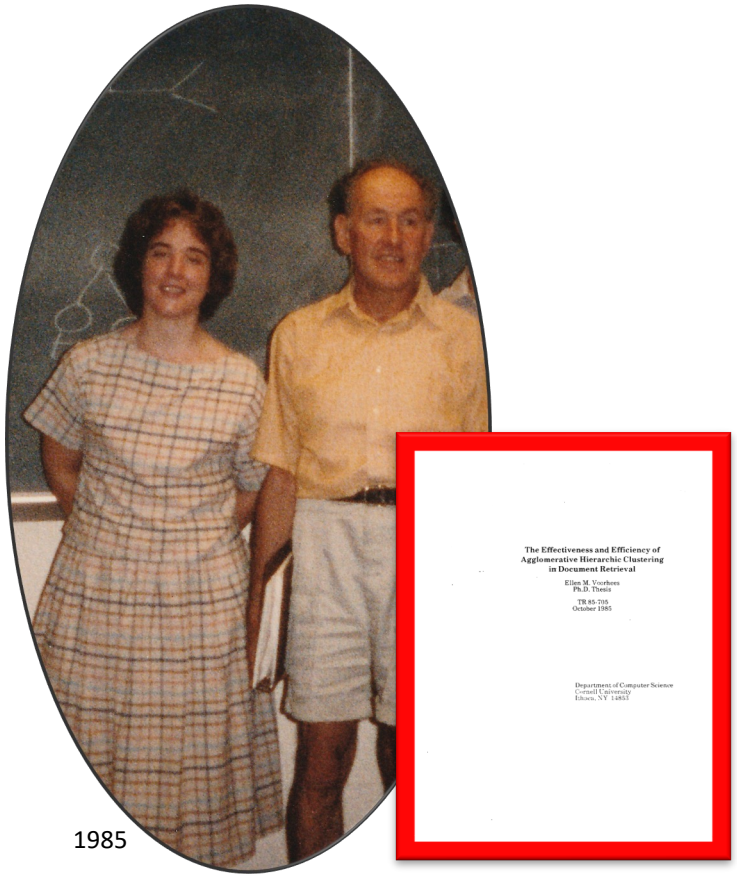
What Gets Measured Gets Done

A Journey of Language Tasks Evaluation

Ellen Voorhees

Image: arielrobin/Pixabay

SIGIR 2024 Washington, D.C.



1985

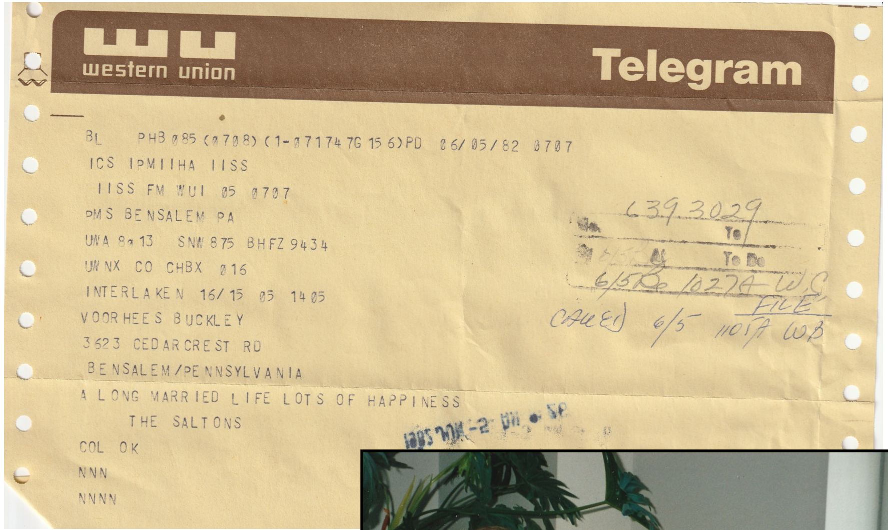
- Particularly honored to be first Salton student to win this award
- Thesis finished in 1985
  - "The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval"
  - main conclusion: clustering not worth it (as compared to inverted index)
  - very definitely not the conclusion Salton was expecting, but he was on sabbatical

- Also married to a Salton student
- SMART group
  - Clem Yu, Harry Wu, Ed Fox…
  - James Allan, Amit Singhal, Mandar Mitra…
- One co-authored publication:

  Salton,Fox,Buckley,Voorhees

  Boolean Query Formulation with Relevance Feedback

  Cornell CS TR 83-539


western union Telegram

BL PHB Ø85 (Ø7Ø8) (1-Ø7174 7G 15 6)PD Ø6/Ø5/82 Ø7Ø7
ICS IPMIIHA IISS
IISS FM WUI Ø5 Ø7Ø7
PMS BENSALEM PA
UWA 8Ø13 SNW 875 BHFZ 9434
UWNX CO CHBX Ø16
INTERLAKEN 16/15 Ø5 14Ø5
VOORHEES BUCKLEY
3623 CEDARCREST RD
BENSALEM/PENNSYLVANIA
A LONG MARRIED LIFE LOTS OF HAPPINESS
THE SALTONS
COL OK
NNN
NNNN

639 3029
CALLED 6/5 1101A WB
FILE


1995

Home Community

image: SIGIR 40th Anniversary photo album

SIGIR 2002 Banquet

No duty is more urgent than that of returning thanks.

-- James Allen (British author)

Image: press 👍 and ⭐/Pixabay

- Award selection committee
- TREC
  - funders
  - participants
  - track coordinators
  - program committee
- Colleagues
  - co-authors
  - NIST
    - esp. Donna Harman, Ian Soboroff

IR Evaluation Timeline

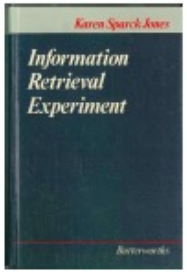
1959 International Conference on Scientific Information

1966 the "Cranfield tests"

1971 Salton publishes collection of SMART papers

1975 Spärck Jones and van Rijsbergen publish report to the British Library on building an 'ideal' information retrieval test collection

1981 Spärck Jones publishes collection of papers on IR tests

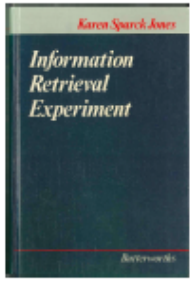


It is arguable that our current understanding of information processing is like that of sixteenth century herbalists: it embodies some observation and insight, but lacks detailed analysis and supporting theory….

The general assumption tends to be that if you know what you want to evaluate, with given evaluation criteria, the appropriate experiment is obvious. Experience shows that this is not the case, because the characteristics of retrieval systems are so difficult to determine and their implication for experiment so difficult to identify.

—KSJ, Introduction

IR Evaluation Timeline



first TREC workshop

1981

lack of comparable results impedes research; lack of realism impedes technology transfer

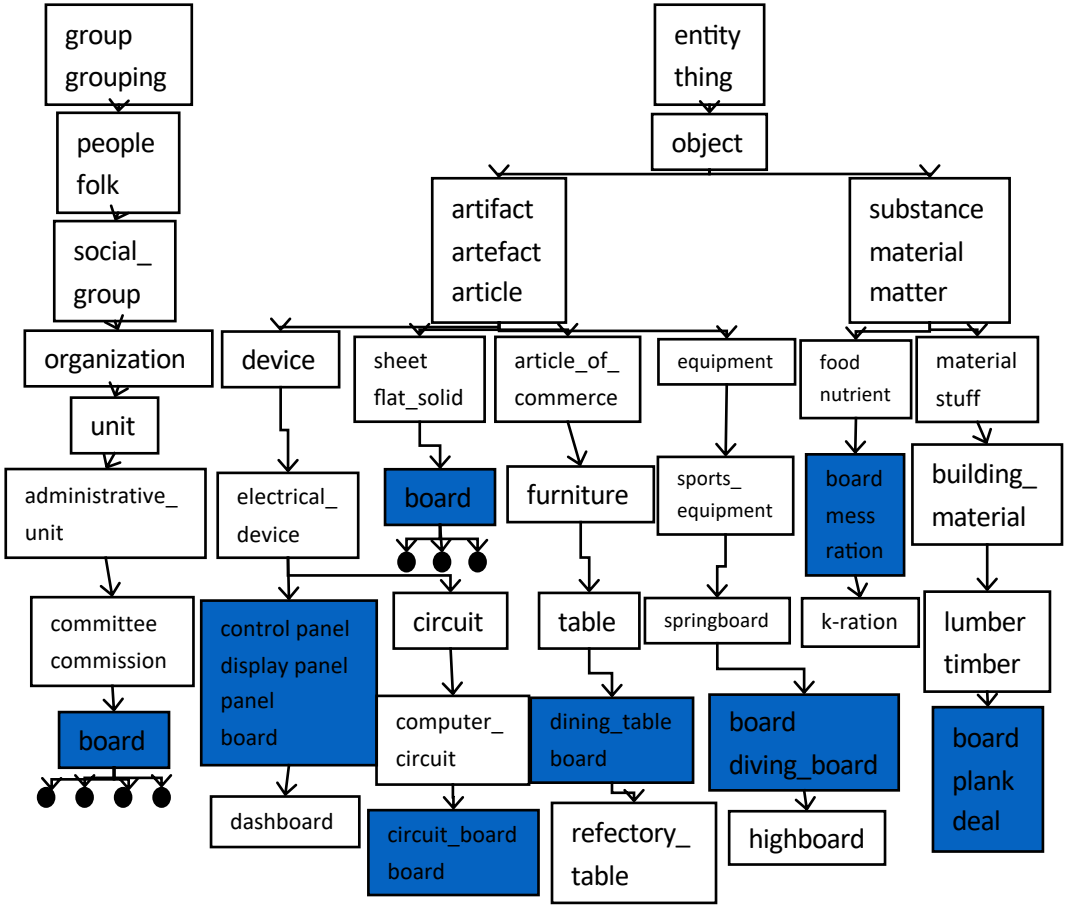
1992

TREC Goals

- ✓ to encourage research in information retrieval based on large test collections;
- • to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- • to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- ✓ to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

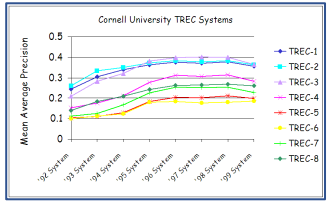
Siemens Decade: 1986--1996

- Use Wordnet to improve retrieval effectiveness?
  - sense resolution? [SIGIR 93]
    - sense-resolved documents didn't match queries
    - IS-A hierarchy doesn't contain enough info to resolve fine distinctions
    - general retrieval similarity functions implicitly do resolution
  - lexical relations provide query expansion terms? [SIGIR 94]
- Software-agent-based distributed retrieval [SIGIR 95]



TREC

Improve the state of the art



The TREC data revitalized research on information retrieval. Having a standard, widely available, and carefully constructed set of data laid the groundwork for further innovation in the field. The yearly TREC conference fostered collaboration, innovation, and a measured dose of competition (and bragging rights) that led to better information retrieval.

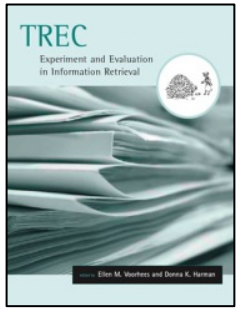
Hal Varian
Google Chief Economist
March 4, 2008

Solidify a research community

This project [the TREC Legal track] can be expected to identify both cost effective and reliable search and information retrieval methodologies and best practice recommendations, which, if adhered to, certainly would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes.

Magistrate Judge Paul Grimm
Victor Stanley v. Creative Pipe

Establish research methodology



TREC is an annual benchmarking exercise that has become a de facto standard in Information Retrieval evaluation.

Stephen Robertson
Microsoft
SIGIR 2007

Facilitate technology transfer

TREC has proven to be a valuable forum in which IBM Research has contributed to an improved understanding of search, while at the same time the insights obtained by participating in TREC have helped to improve IBM's products and services.

Alan Marwick, et al.
IBM chapter of the TREC book
2005

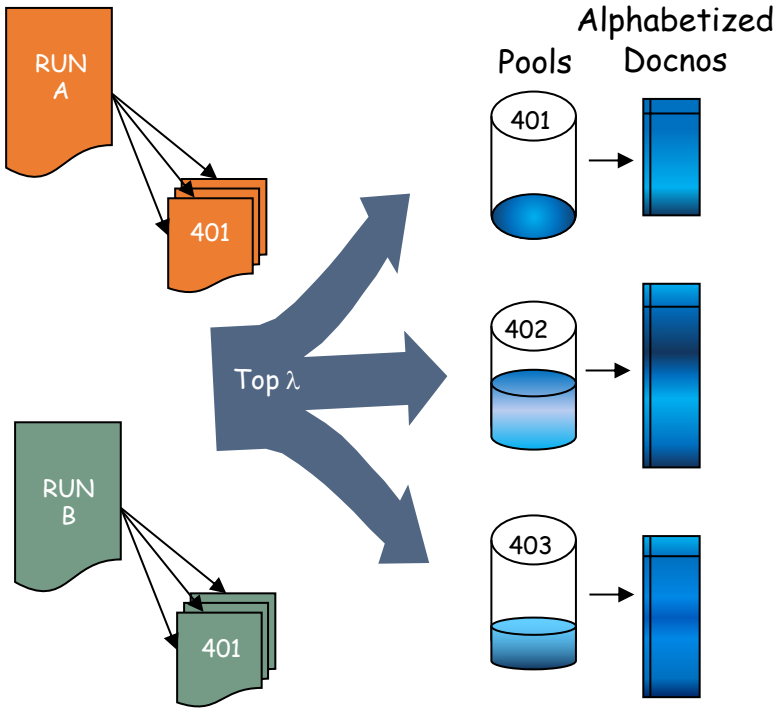
Amortize the costs of infrastructure

In other words, for every $1 NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers…These responses suggest that the benefits of TREC to both private and academic organizations go well beyond those quantified by this study's economic benefits.

RTI International
Economic Impact Assessment of NIST's TREC Program
December 2010

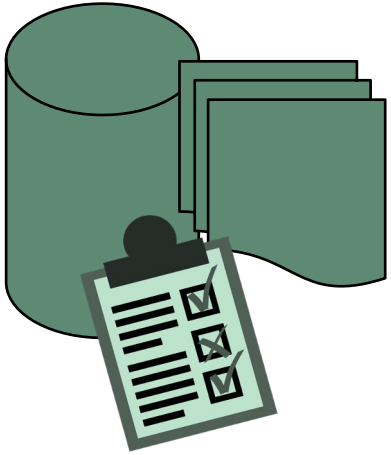
Pooling

- For sufficiently large λ and diverse engines, depth-λ pools produce "essentially complete" judgments
- Unjudged documents are assumed to be not relevant when computing evaluation measures
- Resulting test collections general-purpose and reusable
  1) general-purpose: supports a wide variety of measures and micro tasks
  2) reusable: fair to arbitrary systems, especially those not used in collection construction



K. Spärck Jones and C.J. van Rijsbergen, **Report on the Need for and Provision of an Information Retrieval Test Collection**. British Library Research and Development Report 5266. Computer Laboratory, University of Cambridge. 1975.

Validating Cranfield in the Era of Large(r) Collections



Is the test collection methodology a reliable laboratory tool?

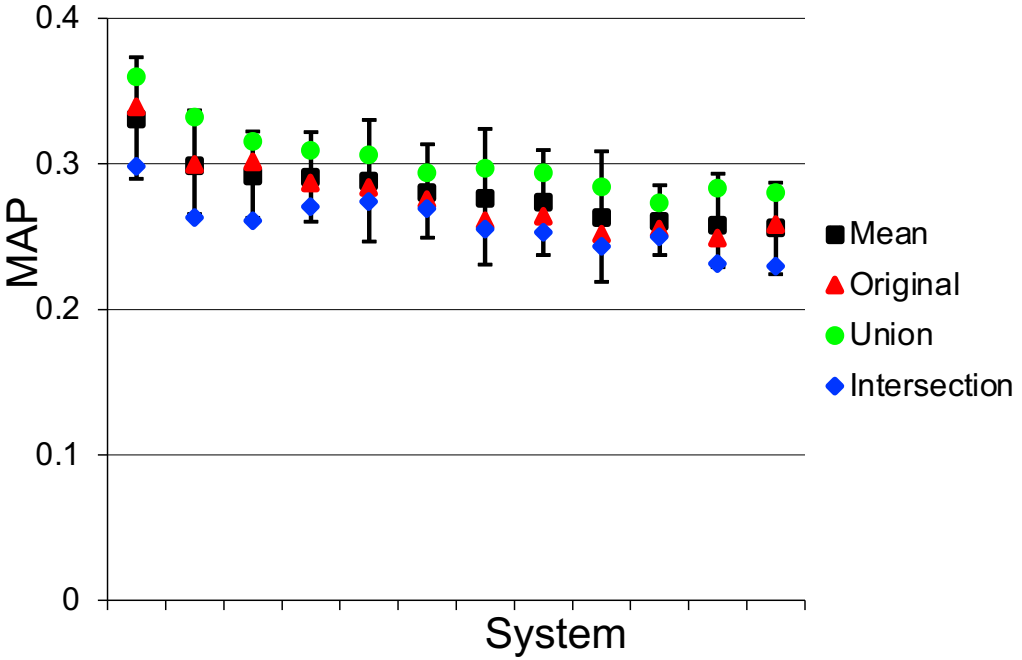
DIFFERENT JUDGES
Robust to different opinions of relevance

INCOMPLETENESS
Unbiased for arbitrary runs

SENSITIVITY
Able to detect (truly) different systems

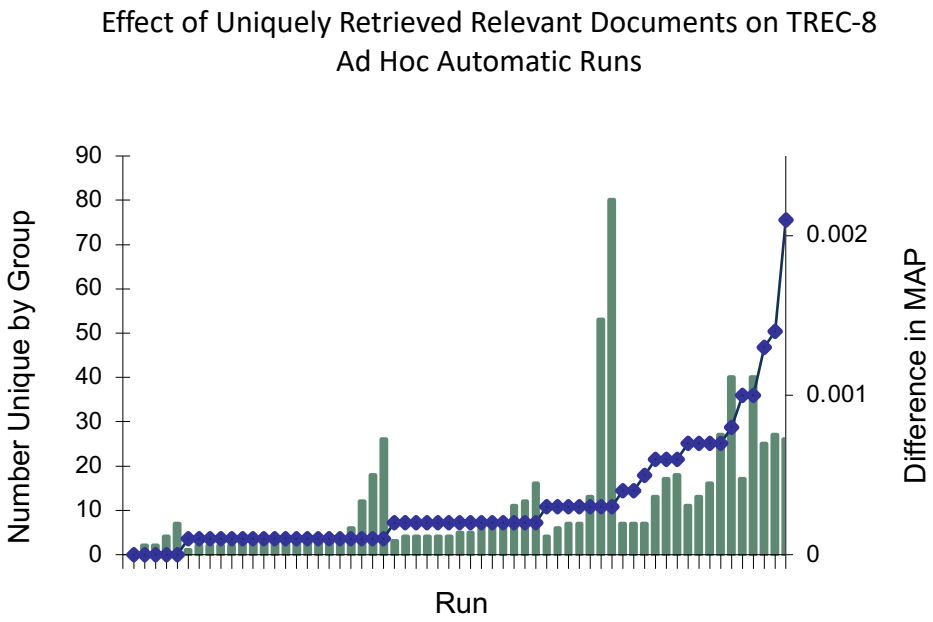
Relevance Judges (and Users!) Disagree

- Experiment
  - three independent sets of judgments for each of 48 TREC-4 topics
  - rank TREC-4 runs when evaluated using different combinations of judgments
- Results
  - judgments do differ
  - comparative results stable
  - true across query sets, measures, kinds of assessors
    - but different grades of relevance are actually different [SIGIR 2001]



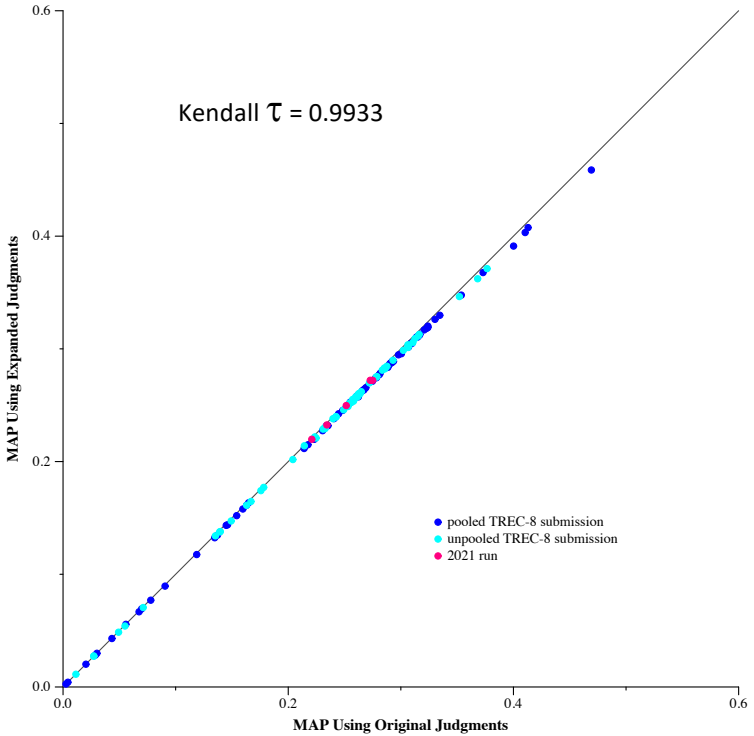
Ellen M. Voorhees. *Variations in relevance judgments and the measurement of retrieval effectiveness*. **Information Processing & Management**, Volume 36, Issue 5, 2000. pp. 697-716.

Reusability



- Zobel's study [SIGIR 98]
  - TREC judgments not complete
  - but ad hoc collections *reusable* (fair to runs not pooled)
  - introduced initial version of leave out uniques (LOU) test
- Modified (more stringent) LOU
  - confirmed results for subsequent ad hoc collections
  - demonstrated importance of diverse, high-quality runs for building collections by pooling

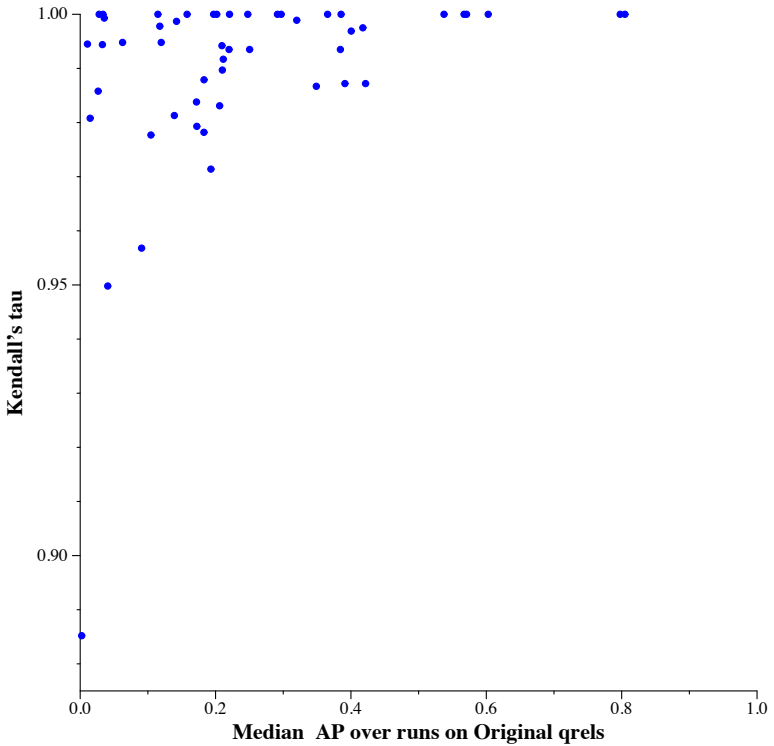
Reusability of TREC-8 Ad Hoc Collection



Voorhees, E.M., Soboroff, I., & Lin, J.J. (2022). Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? *ArXiv, abs/2201.11086*.

- TREC-8 ad hoc (circa 1999)
  - (mostly) newswire collection with approx. 525K documents and 50 test `topics'
  - pooled 71 TREC-8 submissions to depth 100 resulting in 86,830 judgments
- Five new 2021 runs
  - two Anserini BM25 baselines
  - three transformer-based runs
- Pooled 2021 runs plus previously unjudged TREC-8 runs to depth 50
  - 3,842 new judgments in pools ranging from 9—359 documents over 50 queries
  - 158 newly identified relevant documents
  - maximum new relevant in single run: 23

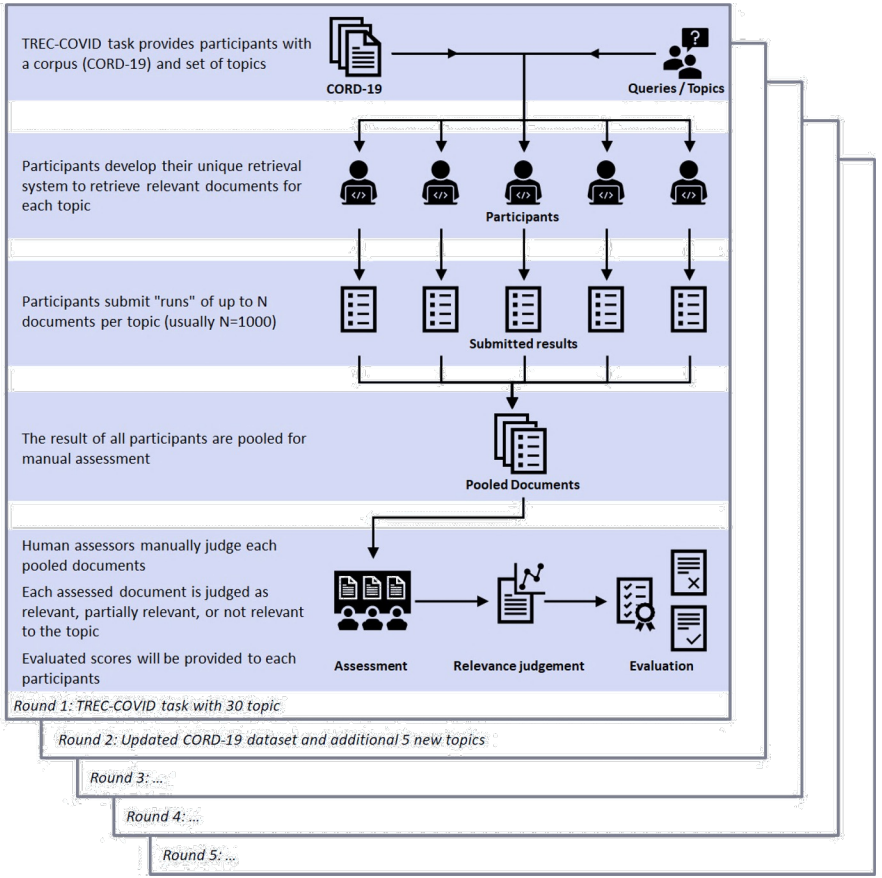
Reusability of TREC-8 Collection



- Even individual topic τ's are stable
  - smallest is 0.8852, and that was caused by many tied scores magnifying the apparent difference
- But… what about some even newer, fancier system?
  - can't conclusively prove it is unaffected unless all documents judged
  - but incredibly unlikely to be significantly unfairly scored
  - to be scored unfairly, system needs to both find sufficiently many new relevant AND rank those new relevant before known relevants

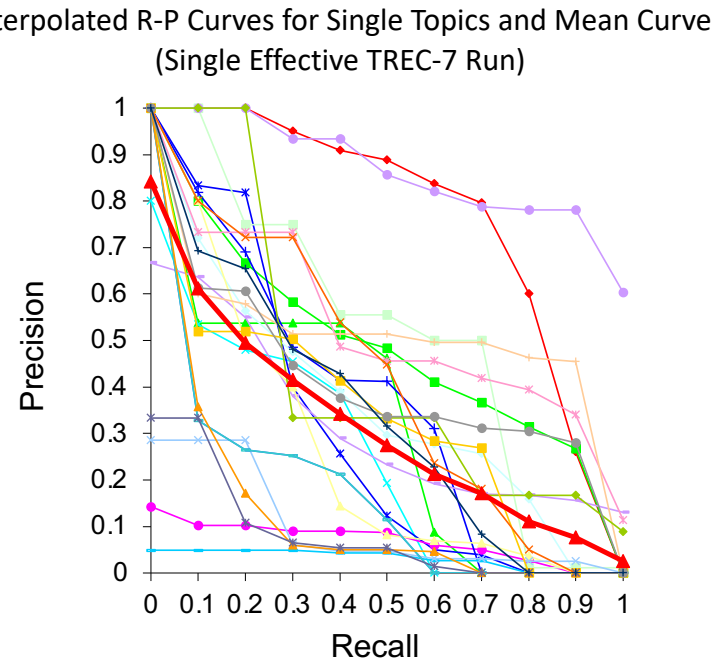
TREC-COVID

- TREC-8: deep pooling over effective runs
  - not particularly actionable, but does produce excellent collections
- TREC-COVID Complete [SIGIR 2021]
  - approximately 69k judgments built from multiple rounds of feedback runs
  - an excellent collection by all diagnostics tests we have
  - pools in individual rounds very shallow, but large and diverse run set
  - almost 1% (enormous) of document collection judged for some topics



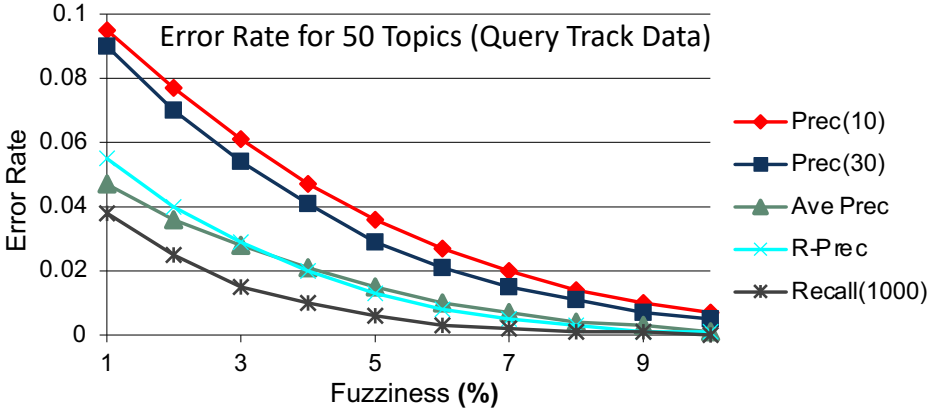
Sensitivity

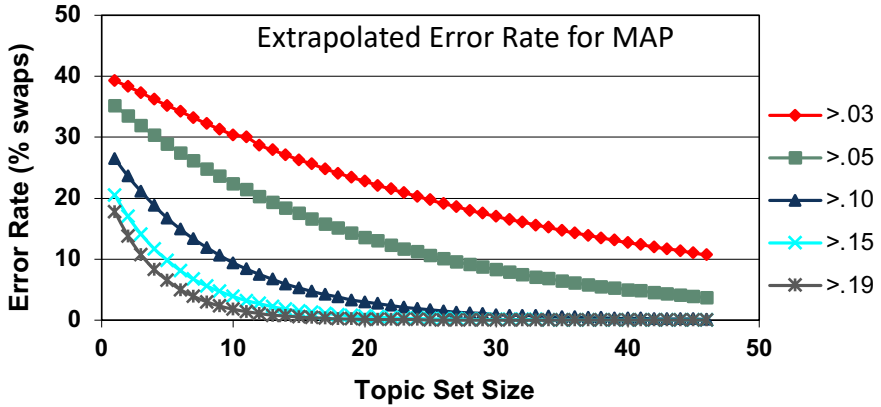
- Variance is rampant
  - topic effect (“user”) is bigger than system effect even given Cranfield’s stark abstraction of search*
  - i.e., what you ask is more important than what system you use
- IR evaluation attempts to detect relatively small difference that can be attributed to the system



* Banks D, Over P, Zhang NF. *Blind men and elephants: Six approaches to TREC data*. **Information Retrieval**. 1999 Jan 1;1(1–2):7–34.

Sensitivity





- Inherent stability of different measures [SIGIR 2000]
  - MAP is stable, informative, less intuitive
  - P(10) is intuitive, relatively unstable, averages poorly
- Empirical relation among topic set size, Δ of measure scores, and error rate [SIGIR 2002]
  - error rate decreases as number of topics increases
  - error rate decreases with larger Δ, but then power is reduced
  - 95% confidence interval requires larger Δ than being seen in the literature

Extending Cranfield

Human-assisted

- Interactive track
- HARD track ("personalization")
- Dynamic Domain

Increased variability means decreased power and generalization, and higher costs

Diversity

- Novelty track
- Diversity tasks in Web track

Measures focus attention on desired behavior, but categories need to be pre-defined and collections less general

Misinformation

- Health Misinformation track

Measures encourage ranking on-topic but incorrect information lower than off-topic. Appeals to central authority for correctness

Fairness

- Fair Ranking track

Explicit formulation of search with multiple stakeholders: producers and consumers. Measures focus on providing producers appropriate (relative to relevance) amount of exposure on average.

TREC QA track through the years

TREC 1999

- question answering track begins; common task for IR and IE
- passage retrieval for specially constructed factoid questions

TREC 2000

- repeat initial task but with natural questions
- study effect of question paraphrases

TREC 2001

- systems must decide whether answer exists in doc set
- new list task

TREC 2002

- factoids require exact answer

TREC 2003

- AQUAINT definition task added
- main task is factoid, list, and definition combination

TREC 2004

- series task (as in NTCIR) begins;
- also add context & evaluation stability

TREC 2005

- series task with events as targets;
- “per-series” scoring decreases variability

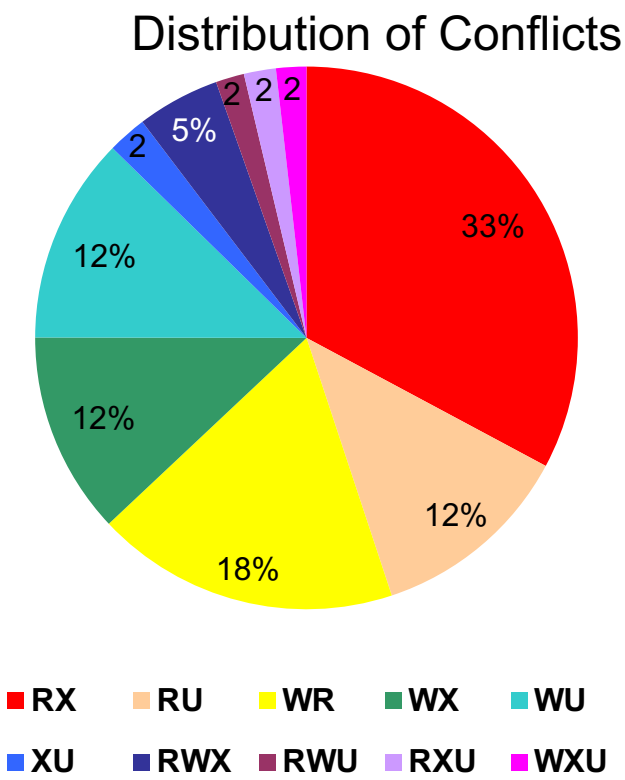
TREC 2006

- series task
- require answers w/ respect to time frame of series
- evaluate via pyramids

TREC 2007

- series using blog documents: malformed language and need for answer validation

Factoids: Assessors Disagree, But System Rankings Stable



- 50% of responses where at least one judgment was not Wrong had disagreements
- Of those, 33% involved disagreements between Right and ineXact
  - well-known granularity issue now reflected here
- For dates and quantities, disagreement among Wrong and ineXact
- Kendall τ scores between system rankings > 0.9

Nugget-based Evaluation of Definition Questions

- Have assessor create list of concepts that definition should contain
  - indicate essential concepts
  - okay concepts
- Mark concepts in system responses
  - mark a concept at most once
  - individual item may have multiple, one, or no concepts

What is a golden parachute?

Assessor nuggets

1. Agreement between companies and top executives
2. Provides remuneration to executives who lose jobs
3. Remuneration is usually very generous
4. Encourages executives not to resist takeover beneficial to shareholders
5. Incentive for executives to join companies
6. Arrangement for which IRS can impose excise tax

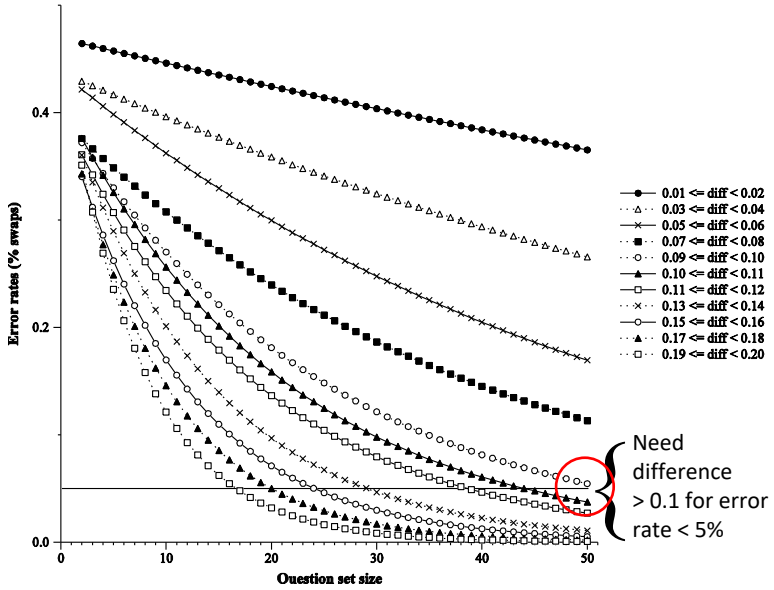
Judged system response

2,3	a	The arrangement, which includes lucrative stock options, a hefty salary, and a “golden parachute” if Gifford is fired
1	b	Oh, Eaton has a new golden parachute clause in his contract
	c	But some, including many of BofA’s top executives joined the 216 and cashed in their “golden parachute” severance packages
6	d	But if he quits or is dismissed during the 2 years after the merger, he will be paid $24.4 million, with Daimler-Chrysler paying the “golden parachute” tax for him and the taxes on the compensation paid to cover the tax.
4	e	After the takeover, as jobs disappeared and BofA’s stock tumbled, many saw him as a bumbler who sold out his bank, walking away with a golden parachute that gives him $5 million a year for the rest of his life.
	f	The big payment that Eyler received in January was intended as a “golden parachute”

Reliability of Nugget-based Evaluation

- Effect of assessor differences
  - different assessors disagree as to correctness
  - Kendall τ correlation among system rankings when questions judged by different assessors of 0.848
- Sample of questions
  - different systems do relatively differently on different questions
  - particular sample of questions can skew results
  - more questions lead to more stable results

TREC 2003 QA Track Definitions Task



QA Track Results

Solidify a community

- enormous growth in QA community
- world-wide interest (e.g., QA tasks in NTCIR, CLEF)

Establish the research methodology

- showed that even "facts" are context-sensitive
- significant steps toward evaluation of complex answers

Facilitate technology transfer

- common architecture for factoid questions

Document the state-of-the-art

- task for which NLP techniques shows real benefit
- rough boundary when IR techniques insufficient
- demonstrated on diverse genres

Amortize the costs of infrastructure

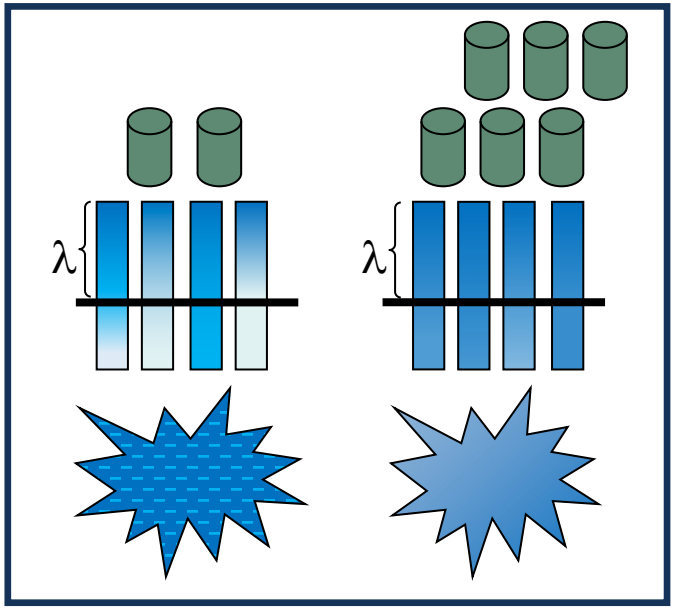
- answer patterns provide partial solution (use with care)

QA Track Evaluation

- Tasks don't produce truly reusable infrastructure
  - can't use judgments to evaluate completely new run
  - evaluations therefore relatively expensive since costs can't be amortized
- Conflict in balancing realism, control
  - e.g., selection of questions across TRECs

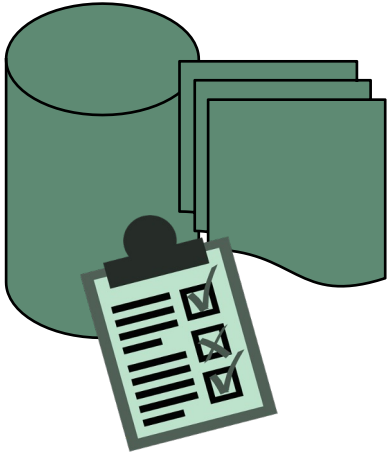
Pooling Bias (TREC Robust track, 2005)

- Traditional pooling takes top λ documents
  - intentional bias toward top ranks where relevant are found
  - λ was originally large enough to reach past swell of topic-word relevant
- As document collection grows, a constant cut-off stays within swell
- Pools cannot be proportional to corpus size due to practical constraints
  - sample runs differently to build unbiased pools
  - new evaluation metrics that do not assume complete judgments



C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. *Bias and the limits of pooling for large collections*. **Information Retrieval**, 10(6):491-508, 2007.

Building Retrieval Test Collections



How do we build **general-purpose**, **reusable** test collections at **acceptable cost**?

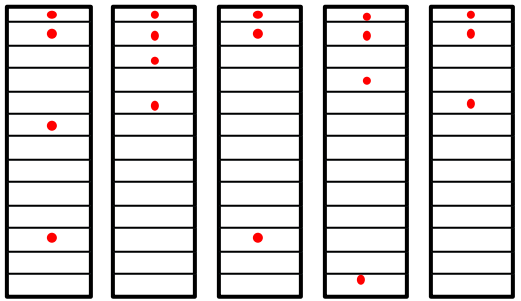
GENERAL PURPOSE
Supports a wide range of measures and search scenarios

REUSABLE
Unbiased for systems not used to build the collection

ACCEPTABLE COST
Cost proportional to number of human relevance judgments needed

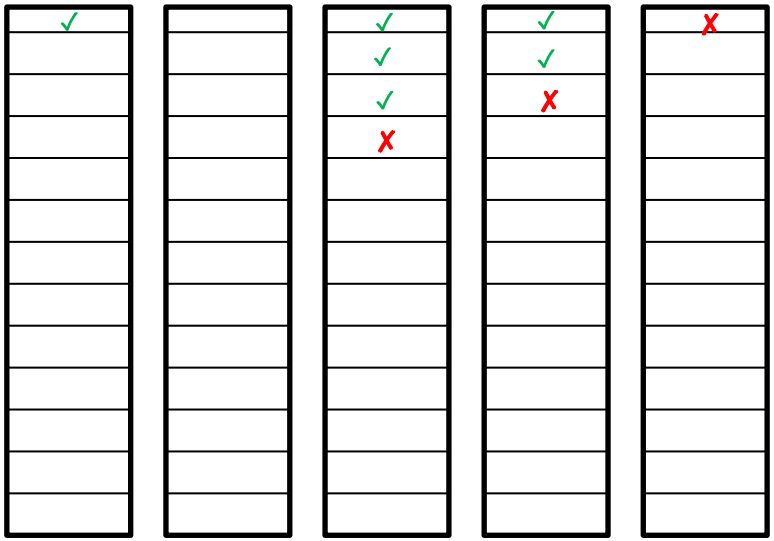
Inferred Measure Sampling

- Stratified sampling where strata are defined by ranks
- Different strata have different probabilities for documents to be selected to be judged
- Given strata and probabilities, estimate AP by inferring which unjudged docs are likely to be relevant
- ✗ Quality of estimate varies widely depending on exact sampling strategy
- Fair, but less general-purpose



E. Yilmaz, E. Kanoulas, and J. A. Aslam. *A simple and efficient sampling method for estimating AP and NDCG*. **SIGIR 2008**, pp.603—610.

Multi-armed Bandit Sampling



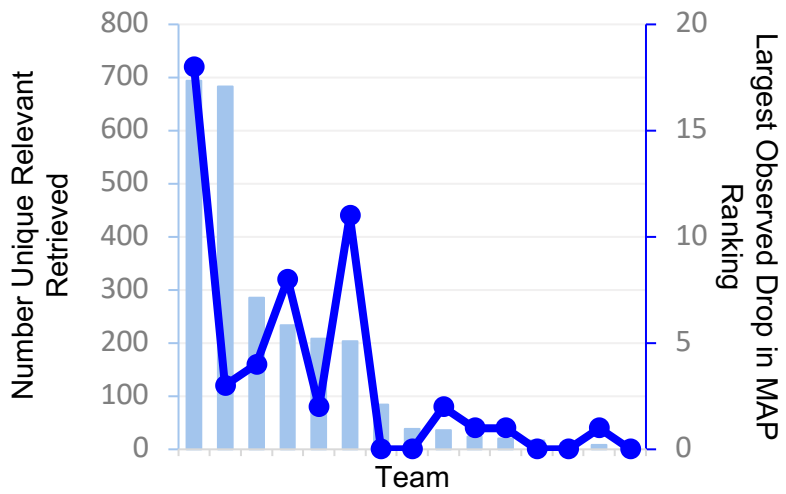
- Bandit techniques trade-off between exploiting known good “arms” and exploring to find better arms. For collection building, each run is an arm, and reward is finding a relevant doc
- Simulations suggest can get similar-quality collections as pooling but with many fewer judgments
- TREC 2017 Common Core track first attempt to build new collection using bandit technique

  bandit selection method:
  2017: MaxMean
  2018: MTF

D. Losada, J. Parapar, A. Barreiro. *Feeling Lucky? Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation*. Proceedings of SAC 2016. pp. 1027-1034.

Collection Quality

- 2017 Common Core collection less *reusable* than hoped (just too few judgments)
- Additional experiments demonstrate greedy bandit methods can be *UNFAIR*



LOU-results for TREC 2017 Common Core collection

	MAP		Precision(10)	
	τ	Drop	τ	Drop
MaxMean	.980	2	.937	11
Inferred	.961	7	.999	1

Fairness test: build collection from judgments on small inferred-sample or on equal number of documents selected by MaxMean bandit approach (average of 300 judgments per topic). Evaluate runs using respective judgment sets and compare run rankings to full collection rankings. Judgment budget is *small enough that R exceeds budget for some topics*.

Example: topic 389 with R=324, 45% of which are uniques; one run has 98 relevant in top 100 ranks, so 1/3 relevant in bandit set came from this single run to the exclusion of other runs.

Bandit Conclusions

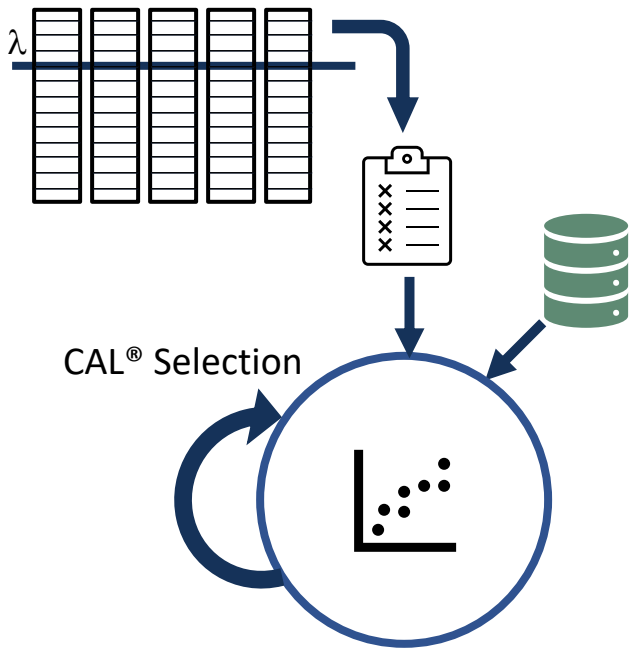
Can be unfair when budget is small relative to (unknown) number of relevant

- must reserve some of budget for quality control, so operative number of judgments is less than B

- Does not provide practical means for coordination among assessors
  - multiple human judges working at different rates and at different times
  - subject to a common overall budget
  - stopping criteria depends on outcome of process

Image: Pascal/flickr

Deep Learning Track

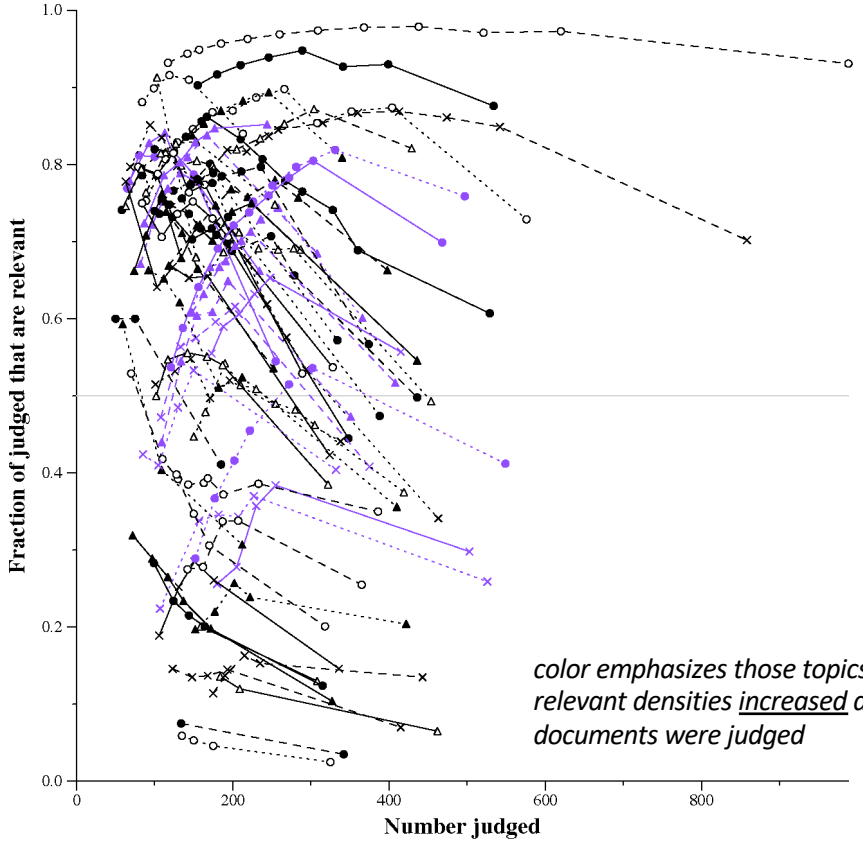


Gordon V. Cormack and Maura R. Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. arXiv:1504.06868 [cs.IR]

- Collections built using shallow pools followed by Continuous Active Learning
  - judge depth-10 pools across submissions
  - given set of relevance judgments, CAL builds model of relevance and orders remaining collection by likelihood of relevance
  - loop on obtaining judgments and running CAL per topic until stopping condition met
    - stopping: few new relevant found or budget exhausted or too many total relevant (so reject)
- Resulted in acceptable collections in 2019 and 2020
  - same process failed to produce acceptable collection in 2021

Judgments

- Four relevance grades
- Two judgment phases
  - track judging in Sept. 2021
    - 13,058 total judgments
    - mean 229.1 [min 75,max 620]
  - supplementary phase in Dec. 2021
    - additional 9255 judgments
    - no CAL; docs selected to support collection experiments
- In track judging, 40/57 topics had relevant densities > 0.5

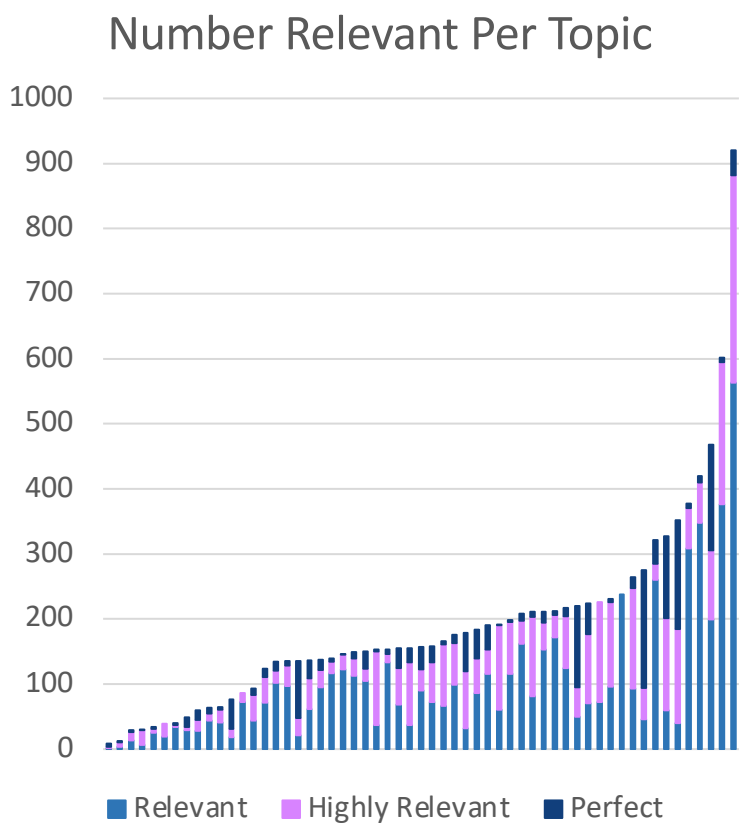


color emphasizes those topics whose relevant densities increased as more documents were judged

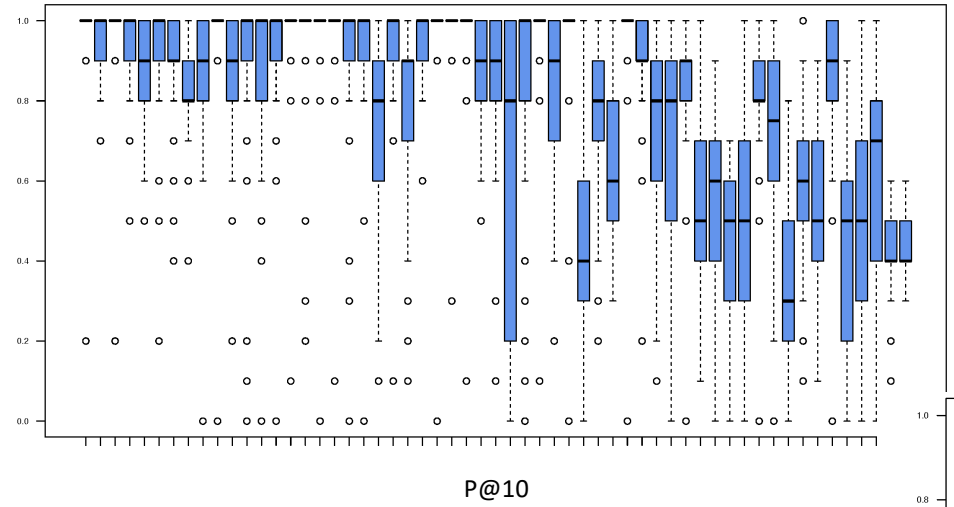
What Happened?

- Corpus size 3.7 times as large in 2021
- Number of relevant increases as corpus size increases, on average*
- These are the “easy to find” relevant documents and there are lots of them
  - collection is not reusable…
  - …but also recall-based measures are unreliable even for track submissions…
  - … and high-precision scores are saturated, so comparisons with them are unstable

* David Hawking and Stephen Robertson. 2003. On Collection Size and Retrieval Effectiveness. **Information Retrieval** 6 (2003), 99–105.



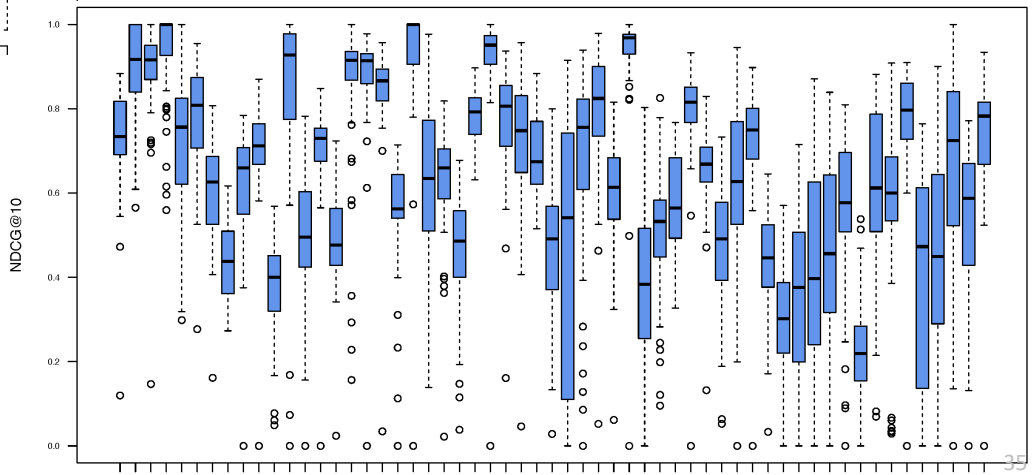
Deep Learning 2021 Scores



P@10

Distribution of Scores across Submissions for Each Topic

NDCG@10



Size is a Perennial Problem

images: OpenClipart-Vectors/Pixabay & Ria Kartika/Pixabay

- Collections are now large enough that “safe” measures are unreliable
  - a collection where it is trivially easy to retrieve k relevant documents makes Prec@k (and related measures) unreliable
- Building a quality collection still depends on size (R)
- Can’t build collection using full judgment budget
  - need more judgments to know quality of collection
  - lack appropriate tools for assessing collection quality
- Still seeking efficient, effective method to coordinate assessing
  - multiple human judges working at different rates and at different times
  - subject to a common overall budget
  - stopping criteria depends on outcome of process

LLM's to the Rescue? [I'm dubious]

- Obtained judgments for TREC-8 ad hoc
  - ~ 80k judgments across 50 topics
  - set of documents TREC assessors judged (depth 100 pools across 71 runs)
- Mean overlap between judgment sets: 0.356
- Kendall's τ of system rankings

  MAP: 0.880

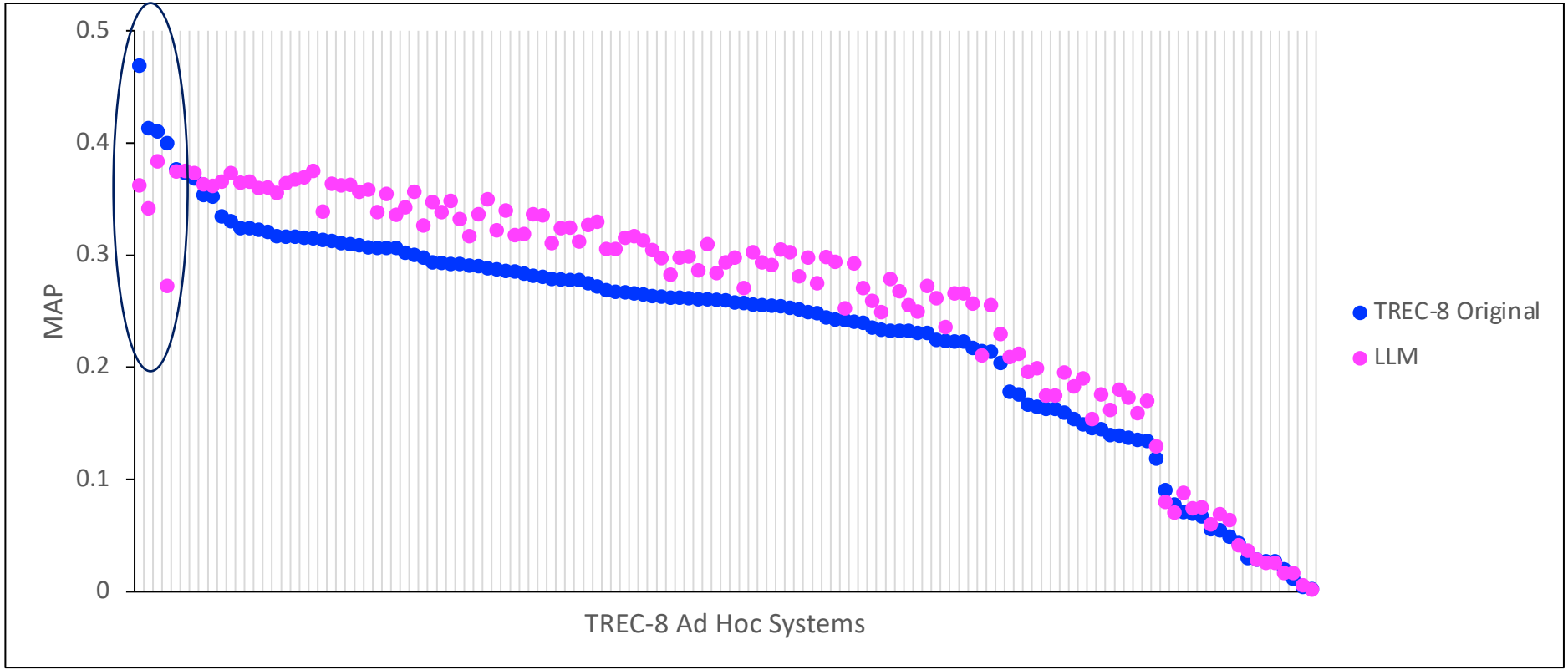
  P(10): 0.883
- BUT huge swings in relative effectiveness of manual runs

  max drop for MAP: 74 (out of 129 runs)

  max drop for P(10): 26

image: Claudia/Pixabay

MAP Scores by Judgment Source



LLM Evaluation

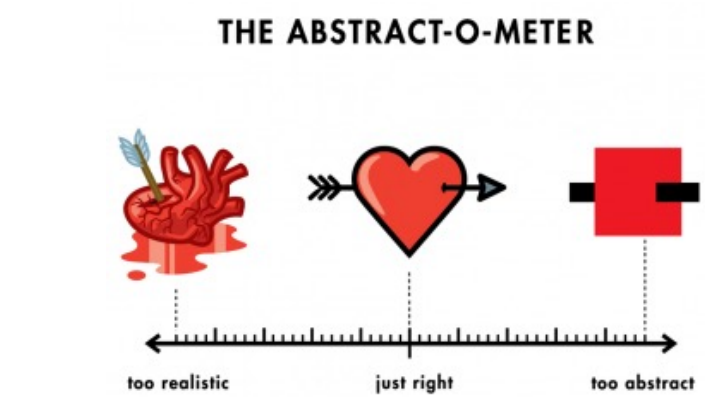


Image: https://computersciencewiki.org/images/thumb/e/e2/Abstract_heart.png/350px-Abstract_heart.png

Appropriate abstraction

- Karen Spärck-Jones' "core competency"
- external validity: measures what was intended to measure
- previous language evals used fluency as proxy, and that is no longer appropriate

Feasible to implement

- accurate, repeatable measures
- resolution of measurements

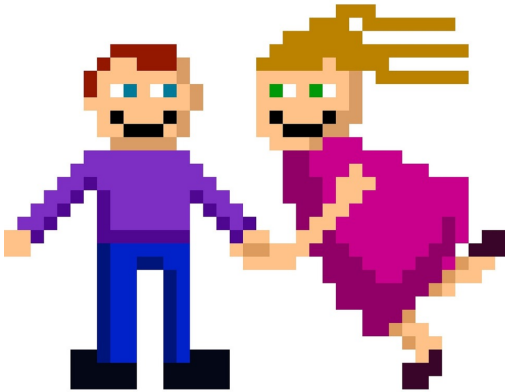


Image: Prawny/Pixabay

What is a Good Evaluation Task?

image: asi24/Pixabay

- Abstraction of real task so variables affecting performance can be controlled…
- …but must capture salient aspects of real task or exercise is pointless
- Metrics must accurately predict relative effectiveness; best if measures are diagnostic.
- Appropriate level of difficulty
- Best if infrastructure is reusable

Test Sets are Measurement Devices

- Any given test set provides a single instance of a measure of interest
- As with any measurement device, each test set has a target property that can be determined to a finite level of resolution
  - make sure test set has credible relation to property of interest
  - make sure measurement limits recognized

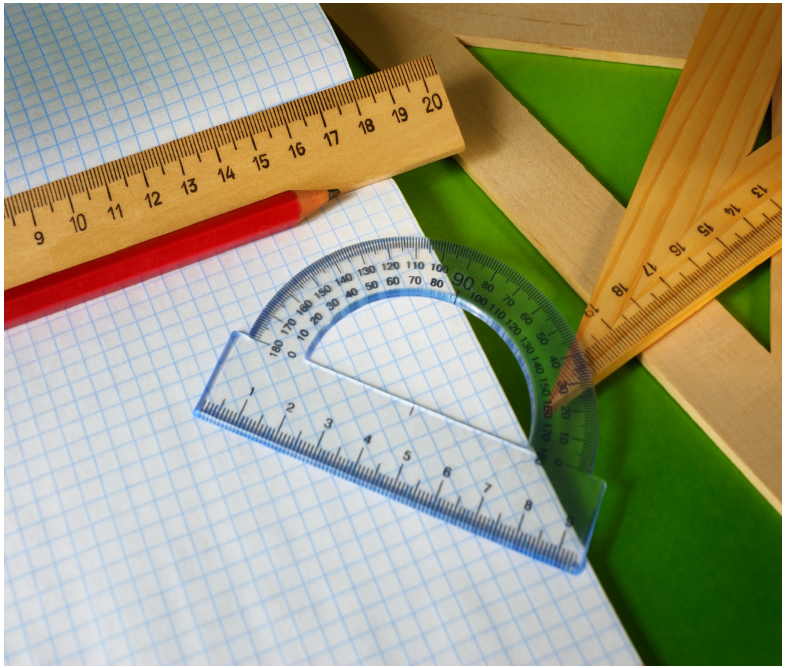


Image: ds_30/pixabay.com

DO NOT INTERPRET MEASUREMENT ERROR AS SUPER-HUMAN EFFECTIVENESS