

Building Retrieval Test Collections

Ellen M. Voorhees



<http://trec.nist.gov>

Cranfield Paradigm



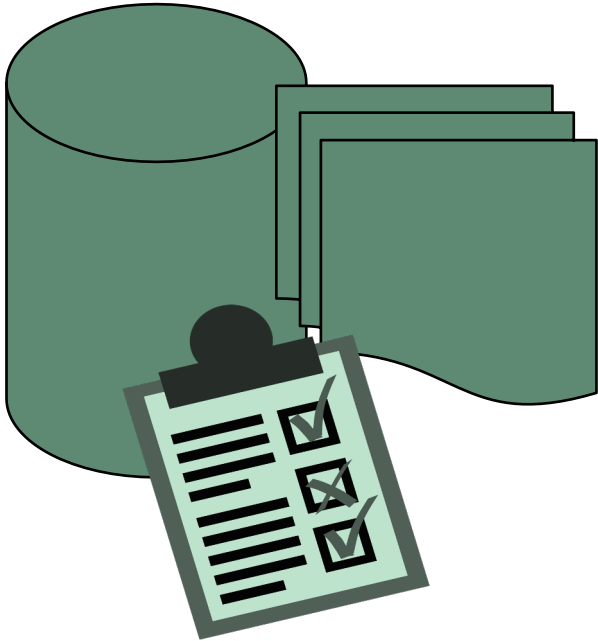
Cyril Cleverdon



a test collection

- Laboratory testing of retrieval systems first done in Cranfield II experiment (1963)
 - fixed document and query sets
 - evaluation based on relevance judgments
- Test collections
 - set of documents
 - set of questions
 - relevance judgments

Rationale for Cranfield



Sufficient fidelity to real user tasks to be informative

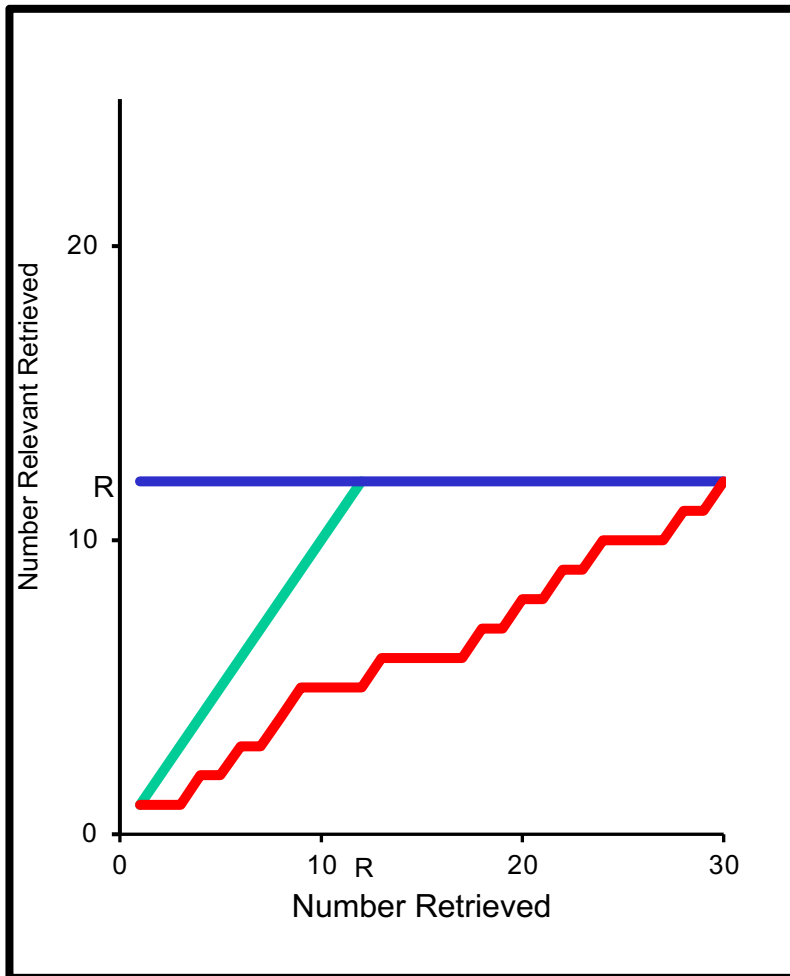


General enough to be broadly applicable, feasible, relatively inexpensive



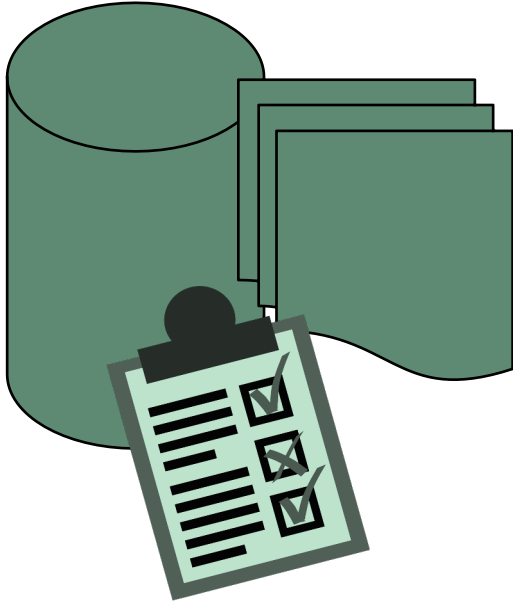
Lose realism to gain control over variables: more experimental power at lower cost

Cranfield Paradigm



- Retrieval system response to a question is a ranked list of documents.
- The ideal output is a list with all relevant documents ranked before any non-relevant document.
- Easy to compute a variety of different evaluation measures from a ranked list once you know the set of relevant documents

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

Text REtrieval Conference (TREC)

Workshop series that builds research infrastructure.



<http://trec.nist.gov>



pioneered use of “pooling” for building large collections



built > 150 test collections for dozens of search tasks



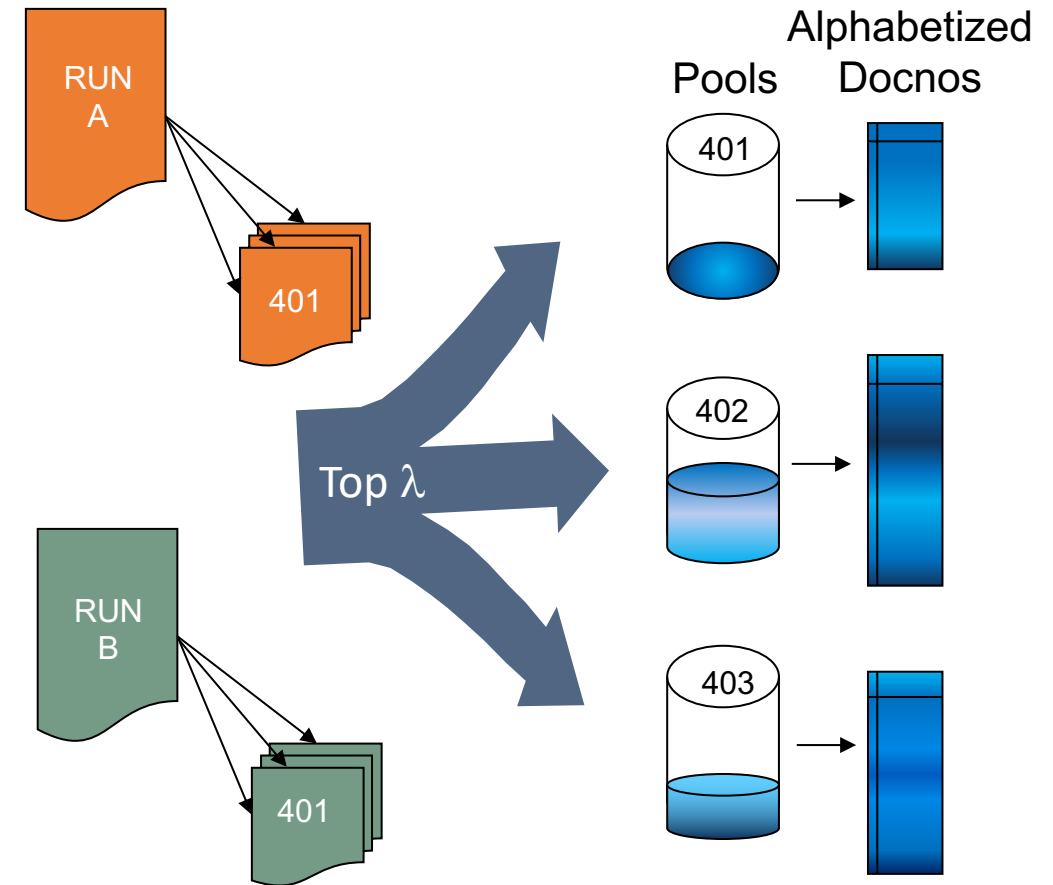
hundreds of participant teams world-wide



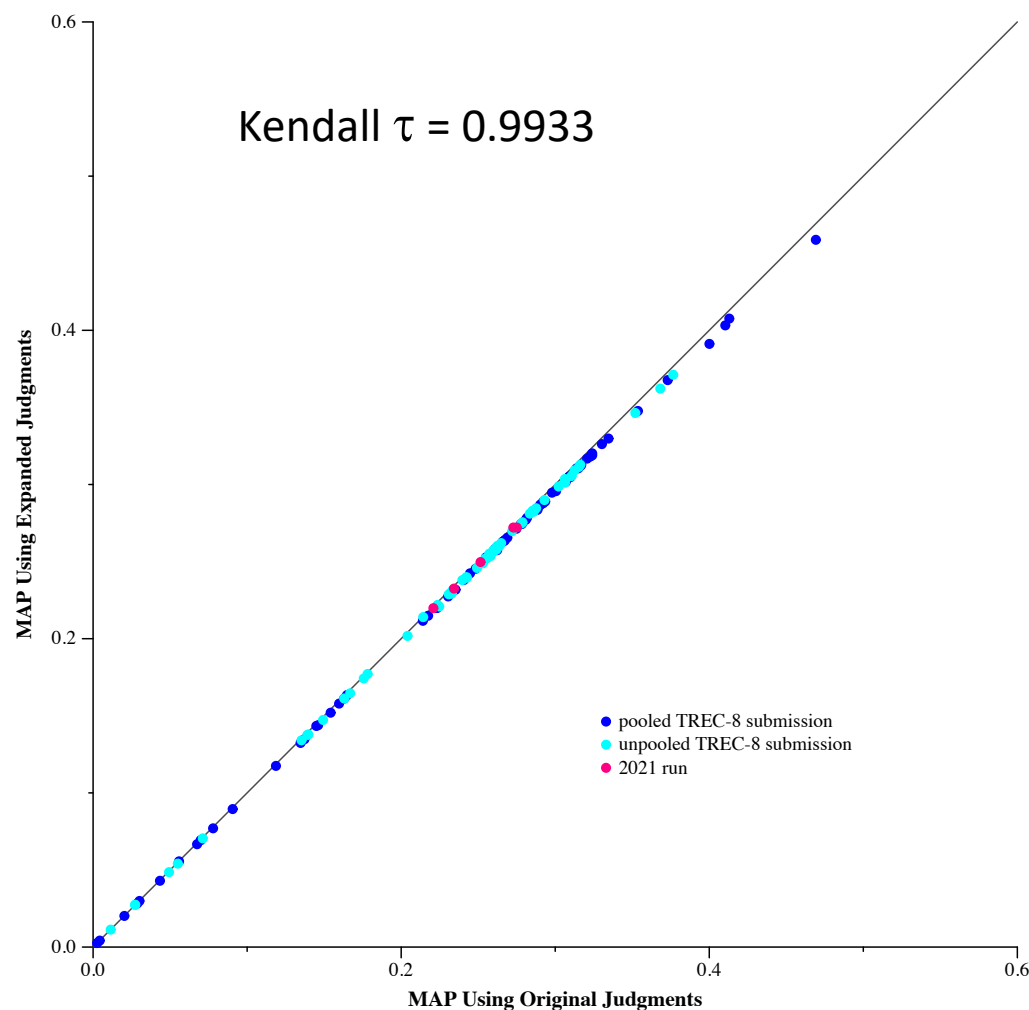
premier venue for determining research methodology

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 have been found to 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

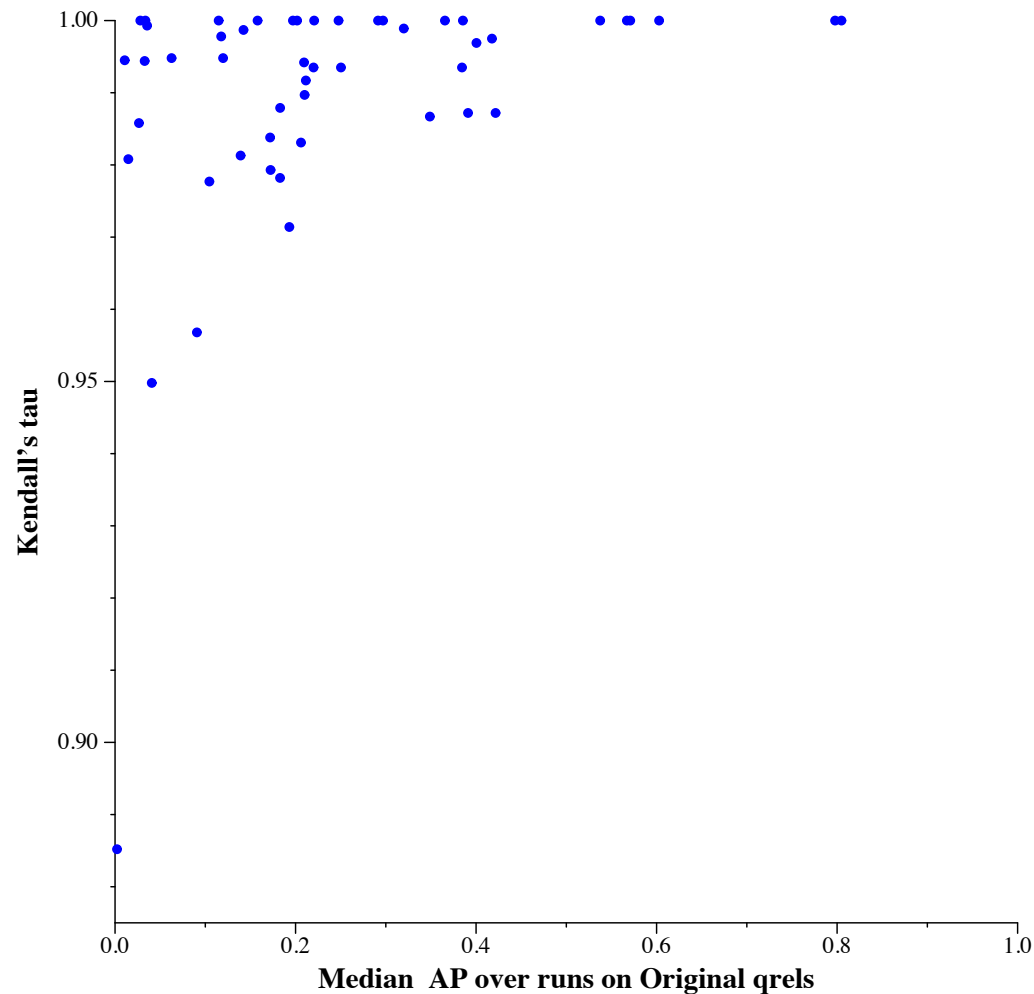


Reusability of TREC-8 Ad Hoc Collection



- 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

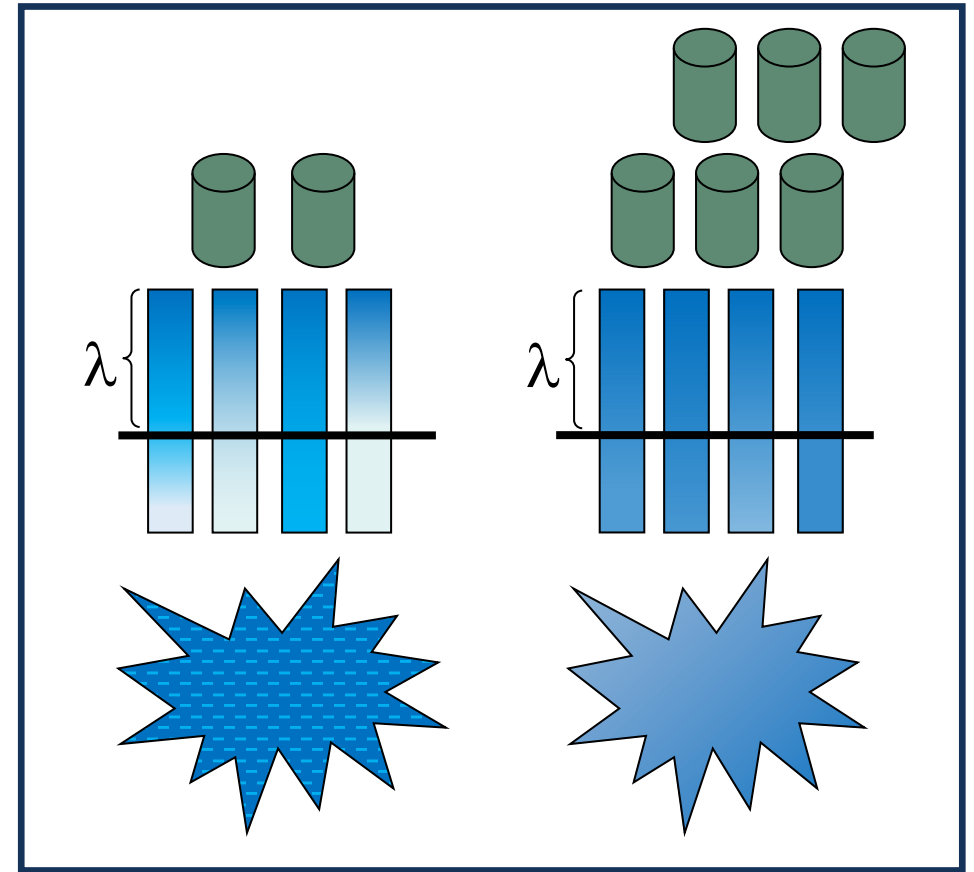
Reusability of TREC-8 Collection



- 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 relevant

Pooling Bias

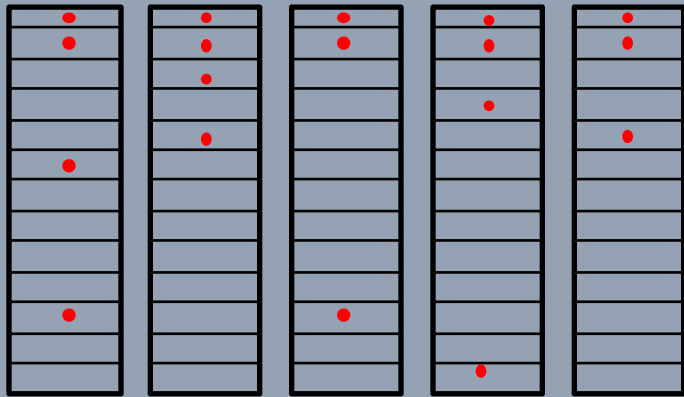
- But TREC-8 has relatively many judgments for modest corpus size
- Pooling assumes top λ documents is sufficient to reach past swell of topic-word relevant
- As document collection grows, a constant cut-off stays within swell
- Pools cannot be proportional to corpus size due to practical constraints
 - 1) sample runs differently to build unbiased pools
 - 2) new evaluation metrics that do not assume complete judgments



Alternate Construction Methods

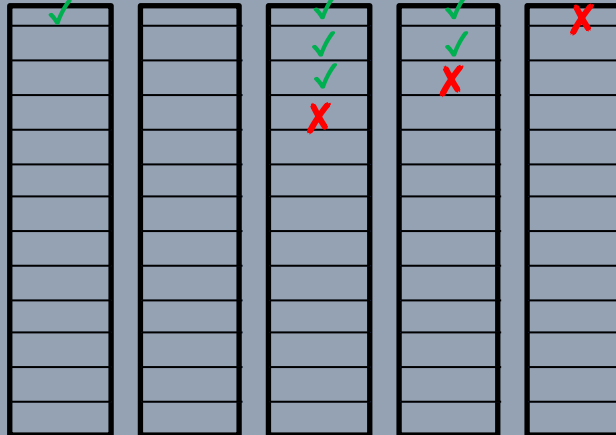
Goal is not to find the largest number of relevant documents possible.
Goal is to find a fair set of relevant documents.

Inferred Measures Sampling



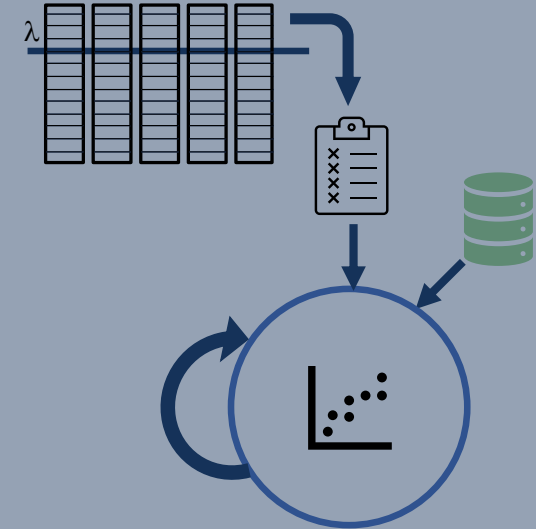
not general-purpose,
quality highly dependent on
definition of strata

Multi-arm Bandit Sampling



not fair, budget hard to control

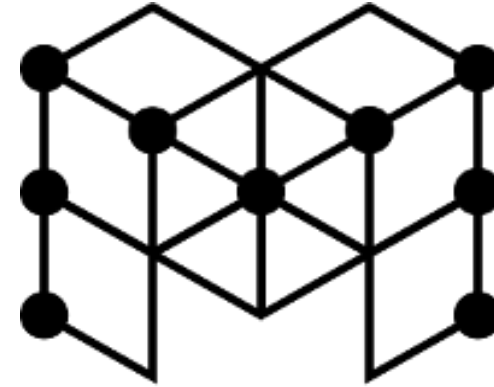
CAL[®] Selection



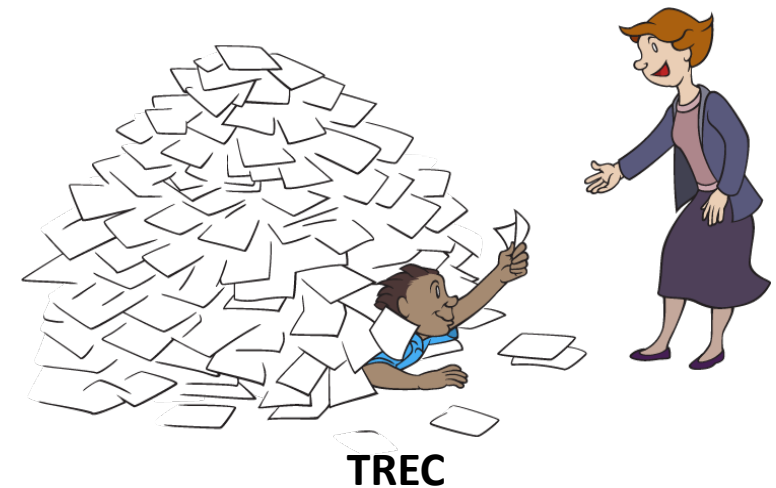
budget hard to control

Deep Learning Track in TREC

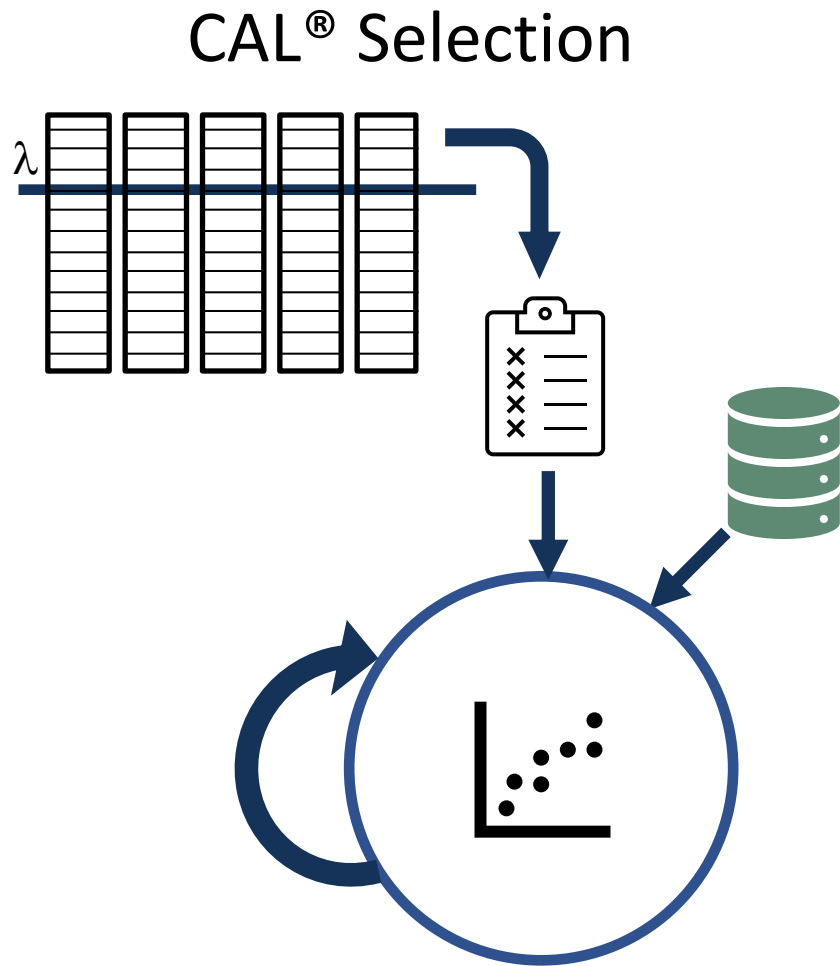
- “Study IR evaluation in a large data regime”
- Coordinated with MS MARCO leaderboard
- TREC track started in 2019
 - build typical TREC test collections for both Documents and Passages corpora: relatively deep judgments for ~50 queries
 - MS MARCO data (hundreds of thousands of queries with ~1 judgment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



MS MARCO



Deep Learning Track



- Collections built using shallow pools followed by Continuous Active Learning process
 - 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

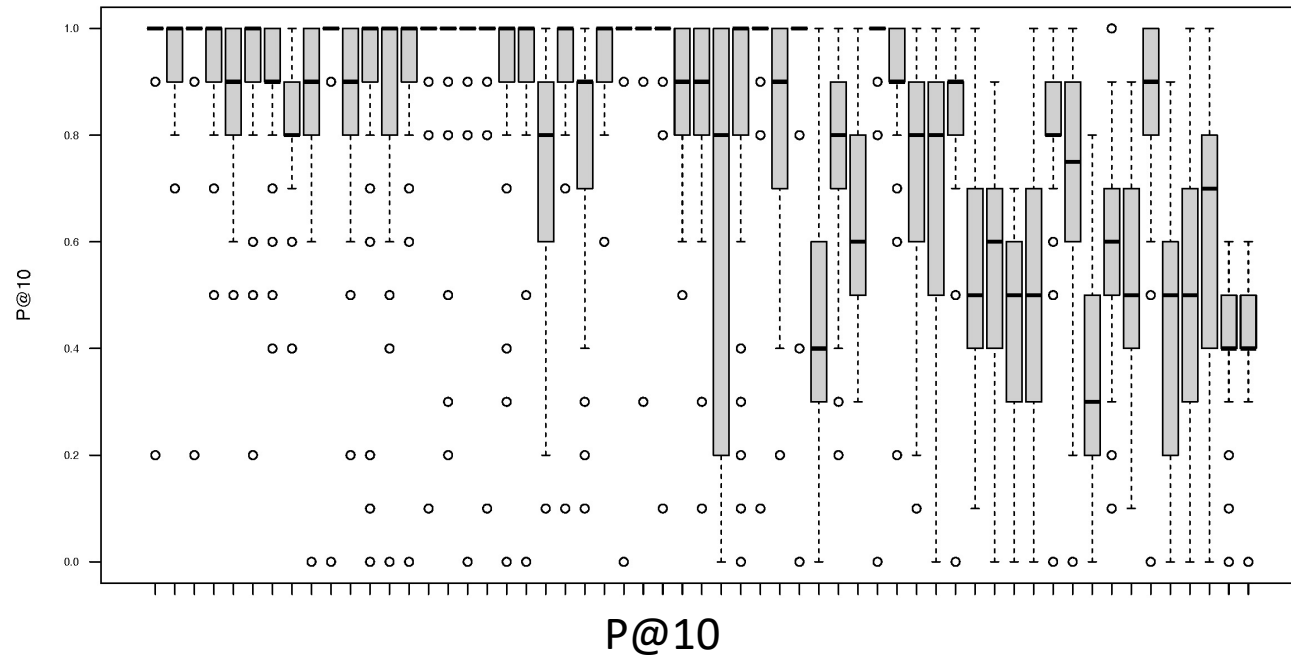
TREC Deep Learning 2021

	Documents	Passages
MS MARCO v. 1	3.2M	8.8
MS MARCO v. 2	12M	138M

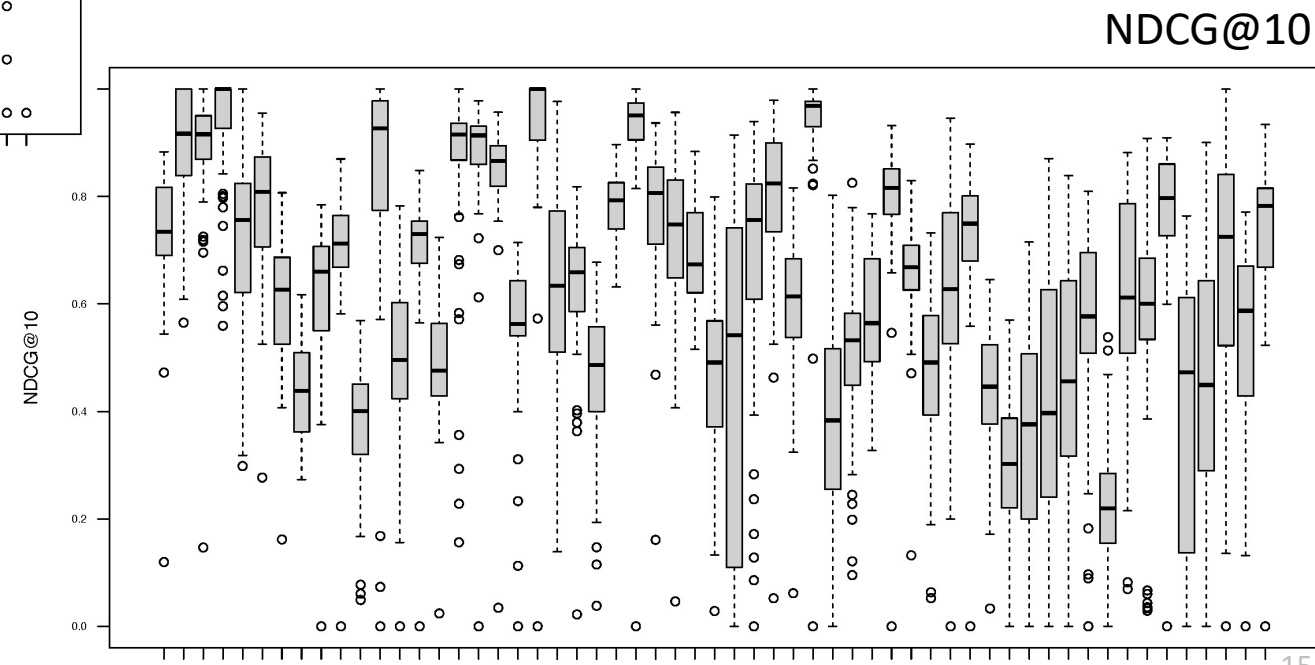
For 2021:	Documents	Passages
# topics judged	57	57
# topics in eval	57	53
min judgments	75	80
max judgments	620	339
mean judgments	229.1	204.3
total judgments	13,058	10,828
mean relevant	143.9	64.7

- Corpora sizes significantly bigger in 2021
 - document corpus 3.7 times as large
 - passage corpus 15.6 times as large
- Result is “too many” relevant documents
 - collections are not reusable
 - recall-based measures for track submissions are unreliable
 - high-precision scores are saturated, so comparisons are unstable
 - collection quality tests depend on finding relevant, so they are also less effective

Deep Learning 2021 Scores

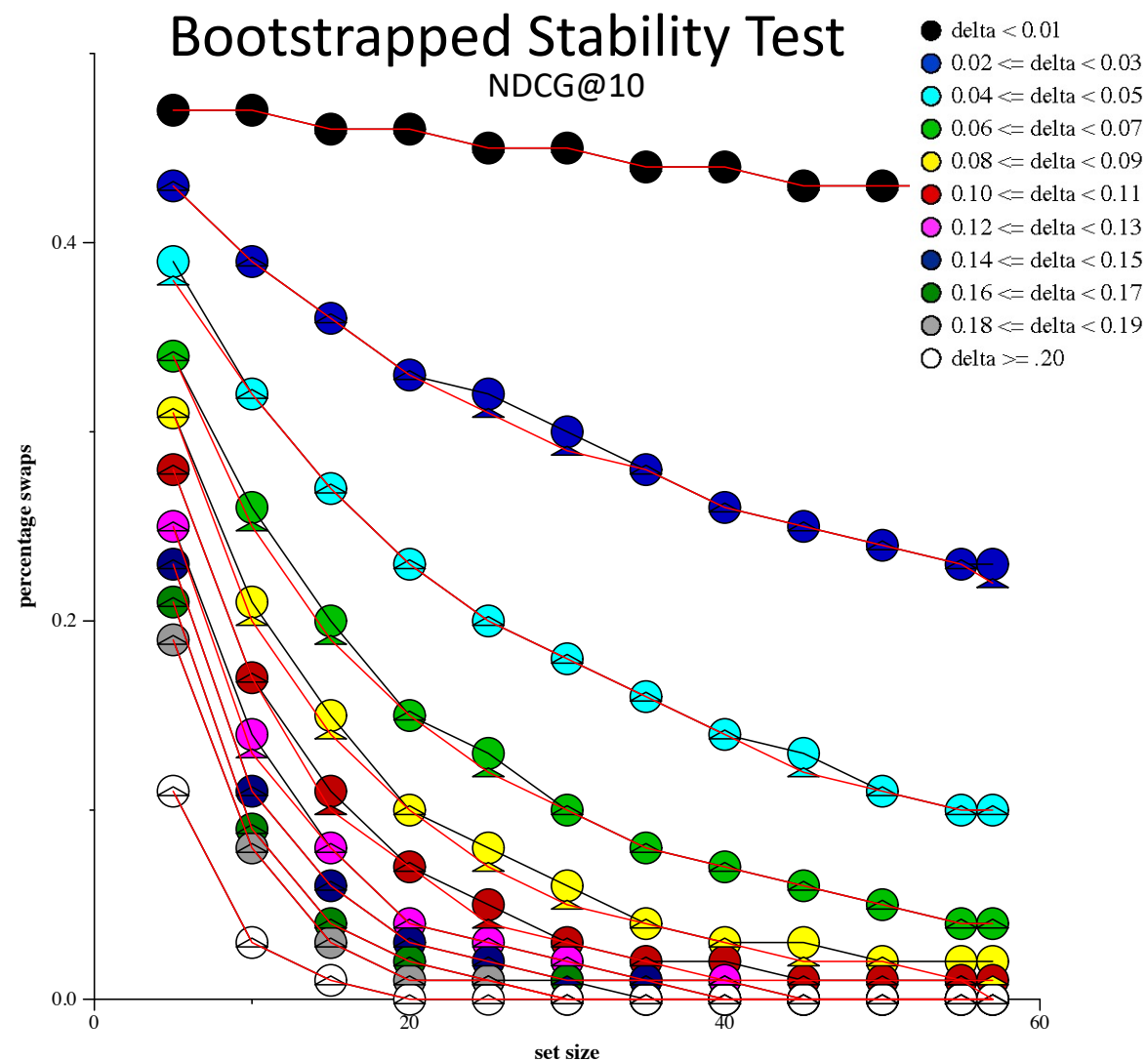
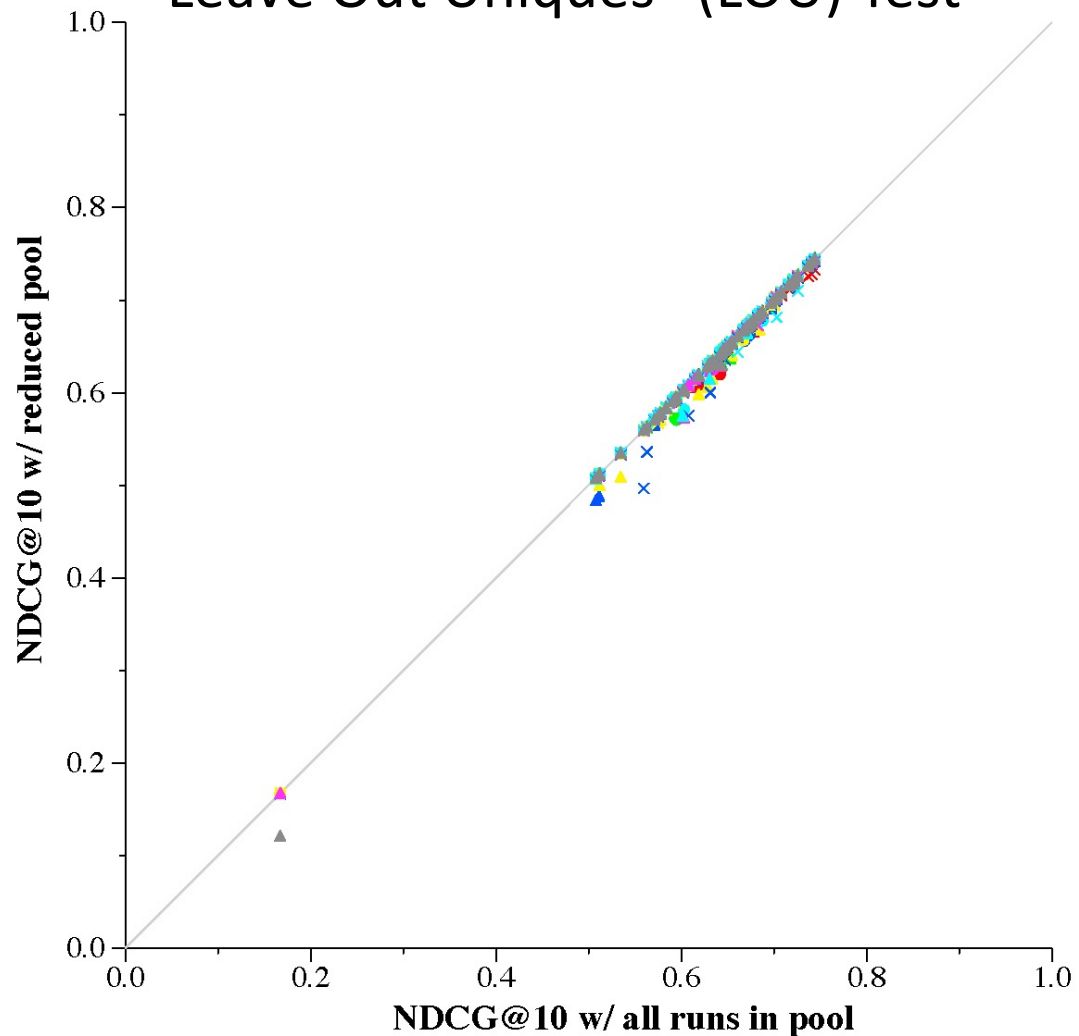


Distribution of Scores across
Submissions for Each Topic
(Document Ranking task)



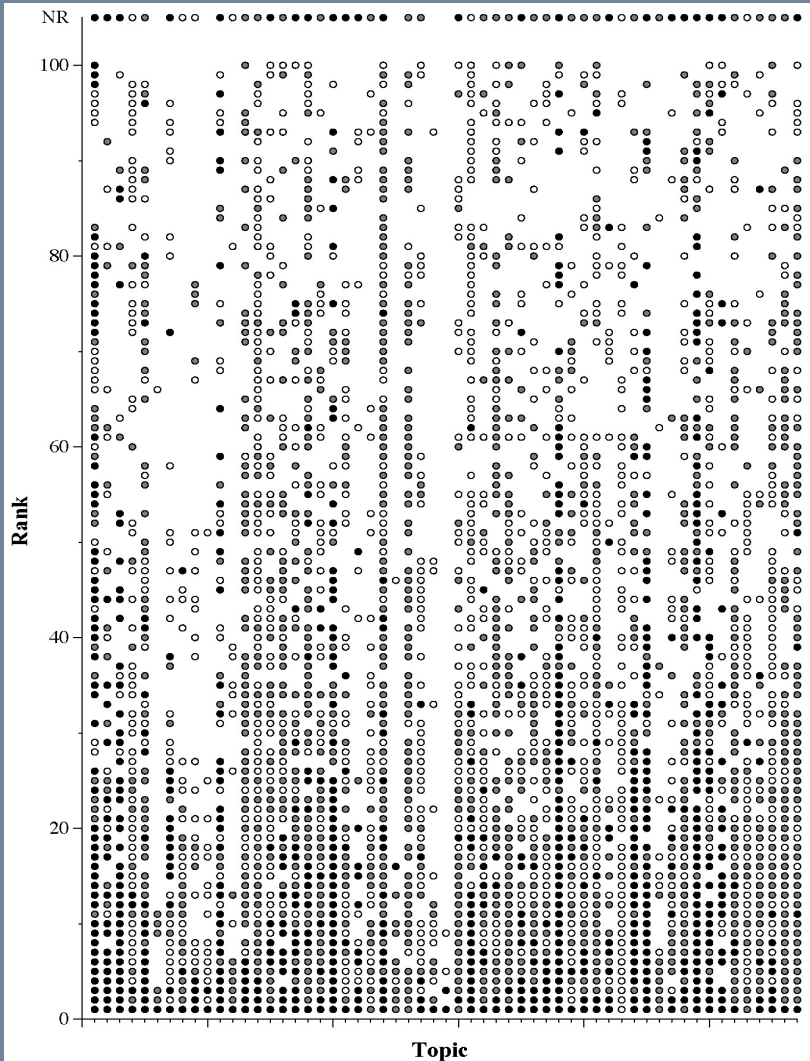
Tests of Collection Quality


“Leave Out Uniques” (LOU) Test



Way Forward?

Distribution of Relevant Documents in Submissions



- Declare success! 
- Arbitrarily narrow definition of relevant not a solution
 - all grades of relevant documents distributed throughout the rankings
 - '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
- Use deeper measures
 - addresses score saturation, but requires bigger budget
 - Rank-Biased Precision (RBP) gives some control over depth plus bounds on uncertainty in score

SOTA in Collection Building



NO SINGLE BEST TECHNIQUE

Quality of collection can depend on factors out of builder's control



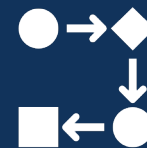
BUDGET

Need to reserve a portion of overall budget for quality control

R

NUMBER OF RELEVANT

Largest factor affecting collection quality



DYNAMIC METHODS

Potential cost savings are often blunted by practicalities in real use

So you want to build your own test collection...

- You sure?
 - existing collections allow comparison to other methods while saving expense of construction
 - seriously consider whether existing collection is sufficiently close abstraction
- Okay, what is your goal?
 - test alternatives to live system?
 - probably use A/B tests instead
 - document current effectiveness?
 - sample current stream for corpus/queries
 - use metrics used in operations
 - drive effectiveness improvements?
 - may want to oversample current challenges
 - high-precision measures unlikely to be sufficient for training



image: succo/Pixabay

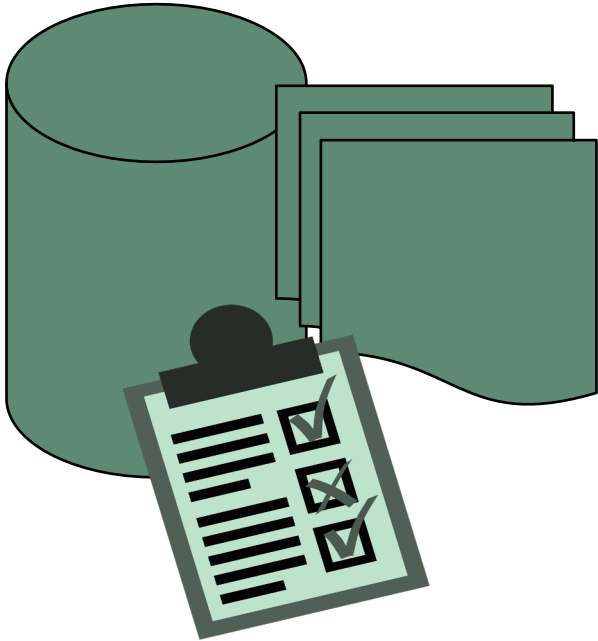
Further Considerations



image: succo/Pixabay

- Methods
 - sufficient number of diverse runs for pooling?
 - Minimum Test Collection (MTC) and variants for small set of known runs
 - CAL or similar active learning schemes
 - number queries vs. number judged per query
- Budget
 - need additional judgments for quality tests
 - novel run (or LOU) test
 - bootstrap stability
 - likely to introduce bias when adding unjudged from a small set of new runs
- Ongoing Use
 - beware overfitting to a single dataset
 - beware repeated test statistical testing

Reprise: Rationale for Test Collections



Sufficient fidelity to real user tasks to be informative *if task properly operationalized*



General enough to be broadly applicable; *feasibility and expense depend on (unknown) number of relevant*



Lose realism to gain control over *some* variables and thus more power; *appropriateness of alternatives context dependent*