Cranfield is Dead; Long Live Cranfield

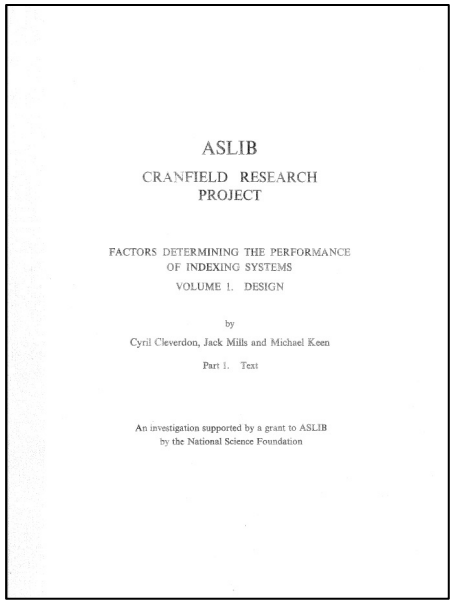
Ellen Voorhees

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

Cranfield Tests (circa mid-1960's)

www.computerhope.com/people/cyril_cleverdon.htm

Cyril Cleverdon



ASLIB

CRANFIELD RESEARCH PROJECT

FACTORS DETERMINING THE PERFORMANCE OF INDEXING SYSTEMS

VOLUME 1. DESIGN

by

Cyril Cleverdon, Jack Mills and Michael Keen

Part 1. Text

An investigation supported by a grant to ASLIB by the National Science Foundation

- Introduced the test collection:
  - set of documents
  - set of information needs
  - relevance judgments that say which docs should be retrieved for each query
- Original focus was on comparing indexing languages
- The use of `relevance' as the basis of an evaluation met with derision
  - vilified in the literature of the time
  - periodic recurrence in the literature ever since

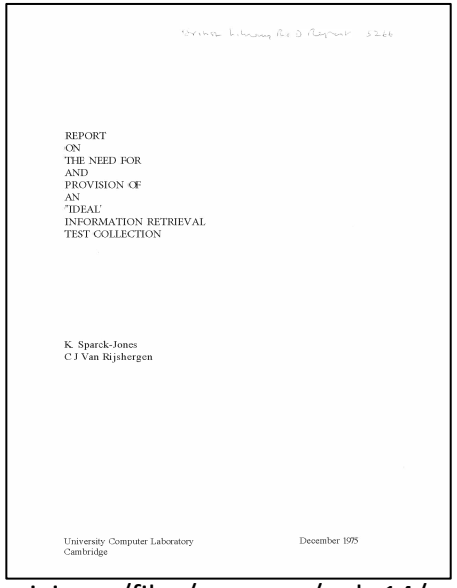
An IDEAL Test Collection, 1975

image: Wikipedia

Karen Sparck Jones

image: www.dcs.gla.ac.uk/~keith/

Keith van Rijsbergen



REPORT
ON
THE NEED FOR
AND
PROVISION OF
AN
"IDEAL"
INFORMATION RETRIEVAL
TEST COLLECTION

K. Sparck-Jones
C J Van Rijshergen

University Computer Laboratory
Cambridge

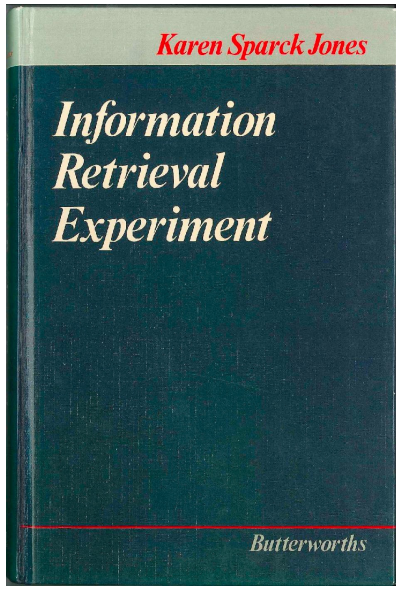
December 1975

sigir.org/files/museum/pub-14/

- Proposal to the British Library to create a purpose-built test collection
- Main features
  - large (30K documents!)
  - 75 requests minimum
  - introduced idea of pooling to get judgments
- Not funded

K. Sparck 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.

IR Experiment, 1981



The volume was dedicated to Cyril Cleverdon, and it was published to reflect the state of IR experimentation 20 years after Cranfield.

image: Wikipedia

Karen Sparck Jones

http://www.staff.city.ac.uk/~sbrp622/

Stephen Robertson

- From the introduction by Spark Jones:

  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.

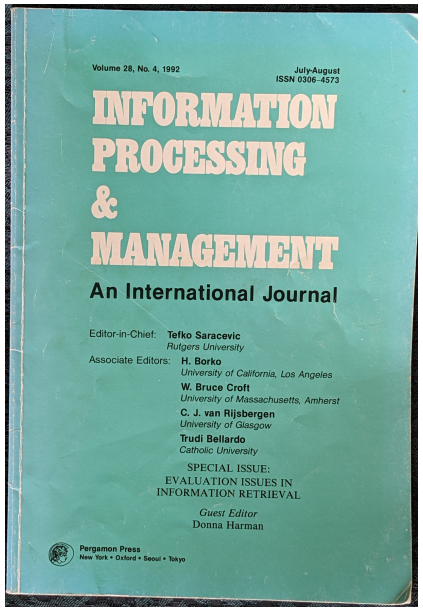
- Steve Robertson argues for 'portable test collections' in his chapter:

  The existence of these collections has had a considerable influence on the direction of research in the field, for the simple reason that some processes (such as automatic indexing from full text) are not possible on these collections as they currently exist. In these circumstances, it is at least arguable that the research community should set up one or more genuinely portable test collections: collections that are designed as general-purpose research tools, rather than taking on that role by accident.

The State of IR Evaluation: Salton, IPM 1992

www.computerhope.com/people/gerard_salton.htm

Gerry Salton



Gerard Salton, The state of retrieval system evaluation, **Information Processing & Management**, Volume 28, Issue 4, 1992, Pages 441-449, ISSN 0306-4573, https://doi.org/10.1016/0306-4573(92)90002-H.

States (& rebuts) the Case Against Cranfield

- Relevance judgments
  - vary too much to be the basis of evaluation
  - topical similarity is not utility
  - static set of judgments cannot reflect user's changing information need
- Recall is unknowable
- Results on test collections are not representative of operational retrieval systems

IR Winter

- Lack of comparable results impeding research
  - research groups reported different measures
  - couldn’t build on one another’s work
- Lack of realism impeding technology transfer
  - test collections too small

image: Vianney Dugrain/Pixabay

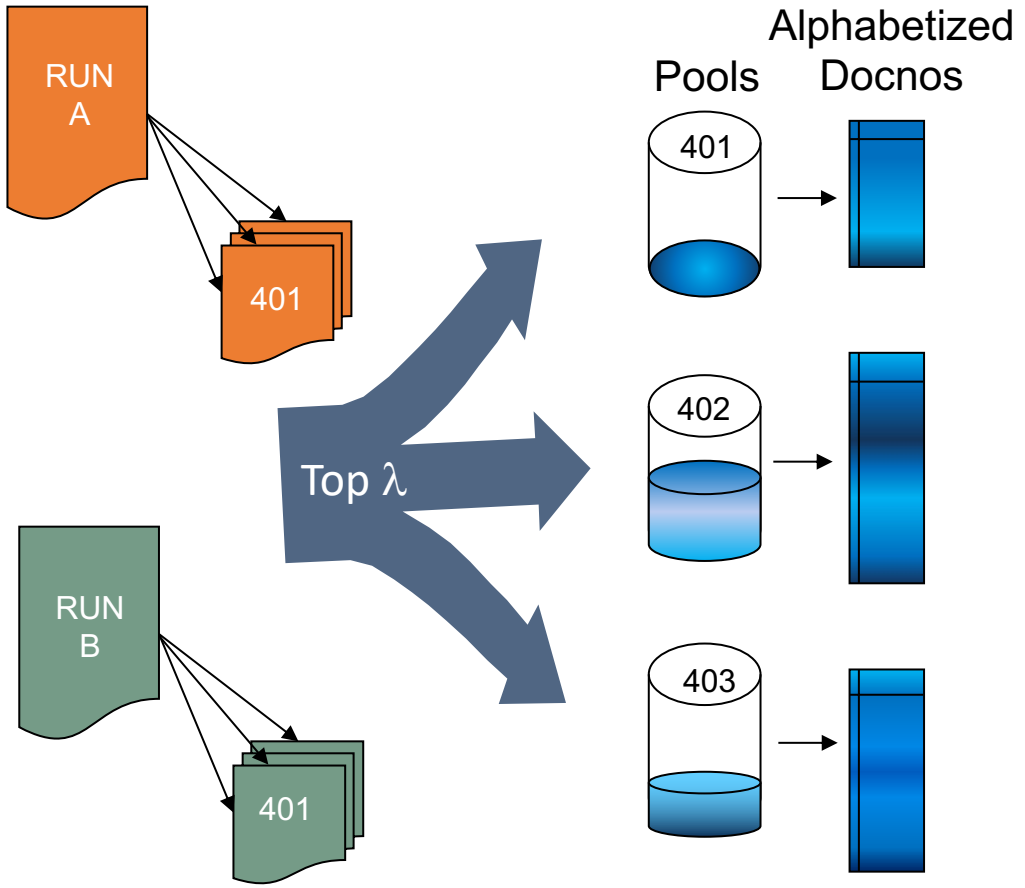
TREC 1992

- 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**;
- 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.

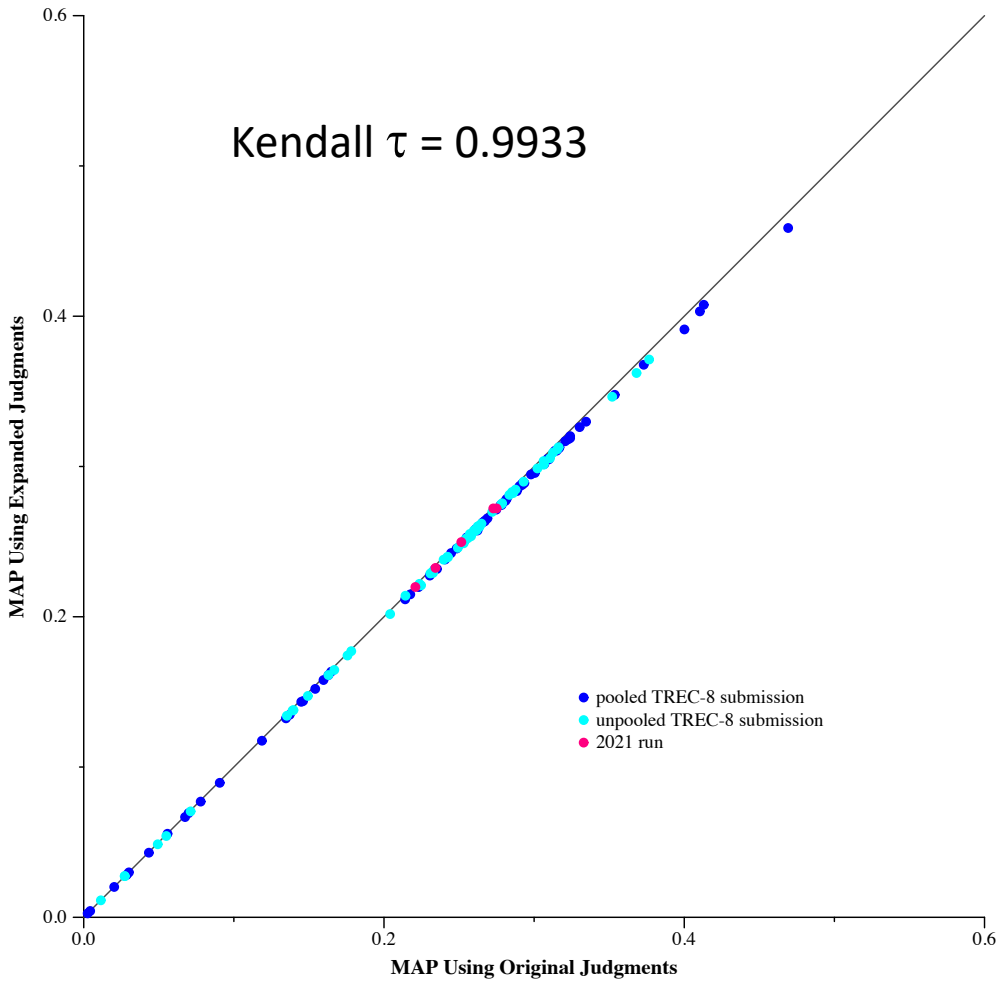
image: trec.nist.gov

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 traditional evaluation measures such as average precision (AP)
- Resulting test collections can be both fair and reusable
  1) fair: no bias against systems used to construct collection
  2) reusable: fair to systems not used in collection construction



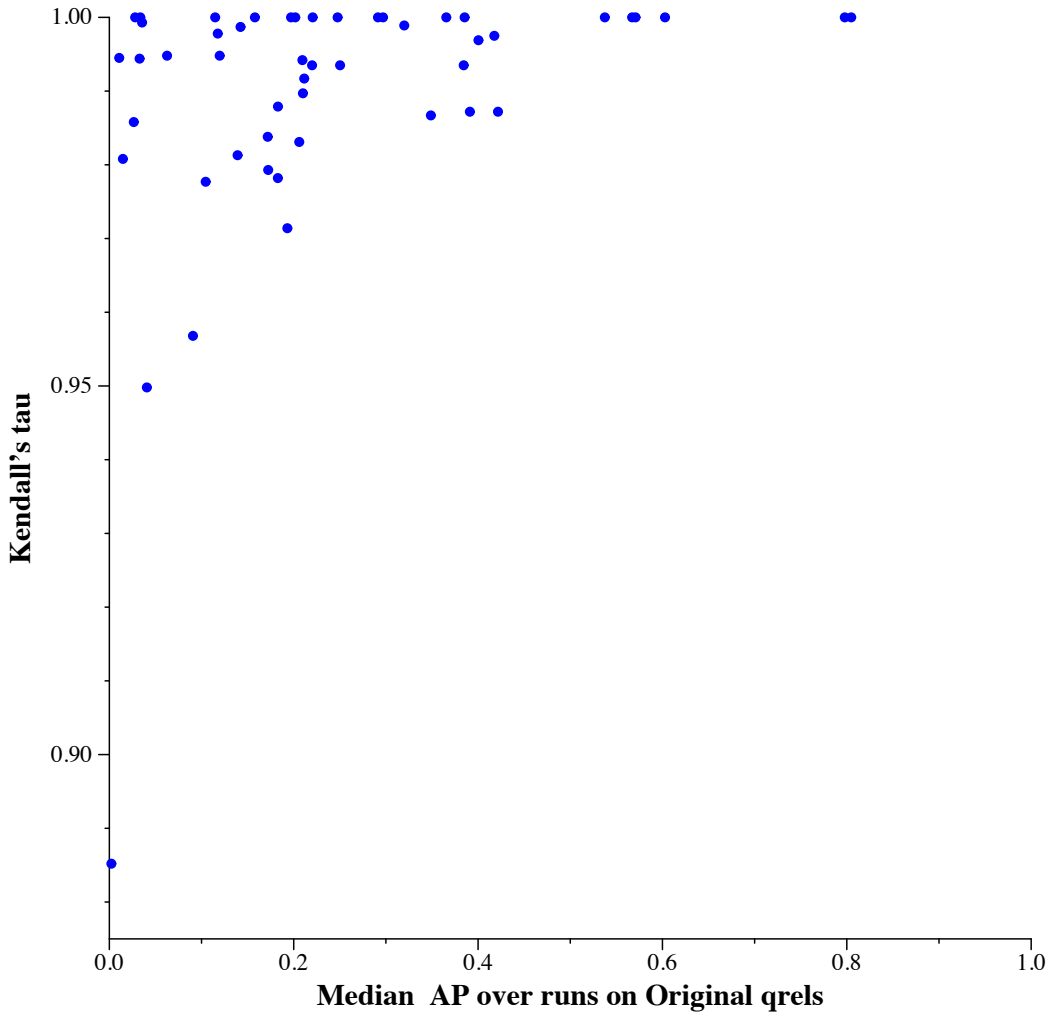
Reusability of TREC-8 Ad Hoc Collection



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

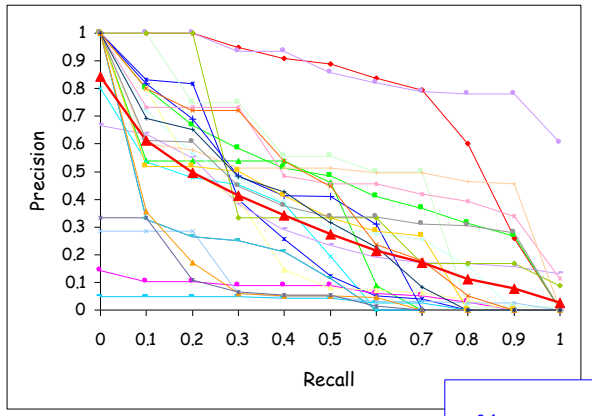
- TREC-8 ad hoc (circa 1999)
  - (mostly) newswire collection with approx. 525K documents and 50 test `topics'
  - pooled 71 TREC-8 submissions to depth 100 resulting in 86,830 judgments
- Five new 2021 runs
  - two Anserini BM25 baselines
  - three transformer-based runs
- Pooled 2021 runs plus previously 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

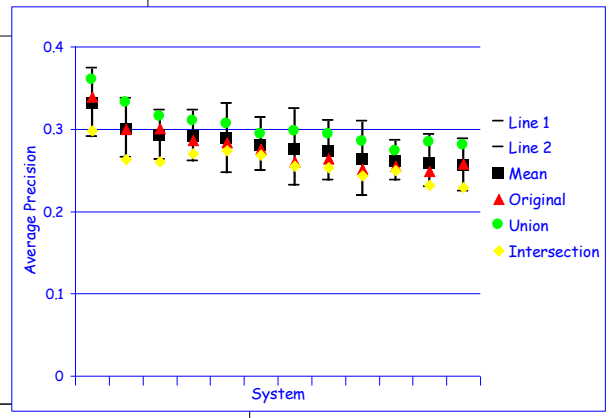
Reusability of TREC-8 Collection

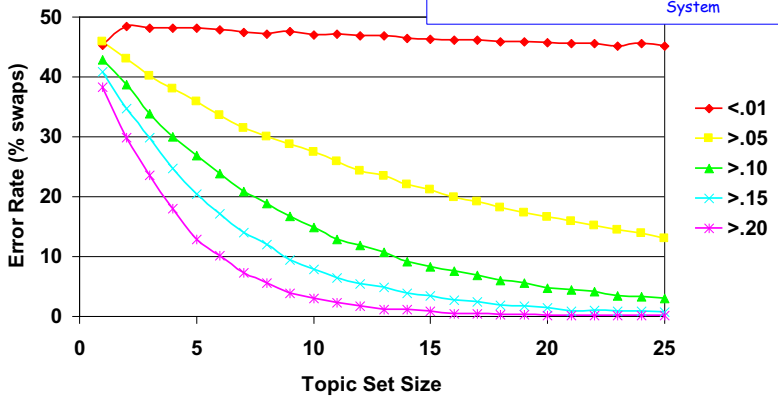


- Even individual topic τ's are stable
  - smallest is 0.8852, and that was caused by many tied scores magnifying the apparent difference
- But… what about some even newer, fancier system?
  - can't conclusively prove it is unaffected unless all documents judged
  - but incredibly unlikely to be significantly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

State of IR Evaluation, 2002





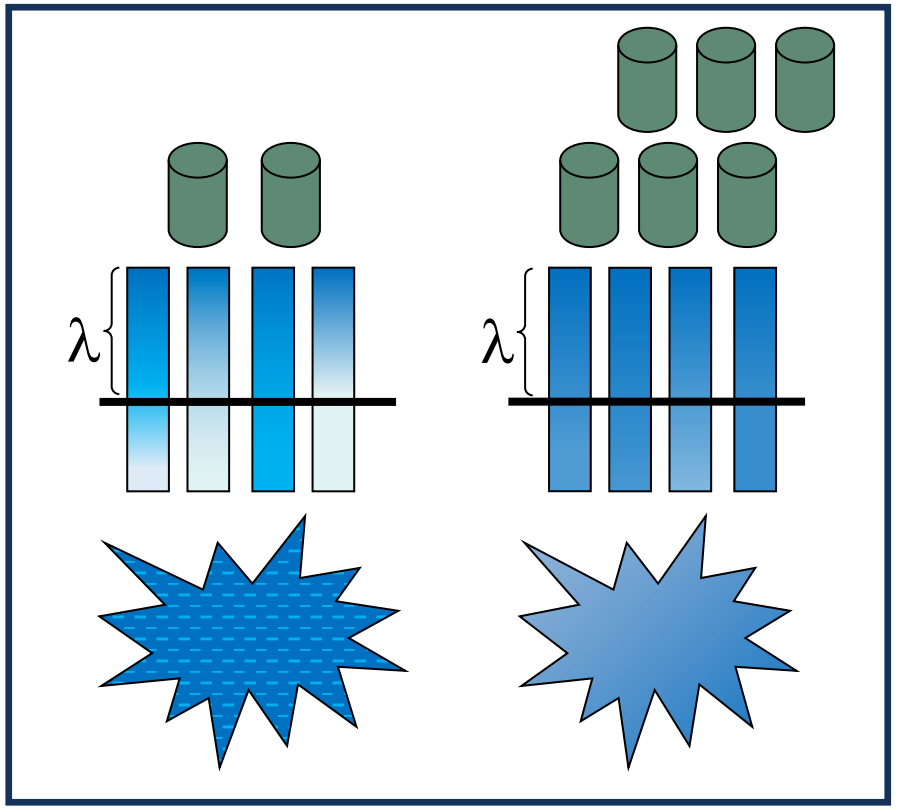


- Cranfield noisy but valid
  - judgments depend on assessor, but rankings stable
  - even with stark user abstraction in Cranfield, user (topic plus judgments) is still the biggest variable
  - number of topics
- Main effect: comparative results only

Voorhees, E.M. (2002). The Philosophy of Information Retrieval Evaluation. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds) Evaluation of Cross-Language Information Retrieval Systems. CLEF 2001. Lecture Notes in Computer Science, vol 2406. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45691-0_34

Pooling Bias (TREC Robust track, 2005)

- Traditional pooling takes top λ documents
  1) intentional bias toward top ranks where relevant are found
  2) 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
  1) sample runs differently to build unbiased pools
  2) 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.

LOU (or LOTO) Test

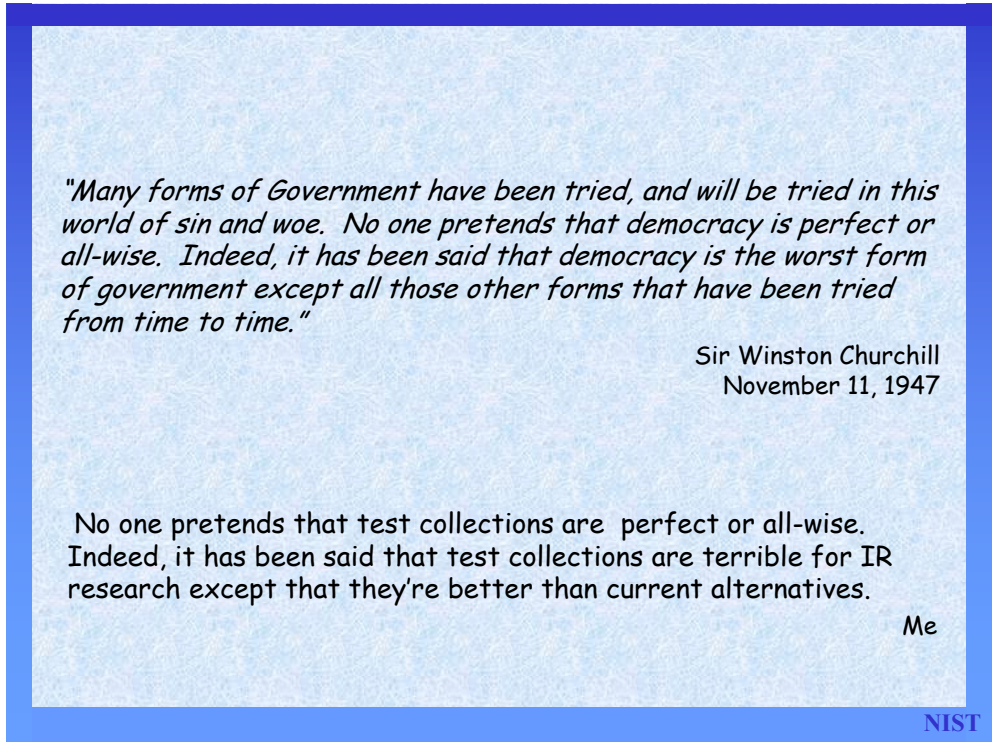
"Leave Out Uniques" test of reusability (variant of Zobel, SIGIR 1998):

examine effect on test collection if some participating team had not done so

Procedure

- create judgment set that removes all uniquely-retrieved relevant documents for one team
- evaluate all runs using original judgment set and again using newly created set
- compare evaluation results
  - Kendall's τ between system rankings
  - maximum drop in ranking over runs submitted by team

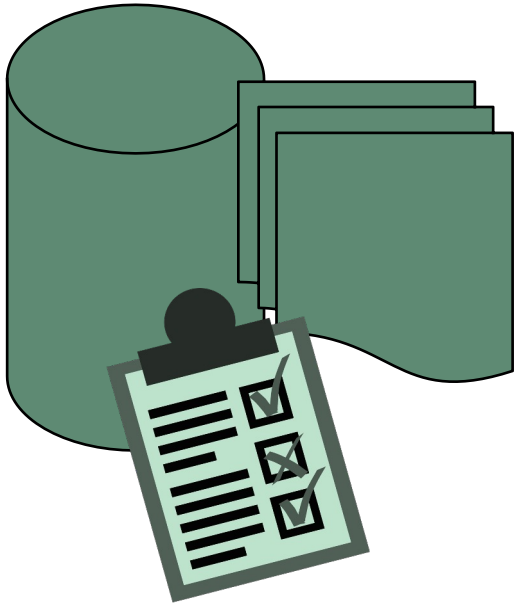
The more things change…



opening slide to my presentation at AIR workshop, Glasgow, 2006

- Relevance judgments are poor models of searchers
  - AIR workshop, Glasgow 2006; http://www.dcs.gla.ac.uk/workshops/air/
- Recall is unknowable
  - Justin Zobel, Alistair Moffat, and Laurence A.F. Park. 2009. Against recall: is it persistence, cardinality, density, coverage, or totality? **SIGIR Forum** 43, 1 (June 2009), 3–8. https://doi.org/10.1145/1670598.1670600
- Results on test collections are not representative of operational retrieval systems
  - Turpin AH, Hersh W (2001) Why batch and user evaluations do not give the same results, **SIGIR 2001**, pp 225–231
- IR evaluation is methodologically unsound
  - Norbert Fuhr. 2018. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. **SIGIR Forum** 51, 3 (December 2017), 32–41. https://doi.org/10.1145/3190580.3190586

Building Retrieval Test Collections



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

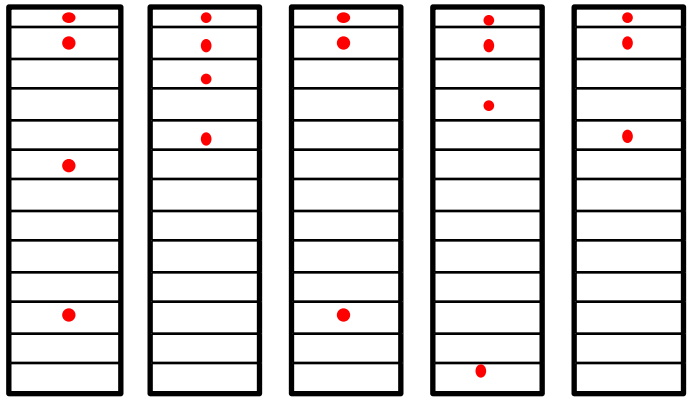
GENERAL PURPOSE
Supports a wide range of measures and search scenarios

REUSABLE
Unbiased for systems not used to build the collection

ACCEPTABLE COST
Cost proportional to number of human relevance judgments needed

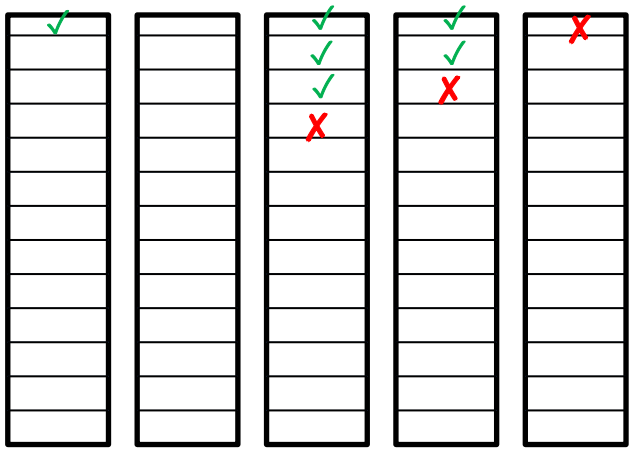
Inferred Measure Sampling

- 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 may be less reusable



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



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.

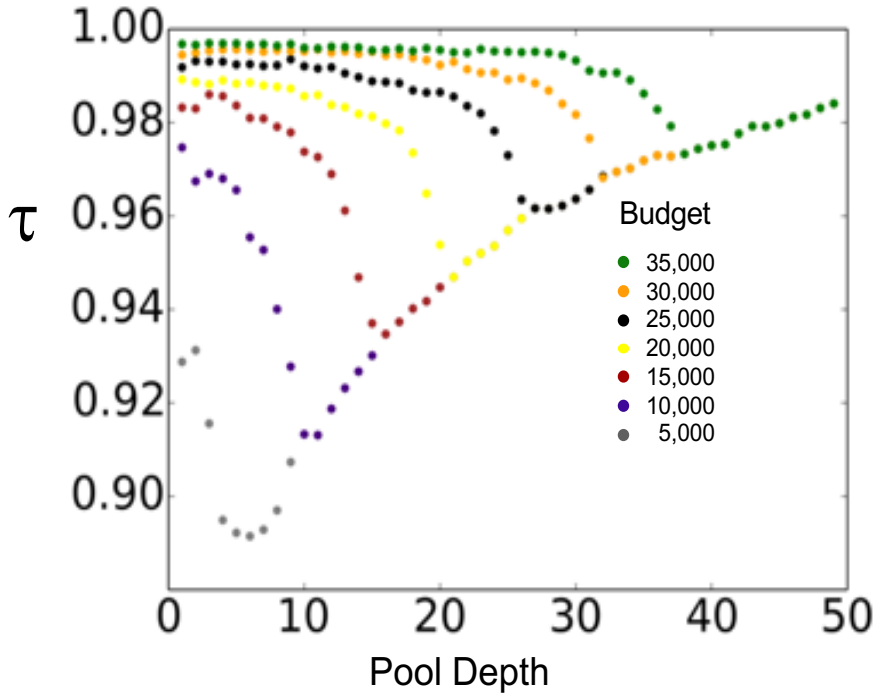
- Bandit techniques trade-off between exploiting known good "arms" and exploring to find better arms. For collection building, each run is an arm, and reward is finding a relevant doc
- Simulations suggest can get similar-quality collections as pooling but with many fewer judgments
- TREC 2017 Common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

bandit selection method:
2017: MaxMean 2018: MTF

Implementing a practical bandit approach

How does assessor learn topic?

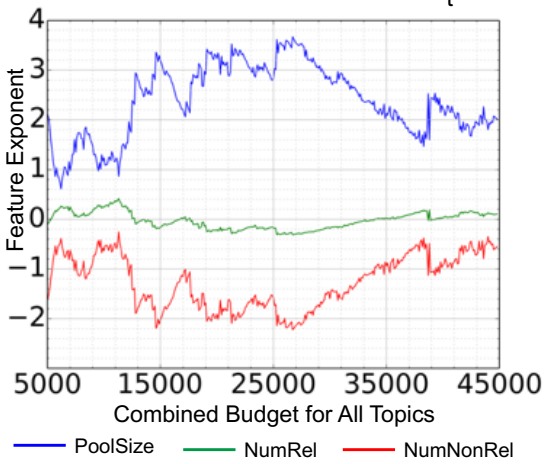
- allocating some budget to shallow pools causes minimal degradation over "pure" bandit method



How should overall budget be divided among topics?

- use features of top-10 pools, to predict per-topic minimum judgments needed

$$\text{Estimate}_t = \frac{\text{PoolSize}_t}{\sqrt{\text{NumNonRel}_t}}$$

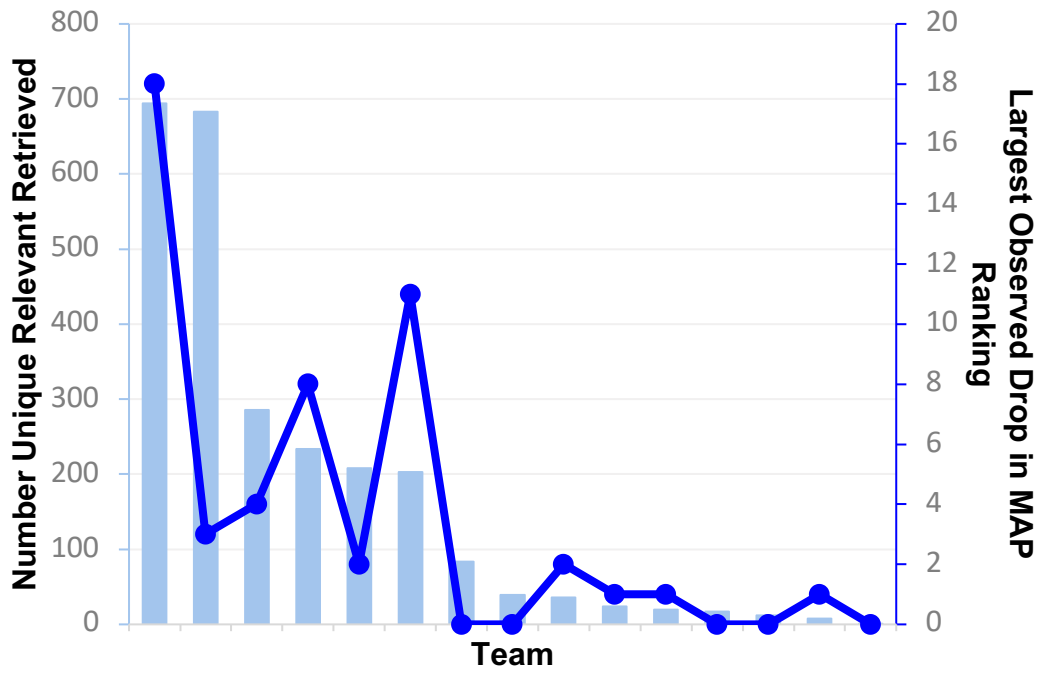


- results in a conservative, but reasonable, allocation of budget across topics for historical collections

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

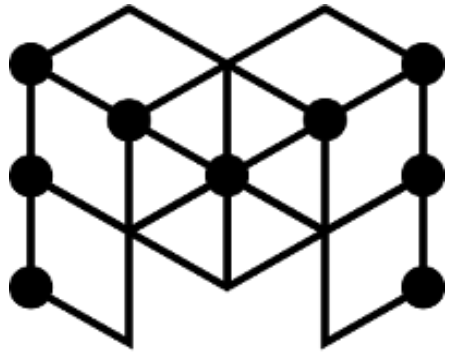
- must reserve some of budget for quality control, so operative number of judgments is less than B

- 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 in TREC

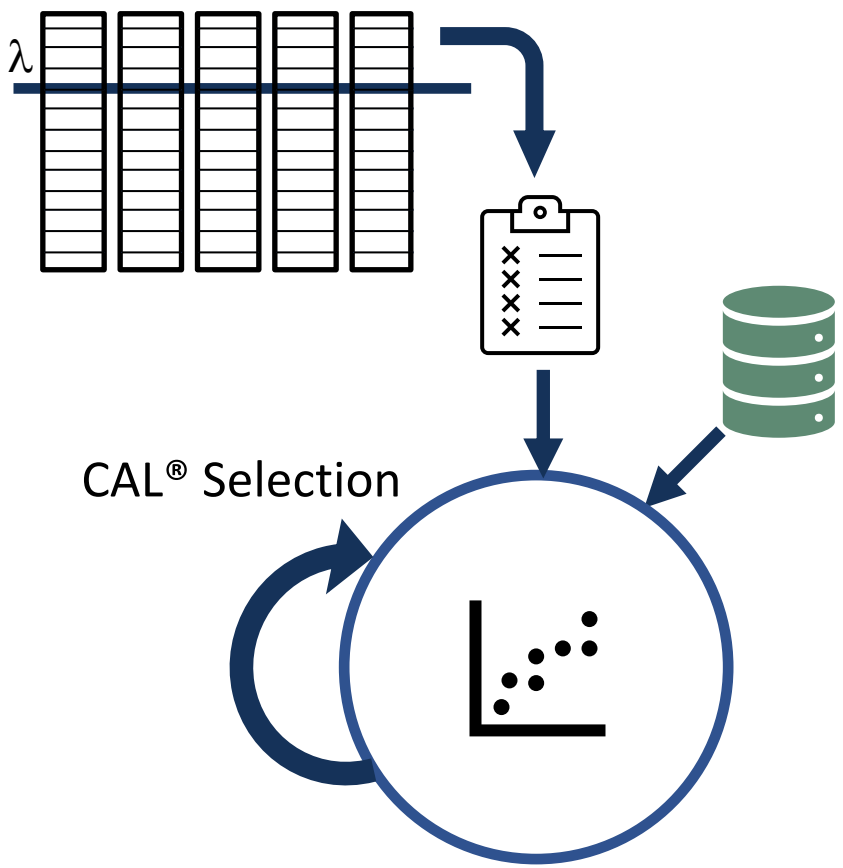
- "Study IR evaluation in a large data regime"
- Coordinated with MS MARCO leaderboard
- TREC track started in 2019
  - build typical TREC test collections for both Documents and Passages corpora: relatively deep judgments for ~50 queries
  - MS MARCO data (hundreds of thousands of queries with ~1 judgment each) available for training
  - ndcg@10 primary measure



MS MARCO

TREC

Deep Learning Track

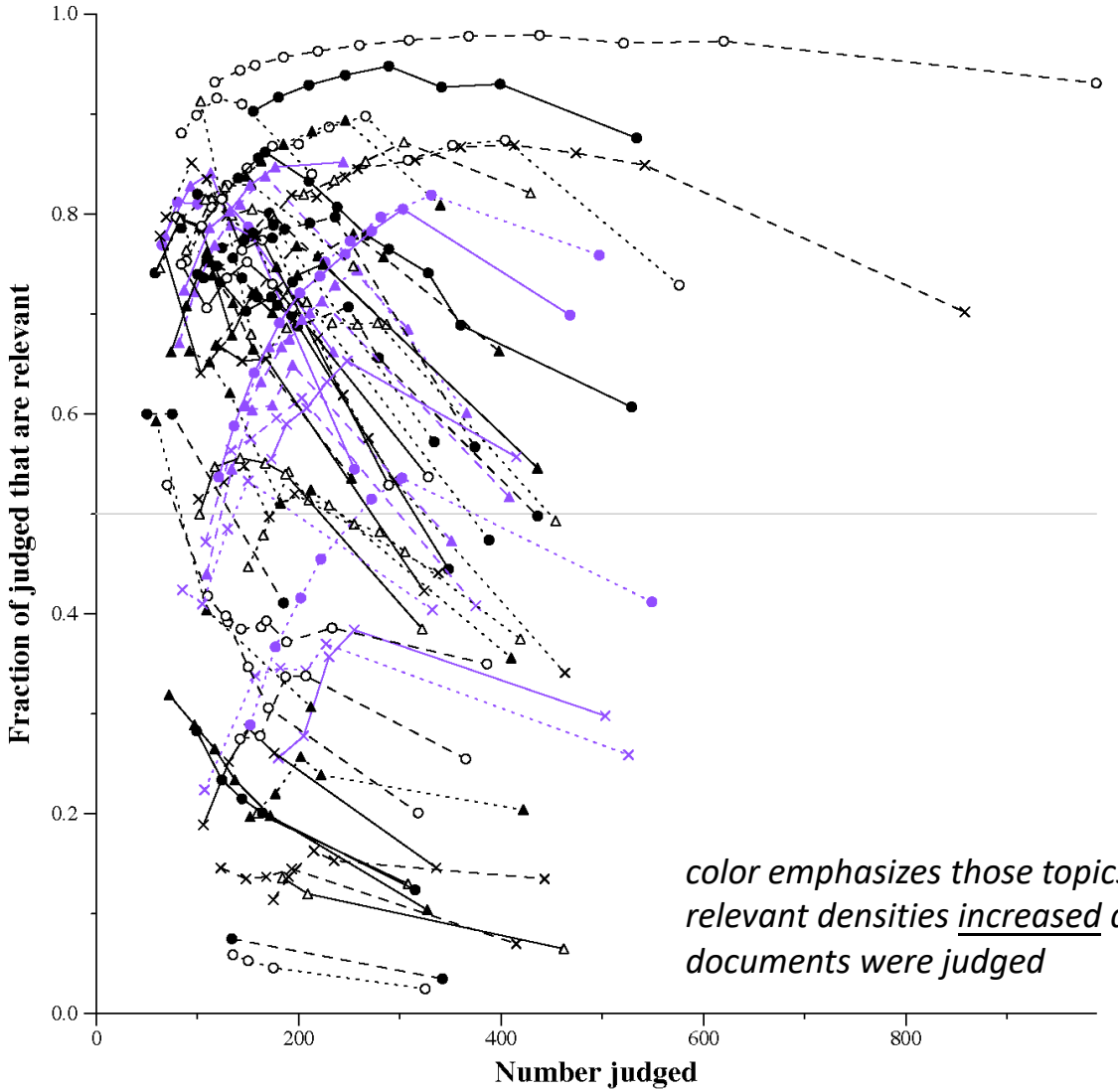


- Collections built using shallow pools followed by Continuous Active Learning
  - judge depth-10 pools across submissions
  - given set of relevance judgments, CAL builds model of relevance and orders remaining collection by likelihood of relevance
  - loop on obtaining judgments and running CAL per 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

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

Judgments

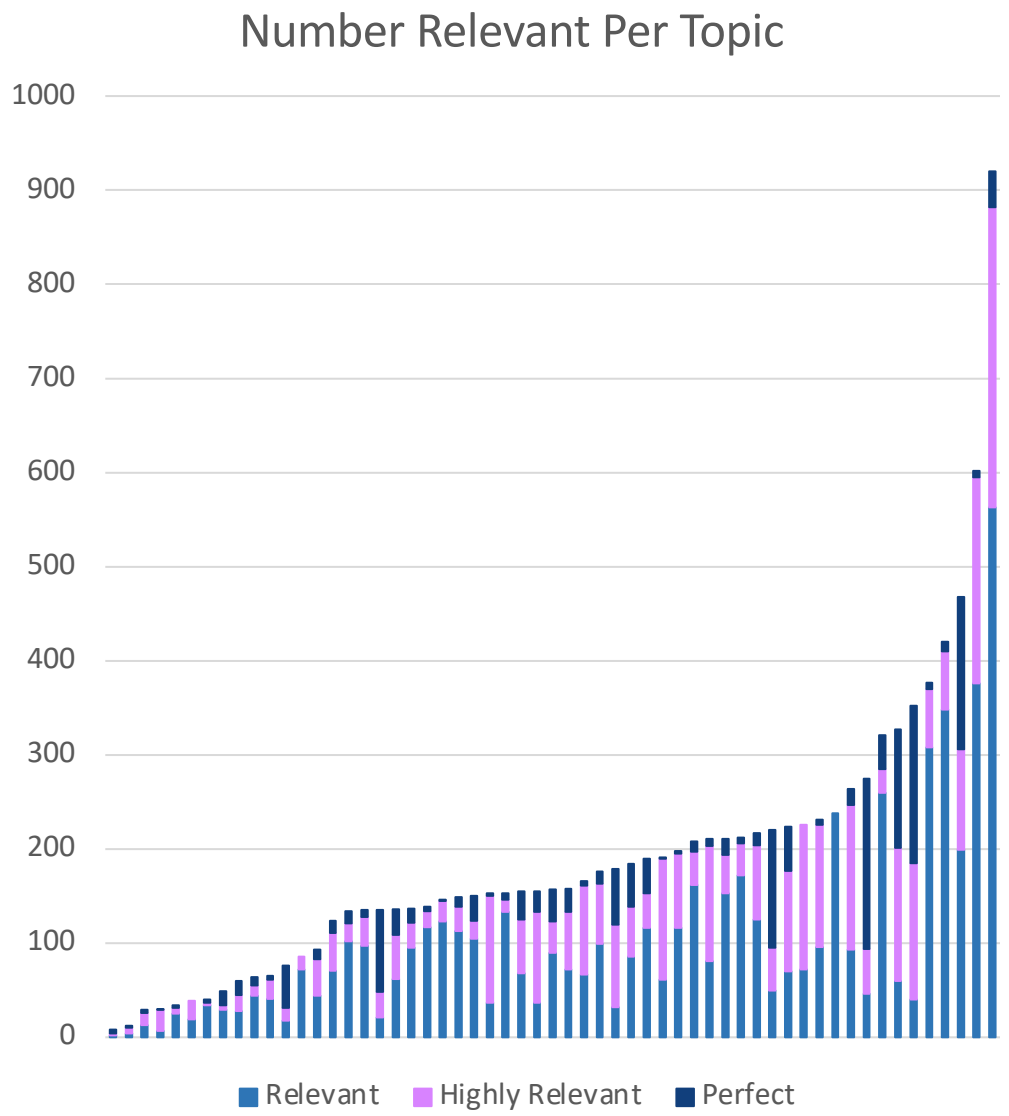
- Four relevance grades
- Two judgment phases
  - track judging in Sept. 2021
    - 13,058 total judgments
    - mean 229.1 [min 75,max 620]
  - supplementary phase in Dec. 2021
    - additional 9255 judgments
    - no CAL; docs selected to support collection experiments
- In track judging, 40/57 topics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



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

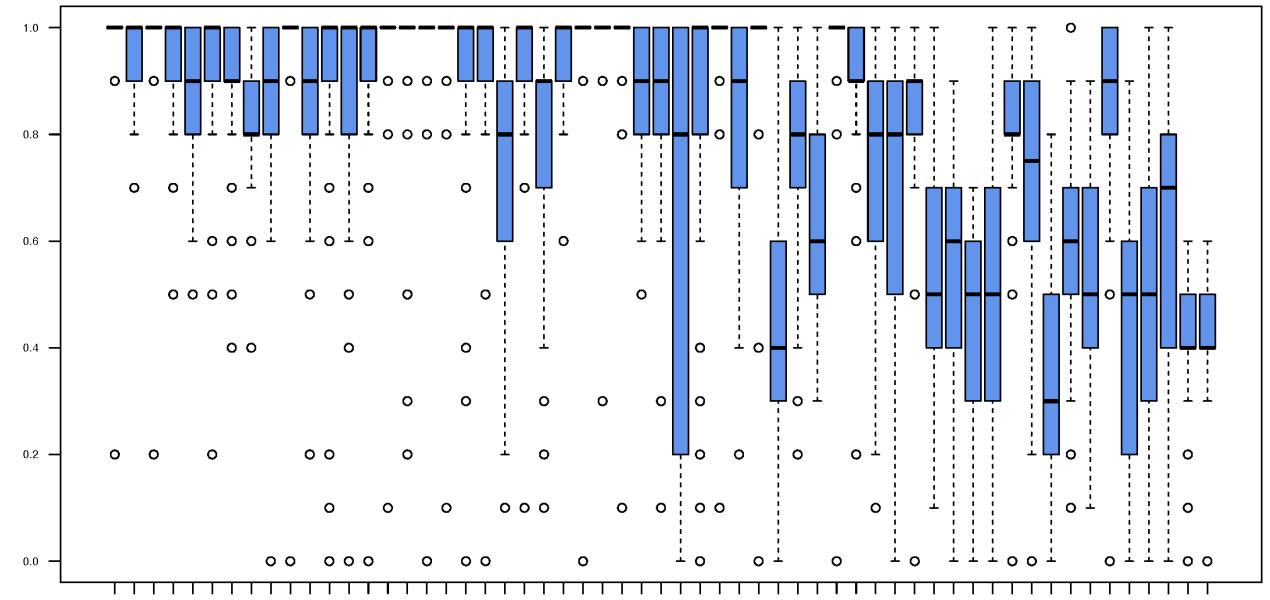
What Happened?

- 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.

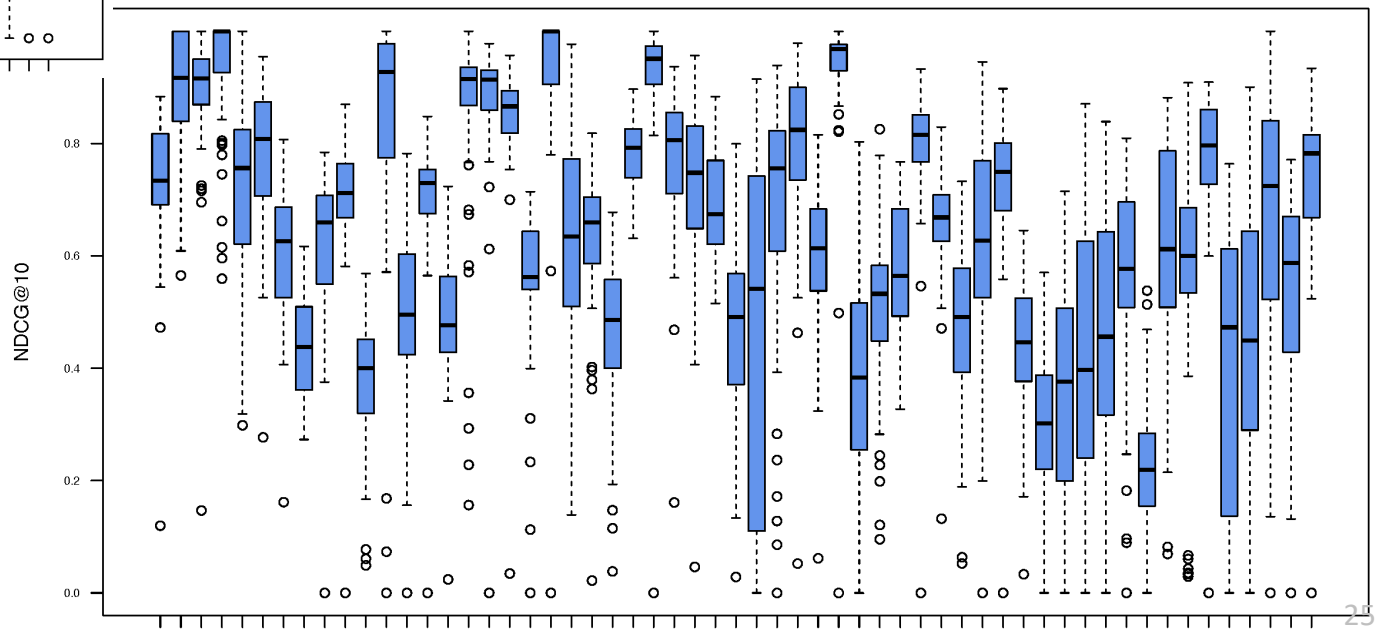
Deep Learning 2021 Scores



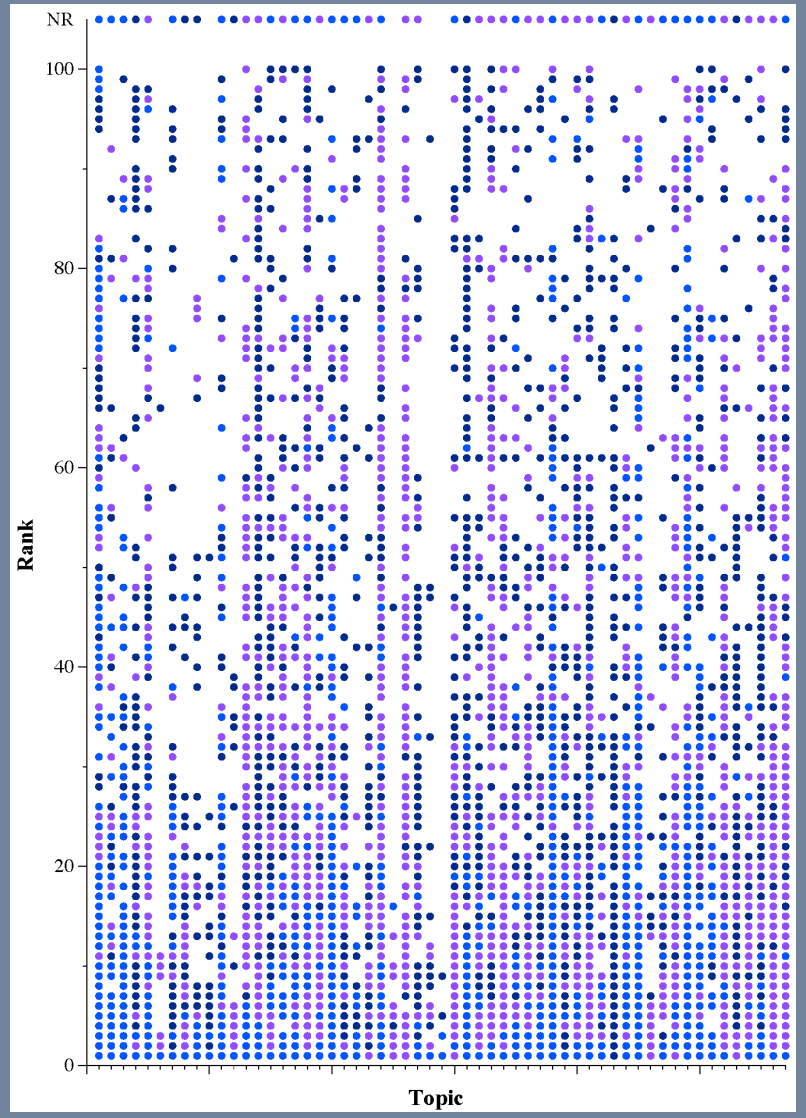
P@10

Distribution of Scores across Submissions for Each Topic

NDCG@10



Just re-define Relevant?



- Arbitrarily narrow definition of relevant not a solution
  - `relevant' needs to be defined by the use case, not the collection characteristics, for collection to be a useful tool
  - all collection-building techniques rely on systems being able to rank relevant docs highly
  - all grades of relevant documents distributed throughout the rankings
- Results suggest using deeper measures a better alternative
  - varying persistence parameter of RBP controls effective depth while also signaling incompleteness effects

Cranfield Paradigm

Defines “core competency” of IR: retrieve relevant before non-relevant

Abstracted task is a necessary but not sufficient proxy for real task

Different effectiveness metrics provide different abstractions of real-world task(s)

Documents, requests, measures should be representative of real use case for good test

Whither Cranfield?

- When pooling was no longer viable, Buckley et al. suggested:
  - form pools differently
  - engineer the topic set
  - down-sample the known relevant to a fair set
- These are all still options, but each needs more research to be actionable

image: Joshua Woroniecki/Pixabay