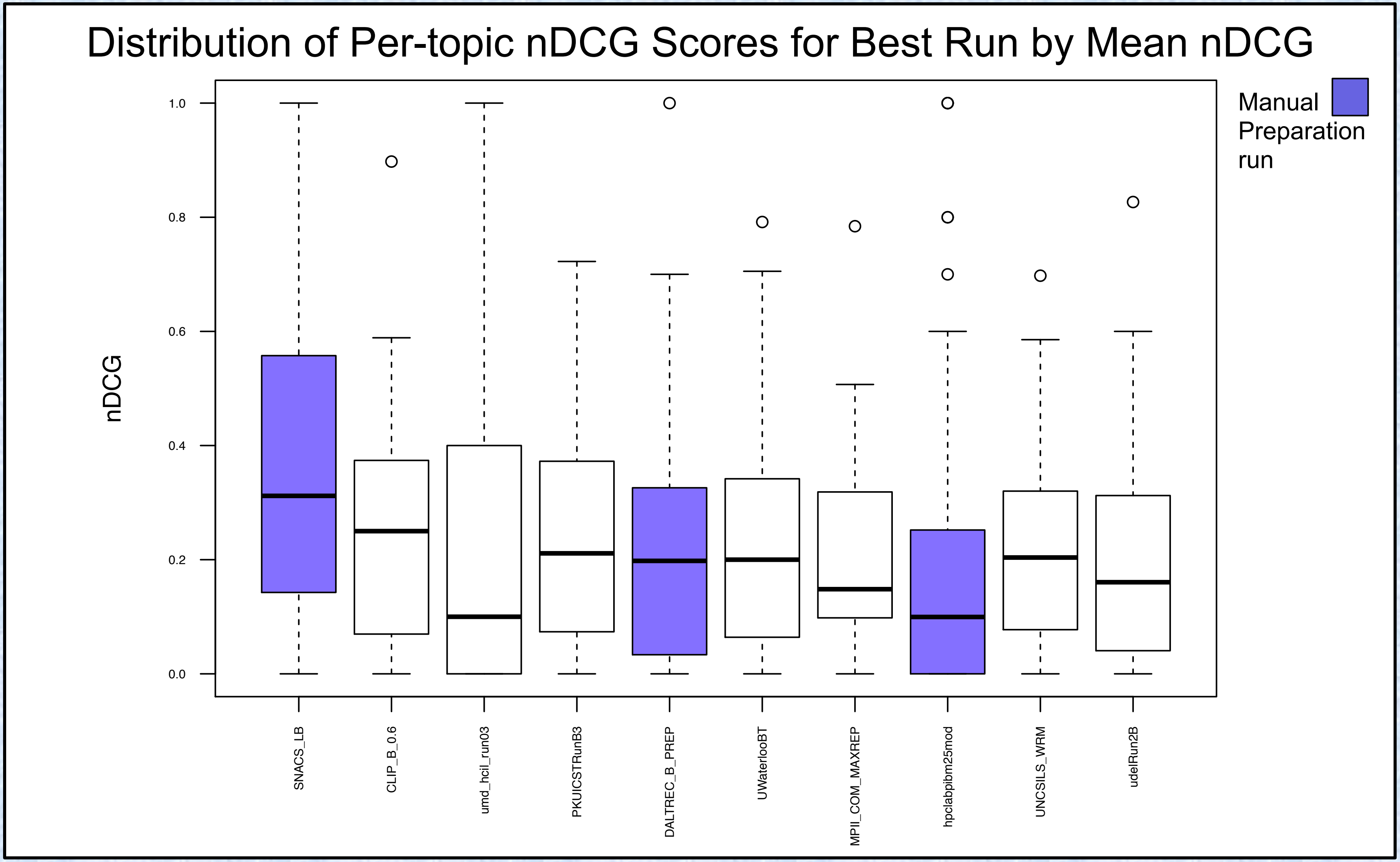
Microblog

- 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

Microblog Task A Results



Microblog Task B Results



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(s) of KBA Stream Corpus

  Filtering & Summarization, Pre-Filtered Summarization, Summarization Only

- 20 topics (events)
  - *each has a single type taken from* {accident, bombing, conflict, earthquake, protest, storm}

start: 1358323140 **end:** 1359619140
query: vauxhall helicopter crash
type: accident

start: 1351296000 **end:** 13518114400
query: cyclone nilam
type: storm

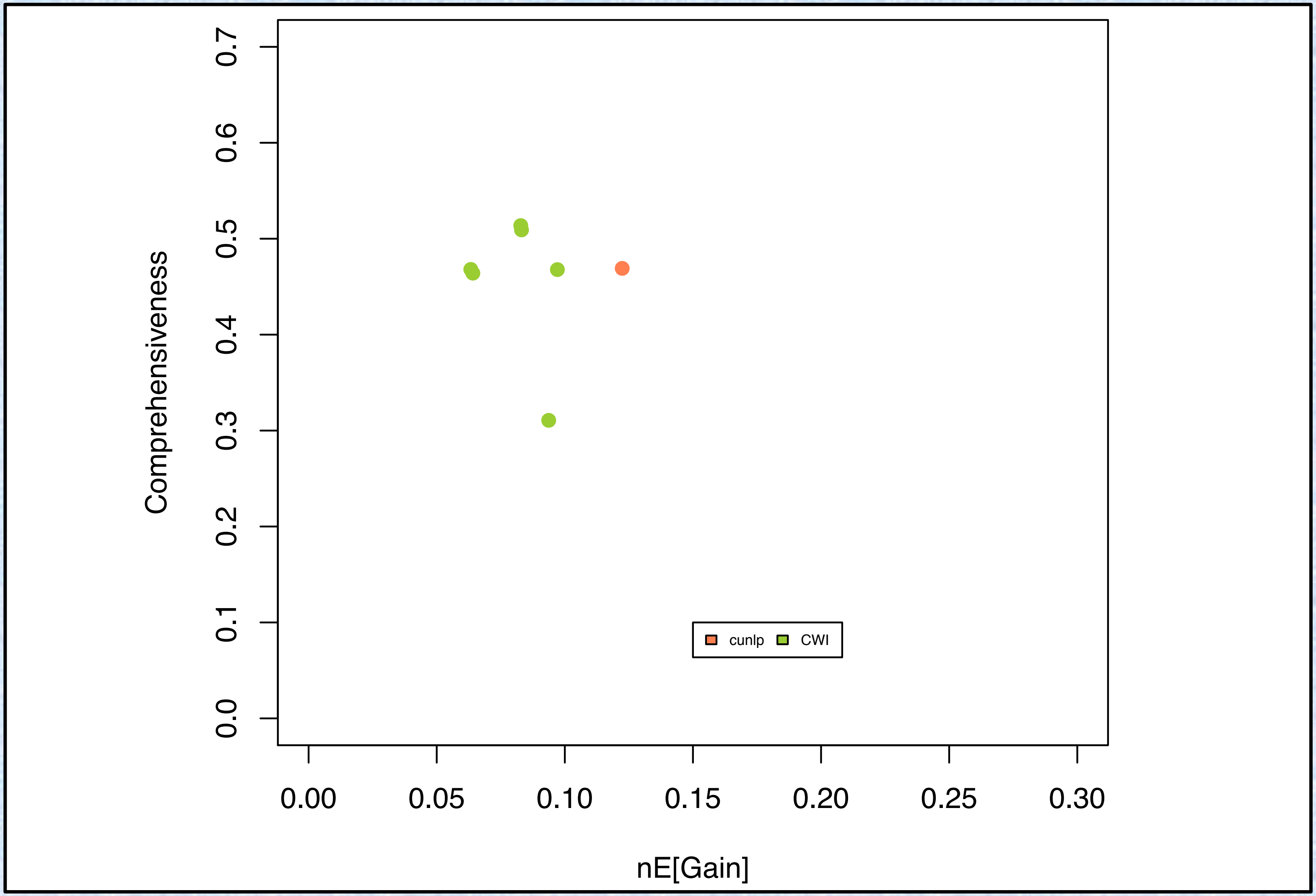
start: 13577776000 **end:** 1358553600
query: konna battle
type: conflict

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

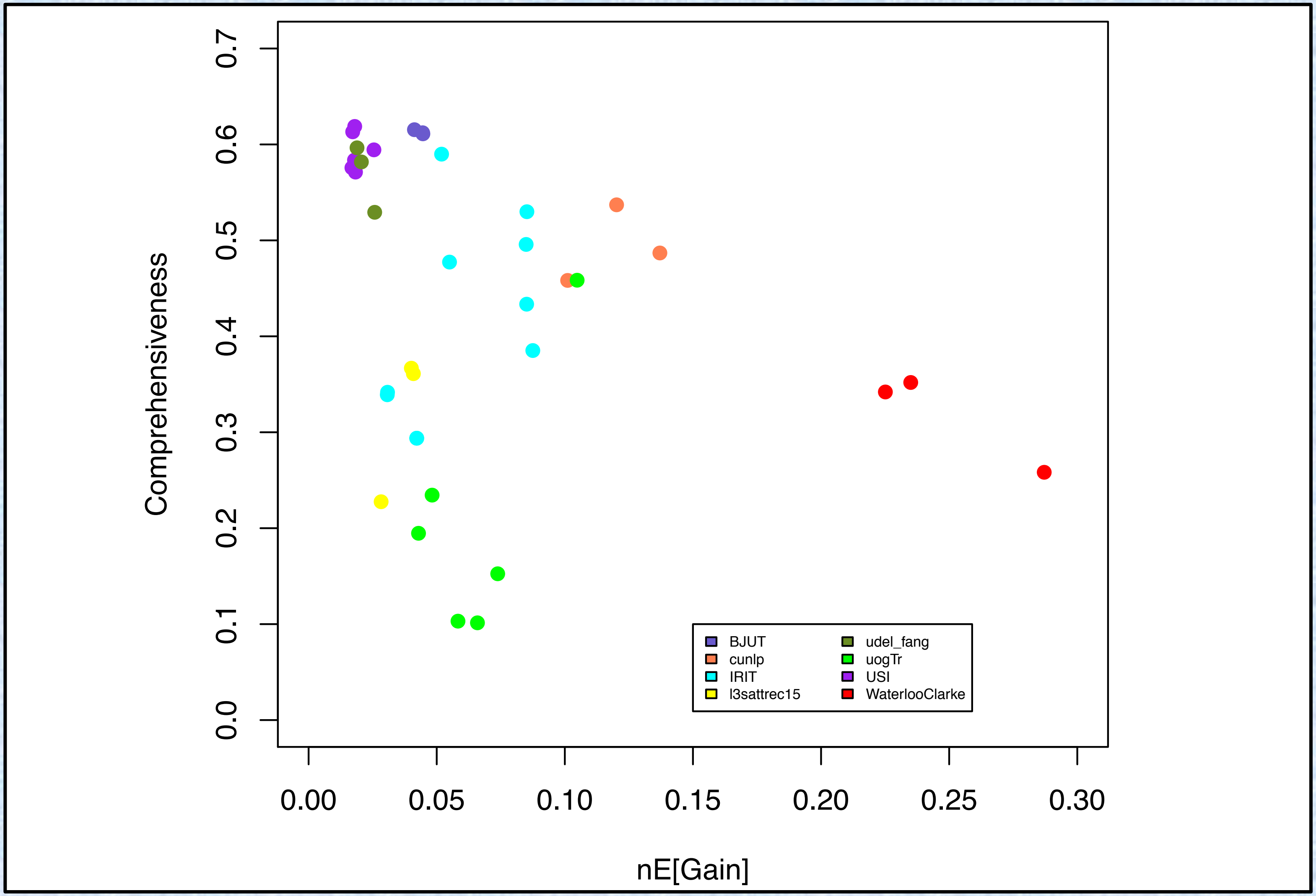
Temporal Summarization

Full Filtering and Summarization Task



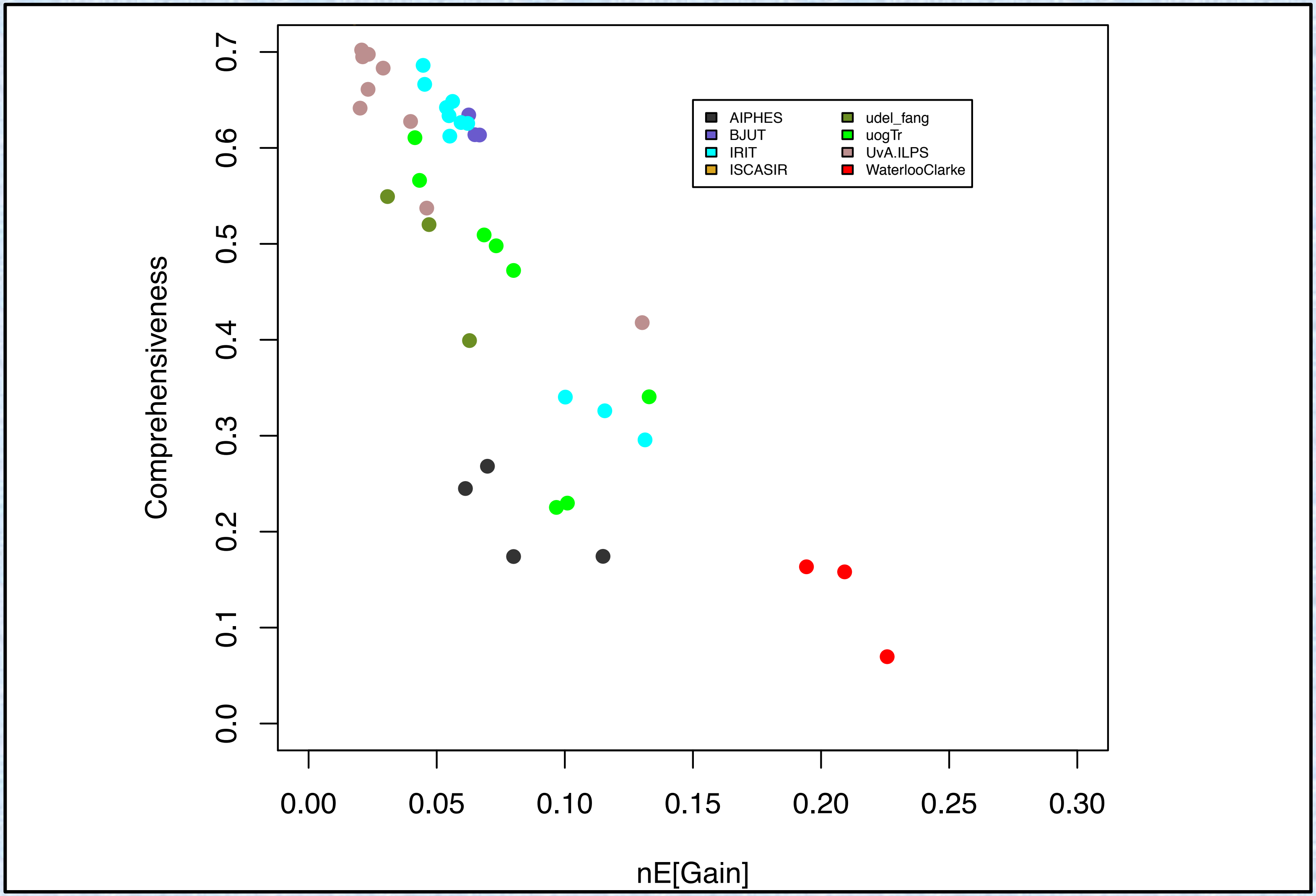
Temporal Summarization

Pre-Filtered Summarization Task



Temporal Summarization

Summarization Only Task



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
  - but only 34 (phrases) or 35 (documents) in evaluation 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: getting organized at work

I need to get organized at work

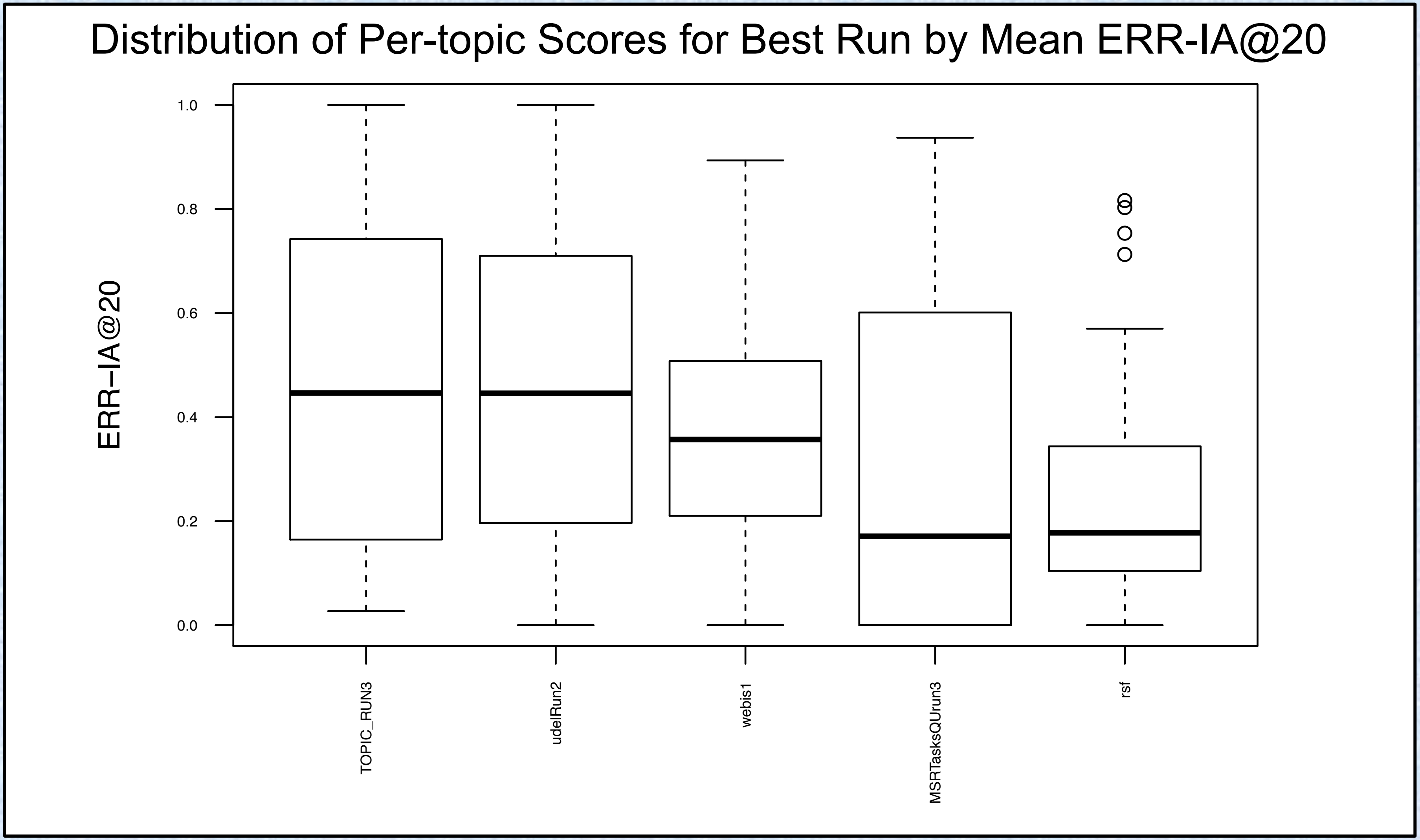
Subtask 1: Checklist for getting organized at work
Subtask 2: How to organize office desk
Subtask 3: Tips for getting organized at work
Subtask 4: Organize schedule at office
Subtask 5: How to create a todo/task list
Subtask 6: How to keep a calendar of scheduled meetings and travel
Subtask 7: How to set deadlines and goals
Subtask 8: How to organize your work space
Subtask 9: How to log the time you spend
Subtask 10: Methods to track your progress towards goals
Subtask 11: How to set up a filing system with a binder or folders

query: disneyland paris

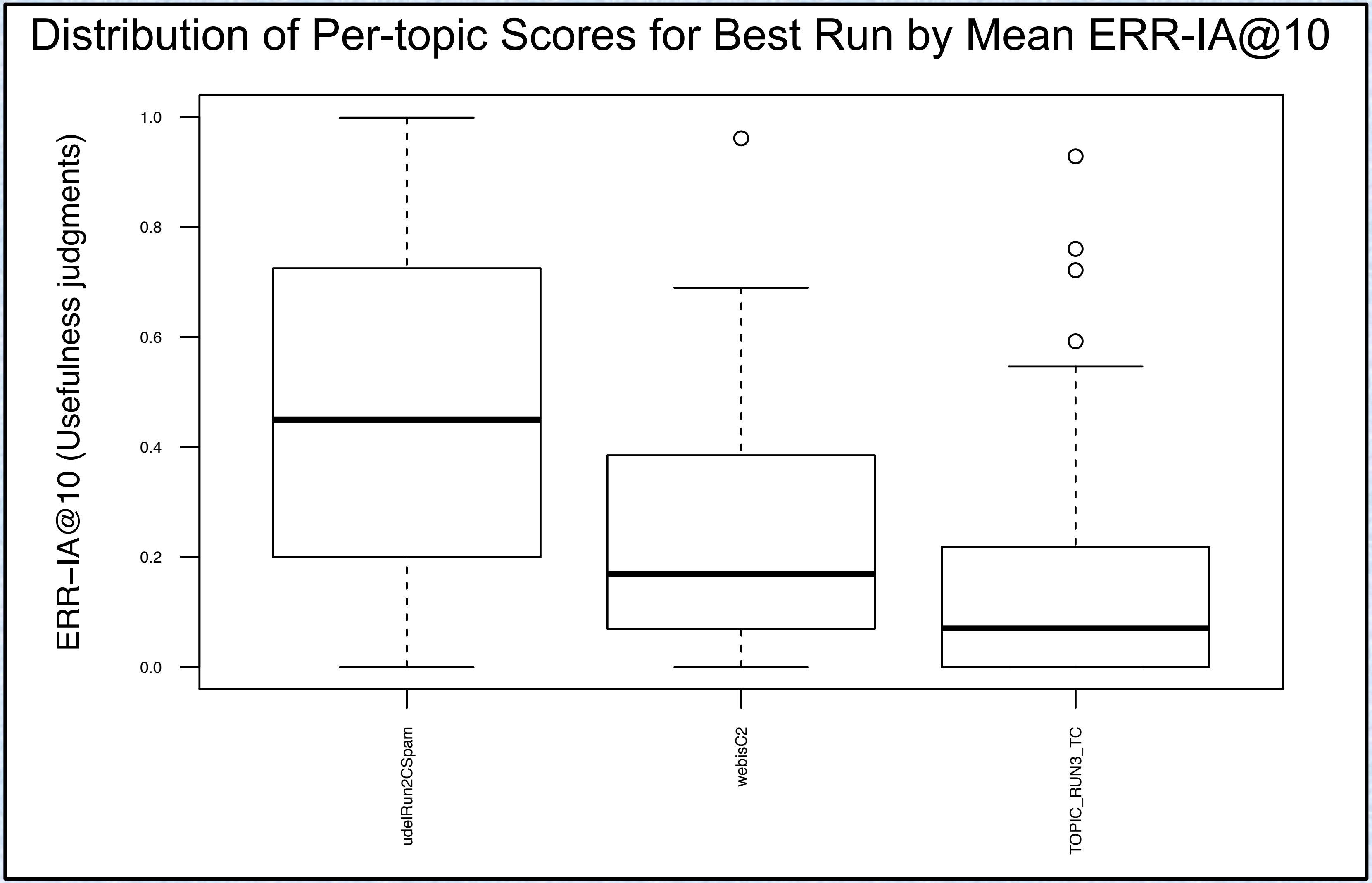
I'm planning my visit to Disneyland Paris.

Subtask 1: Information about Disneyland Paris
Subtask 2: Disneyland Paris entrance fee
Subtask 3: Book a hotel
Subtask 4: Choose the right tickets and buy them
Subtask 5: Book flights/trains
Subtask 6: Avoid queues
Subtask 7: Plan your visit, what to do, when
Subtask 8: Plan meals and drinks in and out of the park

Task Understanding Results



Task Completion Results



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:
  - open access subset of PubMed Central, a database of freely-available full-text biomedical literature
  - contains 733,138 articles in NXML
- 30 topics
  - case narratives developed by NIH physicians plus label designating target clinical task
  - 10 topics for each clinical task type
  - have both "description" & more focused "summary"
  - new for 2015, "B" version of topics gives diagnosis for test and treatment topics

CDS Track

- Judgments
  - judgment sets created using inferred measure sampling (2 strata; ranks 1-20; 20% of 21-100); main measure infNDCG
  - judgments made by physicians coordinated by OHSU
  - up to 3 runs per participant per task
  - all runs contribute to same set of pools

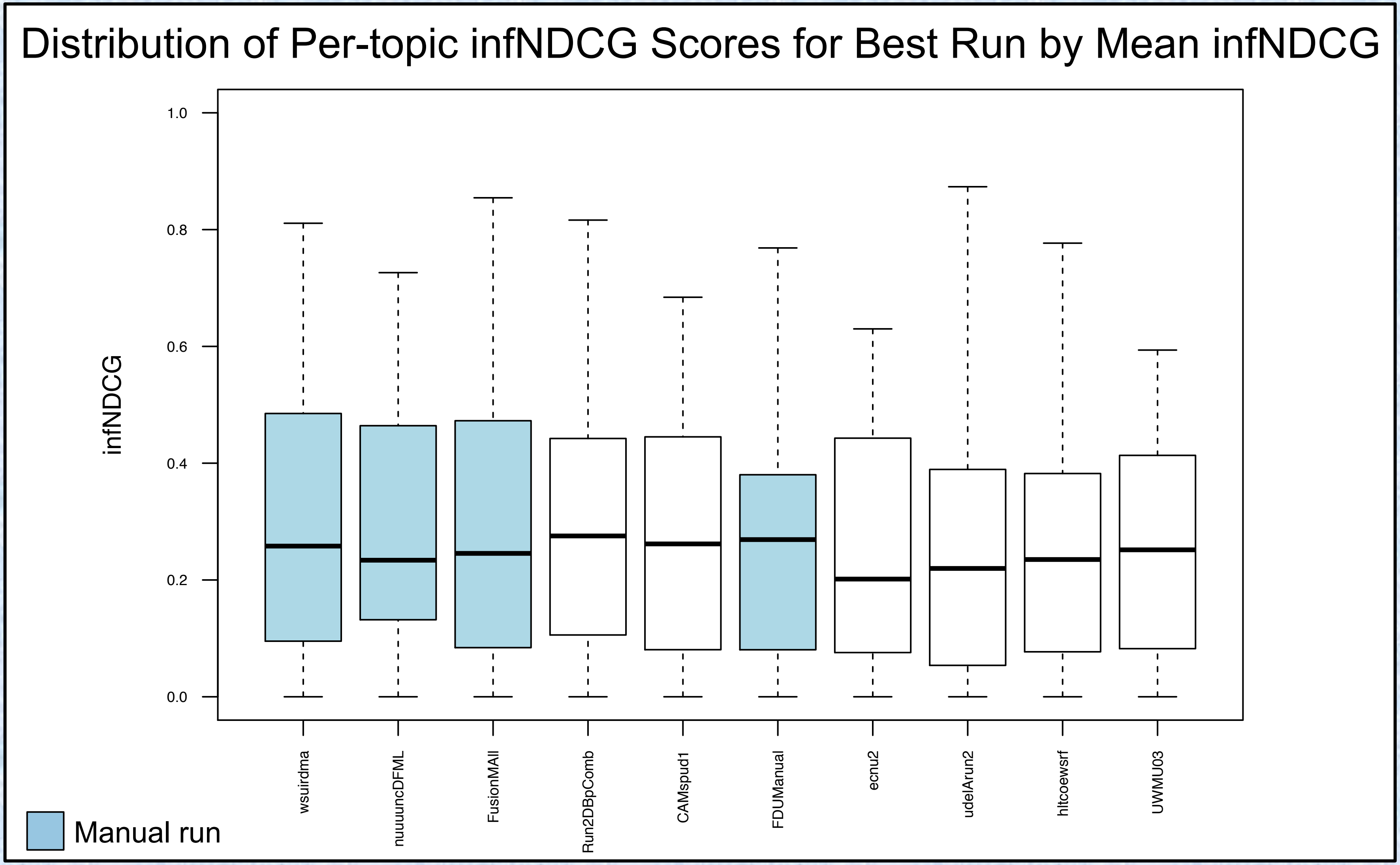
CDS Track Sample Topics

<topic number=“1" type="**diagnosis**”>
Description: A 44 yo male is brought to the emergency room after multiple bouts of vomiting that has a ‘coffee ground’ appearance. His heart rate is 135 bpm and blood pressure is 70/40 mmHg. Physical exam findings include decreased mental status and cool extremities. He receives a rapid infusion of crystalloid solution followed by packed red blood cell transfusion and is admitted to the ICU for further care.
Summary: A 44-year-old man with coffee-ground emesis, tachycardia, hypoxia, hypotension, and cool clammy extremities.

<topic number=”13" type="**test**">
Description: A 5-year-old boy presents to the emergency department with complaints of progressively worsening dysphagia, drooling, fever, and vocal changes. He is toxic-appearing, and leans forward while sitting on his mother’s lap. He is drooling and speaks with a muffled ‘hot potato’ voice. The parents deny the possibility of foreign body ingestion or trauma, and they report that they are delaying some of his vaccines.
Summary: A 5-year-old boy presents with difficulty breathing, stridor, drooling, fever, dysphagia and voice change.
Diagnosis: Epiglottitis

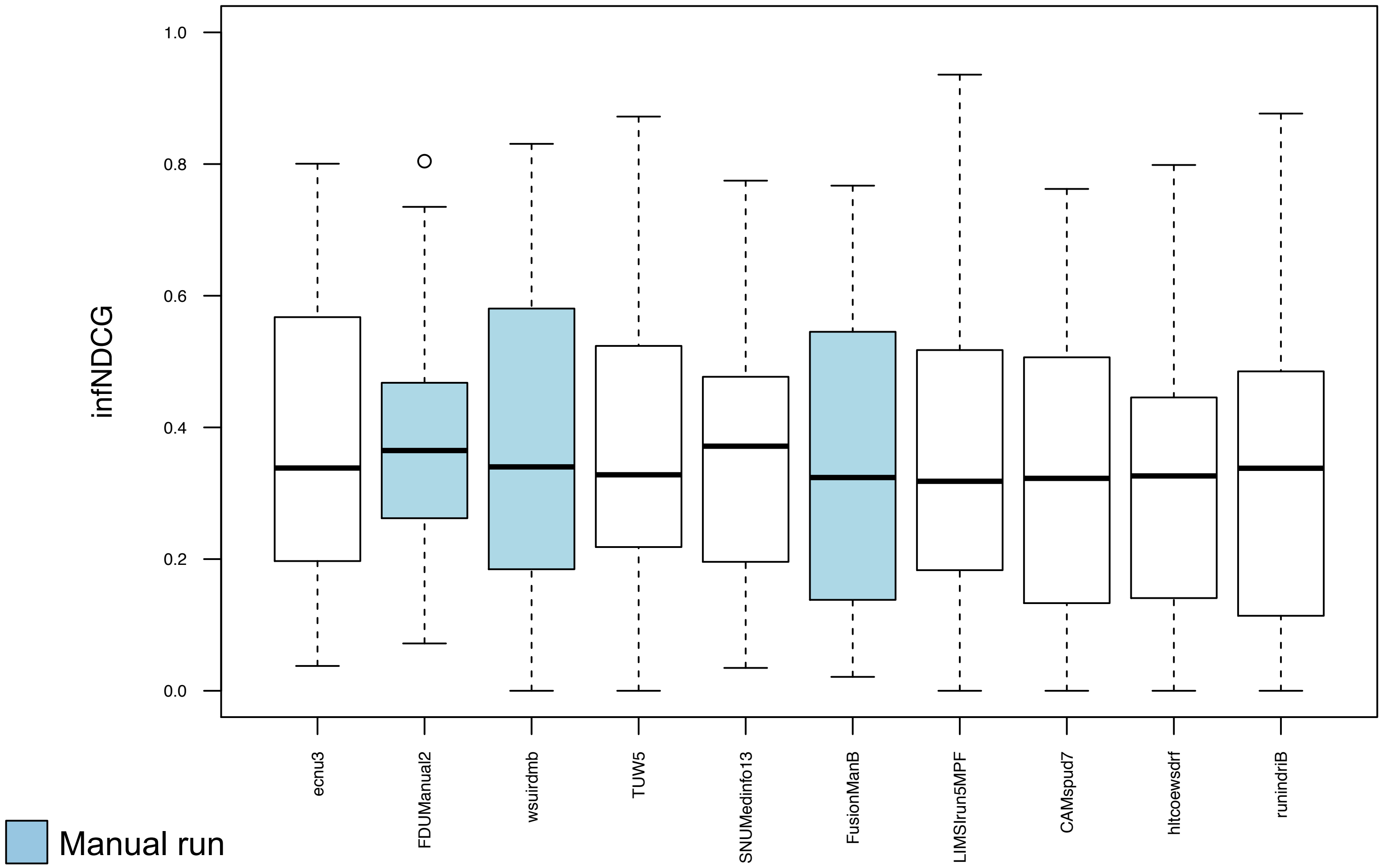
<topic number=”23" type="**treatment**">
Description: An 18-year-old male returning from a recent vacation in Asia presents to the ER with a sudden onset of high fever, chills, facial flushing, pistaxis, and severe headache and joint pain. His complete blood count reveals leukopenia, increased hematocrit concentration and thrombocytopenia.
Summary: An 18 yo male returned from Asia a week ago. He presents with high fever, severe headache, and joint pain. His blood analysis reveals leukopenia, increased hematocrit and thrombocytopenia.
Diagnosis: Dengue

CDS Task A Results


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

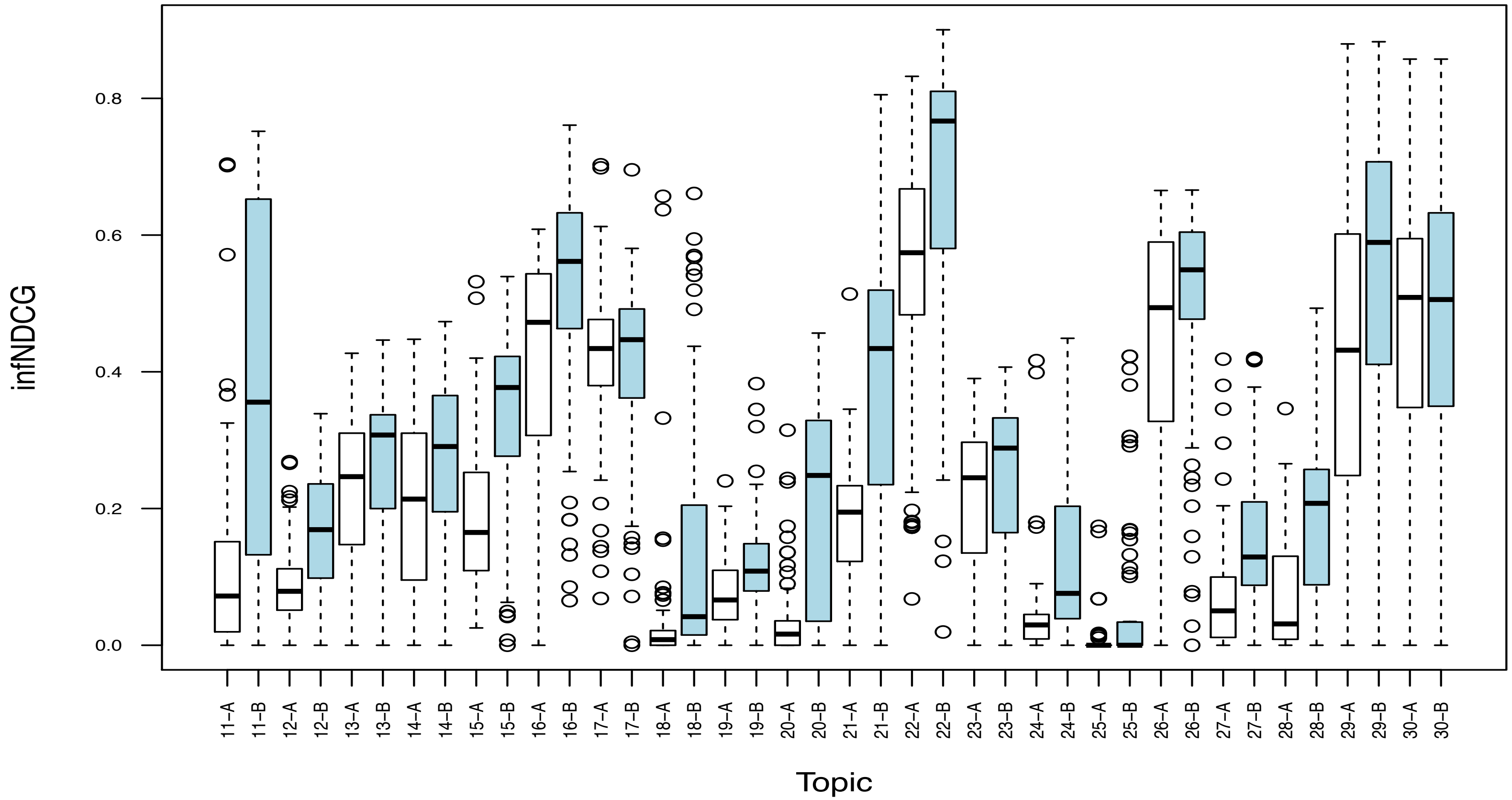
CDS Task B Results

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



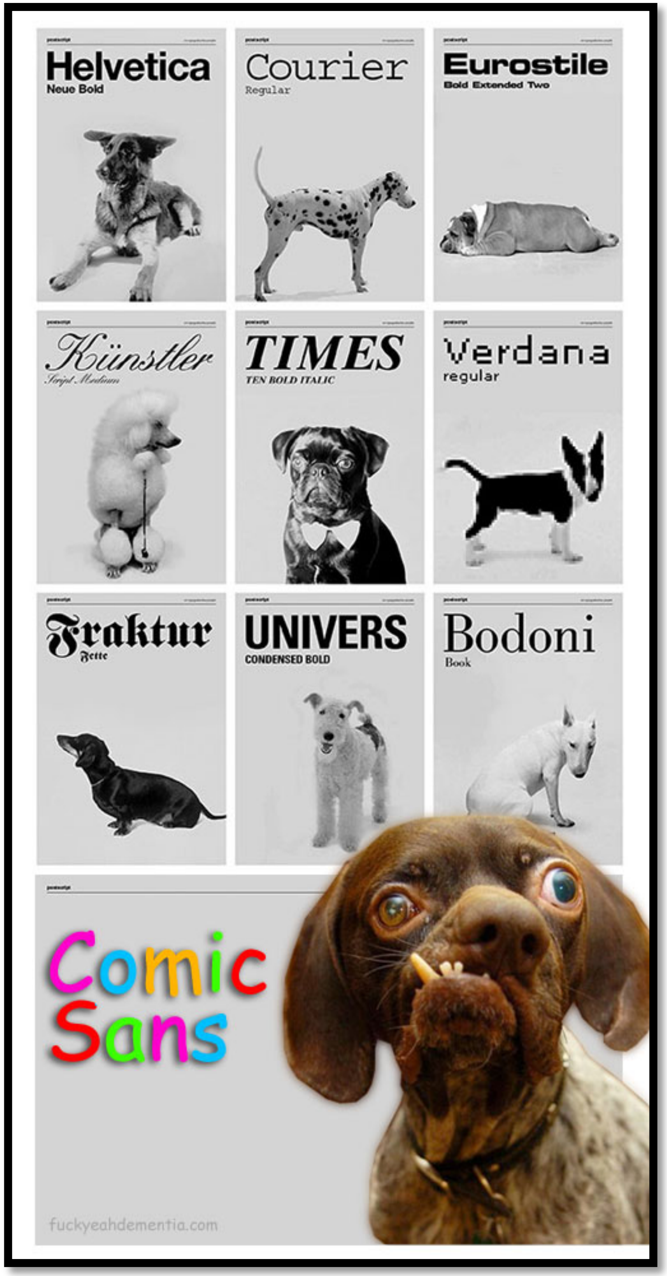
Does the Diagnosis Help?

Distribution of Run infNDCG Scores per Topic, A vs. B



TREC 2016

- Tracks
  - CDS, Contextual Suggestion, Dynamic Domain, Live QA, Tasks, and Total Recall tracks continuing
  - new tracks: Real-time Summarization and Open Search
- TREC 2016 track planning sessions
  - 1.5 hours per track tomorrow (four-way parallel)
  - track coordinators attending 2015
  - you can help shape task(s); make your opinions known



www.geekinheels.com/2001/05/13