Text REtrieval Conference

Ellen M. Voorhees

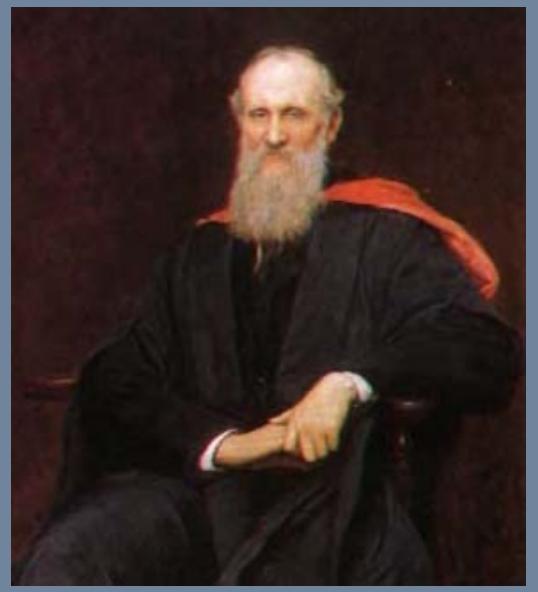
http://trec.nist.gov

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

National Institute of Standards and Technology

To promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

NIST mission statement



If you can not measure it, you can not improve it.

Lord Kelvin

Dr. Kelvin

Symbol: K

SI unit of
TEMPERATURE

Candela

Symbol: cd

SI unit of
BRIGHTNESS

Monsieur Kilogram

Symbol: kg

SI unit of
MASS

Major Uncertainty

Symbol: ?

ERROR!

Meter Man

Symbol: m

SI unit of
LENGTH

The Mole

Symbol: mol

SI unit of
AMOUNT OF SUBSTANCE

Ms. Ampere

Symbol: A

SI unit of
ELECTRIC CURRENT

Professor Second

Symbol: s

SI unit of
TIME

Text REtrieval Conference (TREC)

Workshop series that builds research infrastructure.

http://trec.nist.gov

TREC 2021 is 30th in series.

- 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

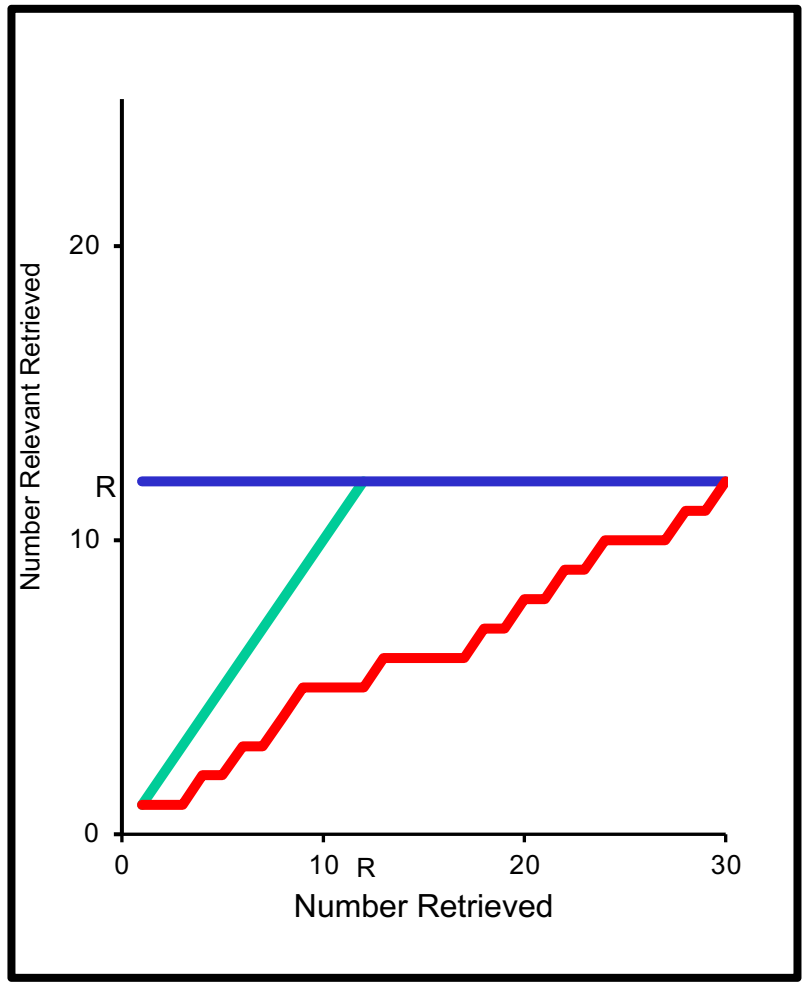
Cranfield Paradigm

Cyril Cleverdon

a test collection

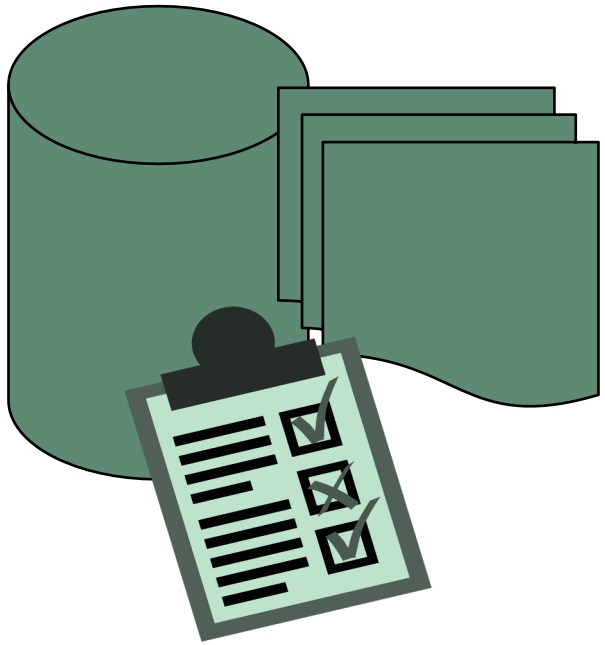
- Laboratory testing of retrieval systems first done in Cranfield II experiment (1963)
  - fixed document and query sets
  - evaluation based on relevance judgments
- Test collections
  - set of documents
  - set of questions
  - relevance judgments

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

Rationale for Cranfield



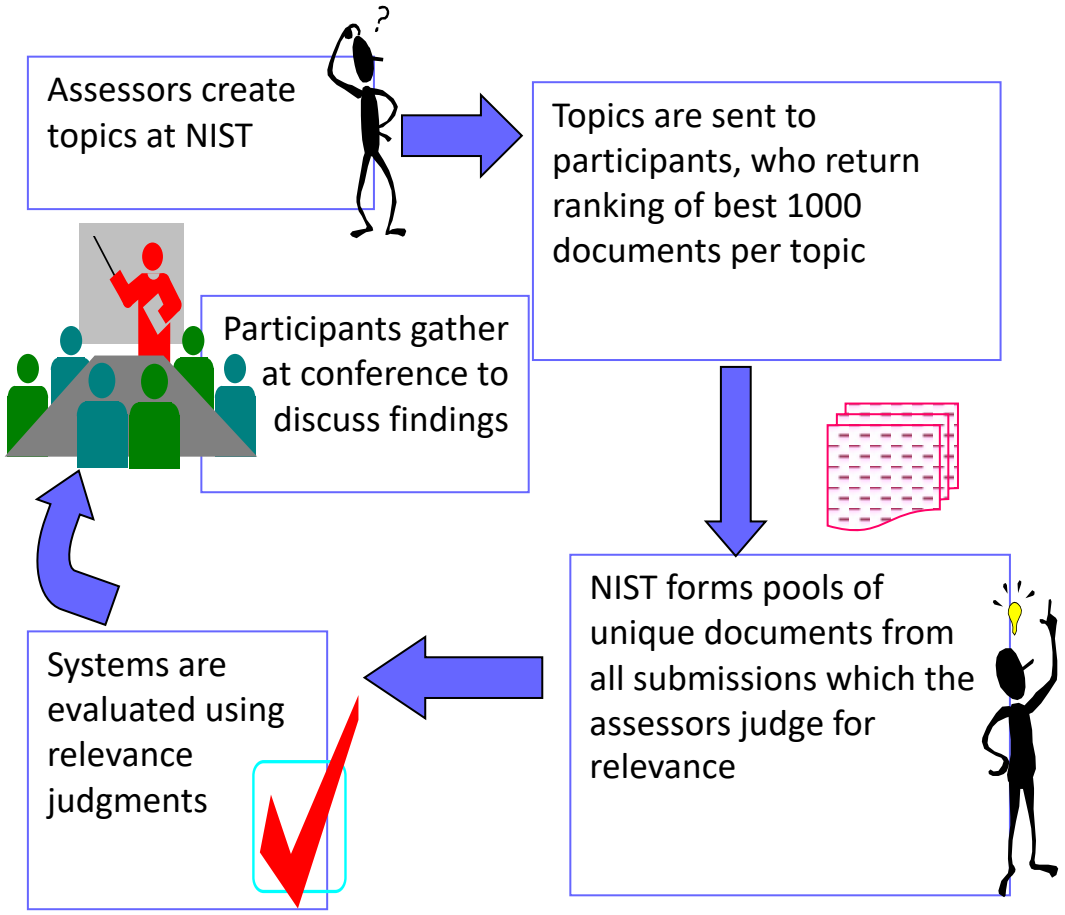
Sufficient fidelity to
real user tasks to be informative

General enough to be broadly applicable,
feasible, relatively inexpensive

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

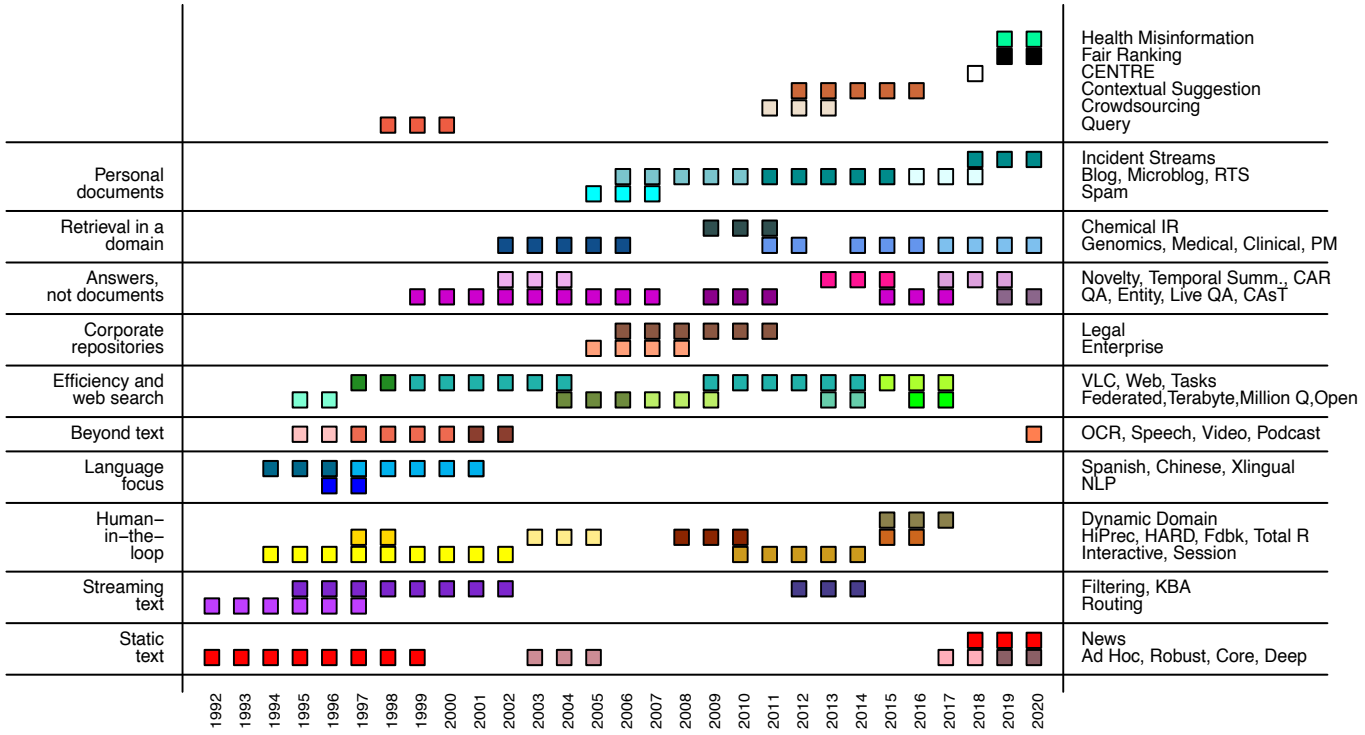
TREC Philosophy

- TREC is a modern example of the Cranfield tradition
  - system evaluation based on test collections
- Emphasis on advancing the state of the art from evaluation results
  - TREC's primary purpose is not competitive benchmarking
  - experimental workshop: sometimes experiments fail!



TREC Tracks

- A track is a task that focuses on a particular subproblem of information access
- Tracks invigorate TREC & keep it ahead of the state-of-the-art
  - specialized collections support research in new areas
  - first large-scale experiments debug what the task <u>really</u> is
  - provide evidence of technology's robustness



TREC Tracks

- Set of tracks in a particular TREC depends on:
  - interests of participants
  - appropriateness of task to TREC
  - needs of sponsors
  - resource constraints
- Need to submit proposal for new track in writing to NIST

TREC 2021 Tracks

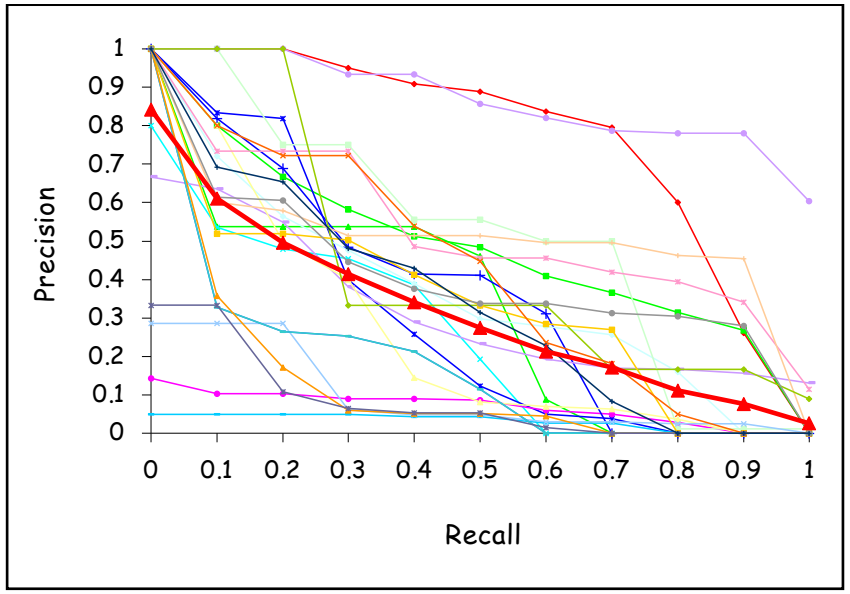
- Clinical Trials
- Conversational Assistance
- Deep Learning
- Fair Ranking
- Health Misinformation
- Incident Streams
- News
- Podcast

Public Data Sets

- Data needs to have high-fidelity to salient aspects of the true task. For search:
  - size, variability, naturalness are all important
  - can't model content well enough to generate synthetic data sets
- Human data: privacy issues difficult to recognize and compensate for
- Privacy is not the only issue: intellectual property, proprietary advantage can encumber data
- The more encumbered data is the less it will be used, but some data is (usually) better than none.

Search Evaluation is Hard

humans differ



system differences is variable that affects search effectiveness the least

size matters

Relevance Judgments are Made by Humans

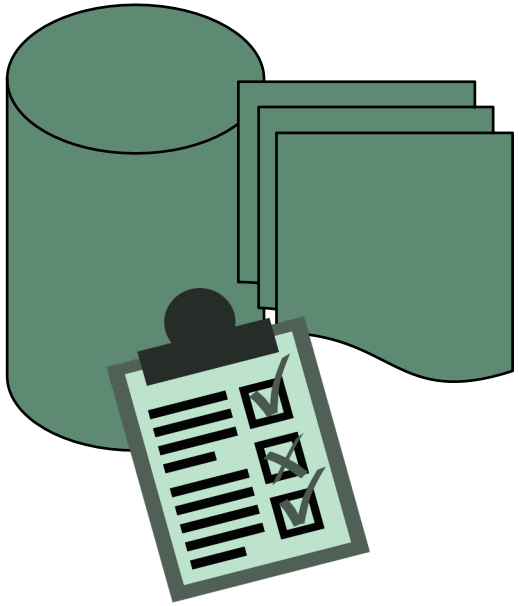
500,000 documents, judged 1/minute

$$\frac{500{,}000 \text{ mins}}{60 \text{ mins}} = 8333 \text{ hours}$$

$$= 347 \text{ days}$$

Not Possible.

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

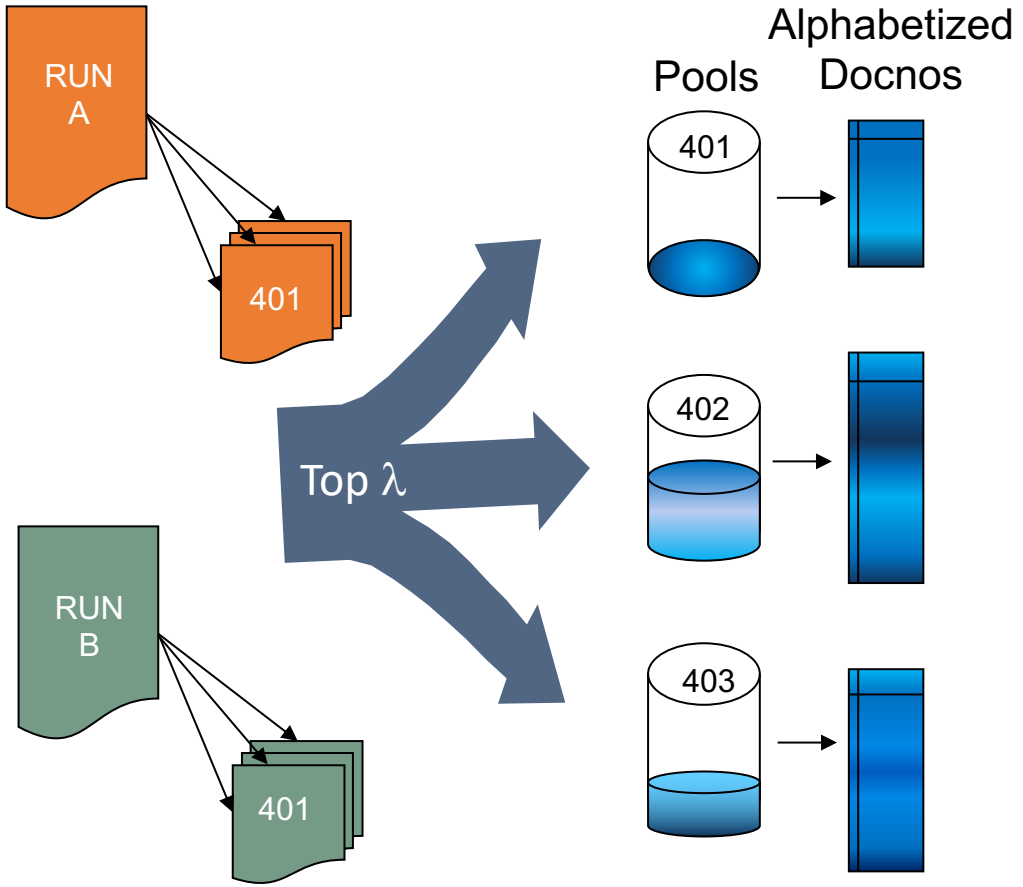
Unbiased for systems not used to build the collection

ACCEPTABLE COST

Cost proportional to number of human relevance judgments needed

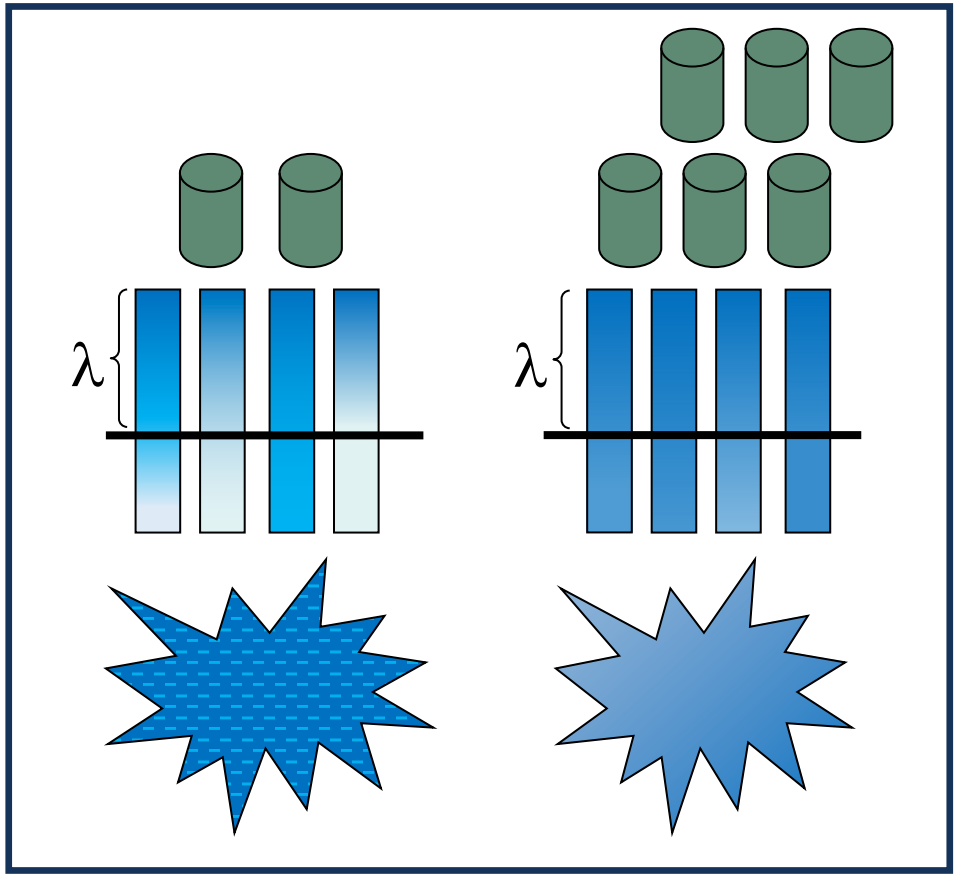
Pooling

- 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



Pooling Bias

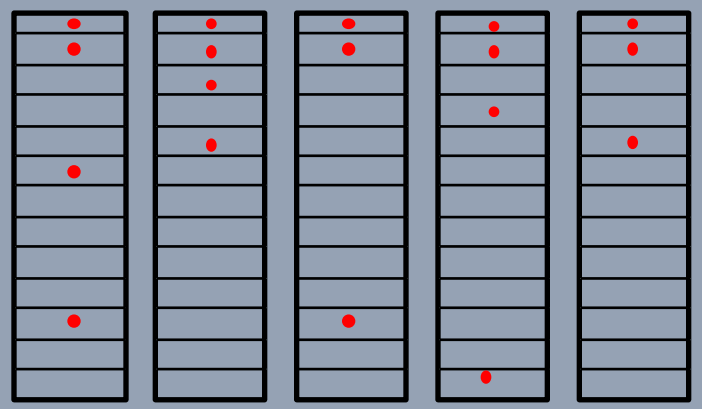
- Traditional pooling takes top λ documents
  1) intentional bias toward top ranks where relevant are found
  2) λ was originally large enough to reach past swell of topic-word relevant
- As document collection grows, a constant cut-off stays within swell
- Pools cannot be proportional to corpus size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



Alternate Construction Methods

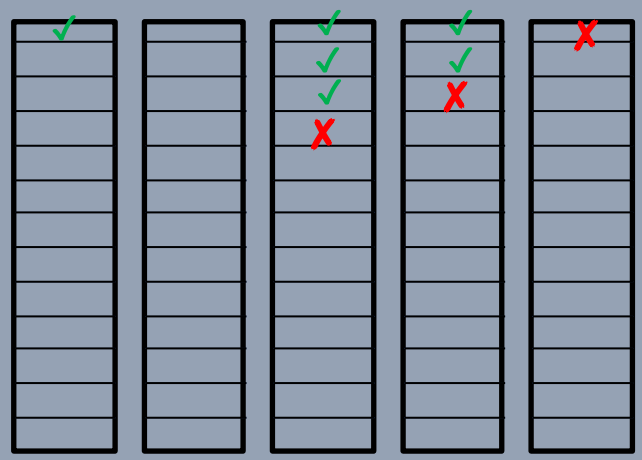
Goal is not to find the most relevant documents possible.
Goal is to find fair set of relevant documents.

Inferred Measures Sampling



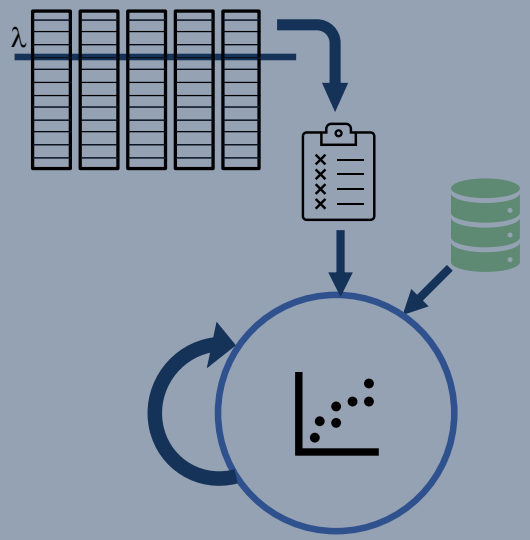
not general-purpose

Multi-arm Bandit Sampling



not fair, budget hard to control

CAL® Selection



budget hard to control

SOTA in Collection Building

NO SINGLE BEST TECHNIQUE

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

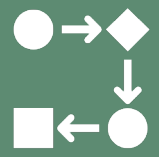
BUDGET

Need to reserve a portion of overall budget for quality control

R

NUMBER OF RELEVANT

Largest factor affecting collection quality

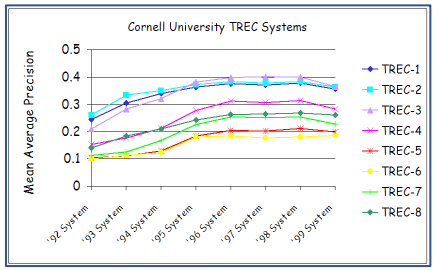


DYNAMIC METHODS

Potential cost savings are often blunted by practicalities in real use

TREC Impact

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.

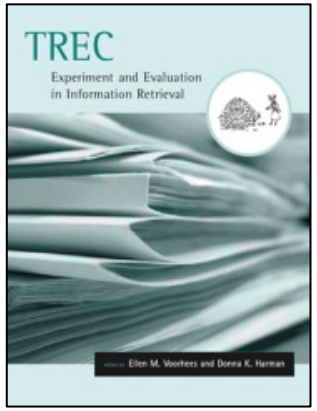
Hal Varian
Google Chief Economist
March 4, 2008

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