Retrieval Evaluation

- Reference Collections

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

1. *Modern Information Retrieval*. Chapter 3
2. *Text REtrieval Conference*. http://trec.nist.gov/
3. *Search Engines: Information Retrieval in Practice,* Chapter 8

Premises

- Research in IR has frequently been criticized on two fronts
  - *Lack a solid formal framework* as a basic foundation
    - The inherent *degree of psychological subjectiveness associated with the task decides the relevance* of a given document
      - Difficult to dismiss entirely
    - Relevance can be binary or graded
      - Binary relevance: relevant and not relevant
      - Graded relevance: e.g., highly relevant, relevant and not relevant
  - *Lack robust and consistent testbeds and benchmarks*
    - Small test collections did not reflect real-world application
    - No widely accepted benchmarks
      - Comparisons between various retrieval systems were difficult (different groups focus on different aspects of retrieval)

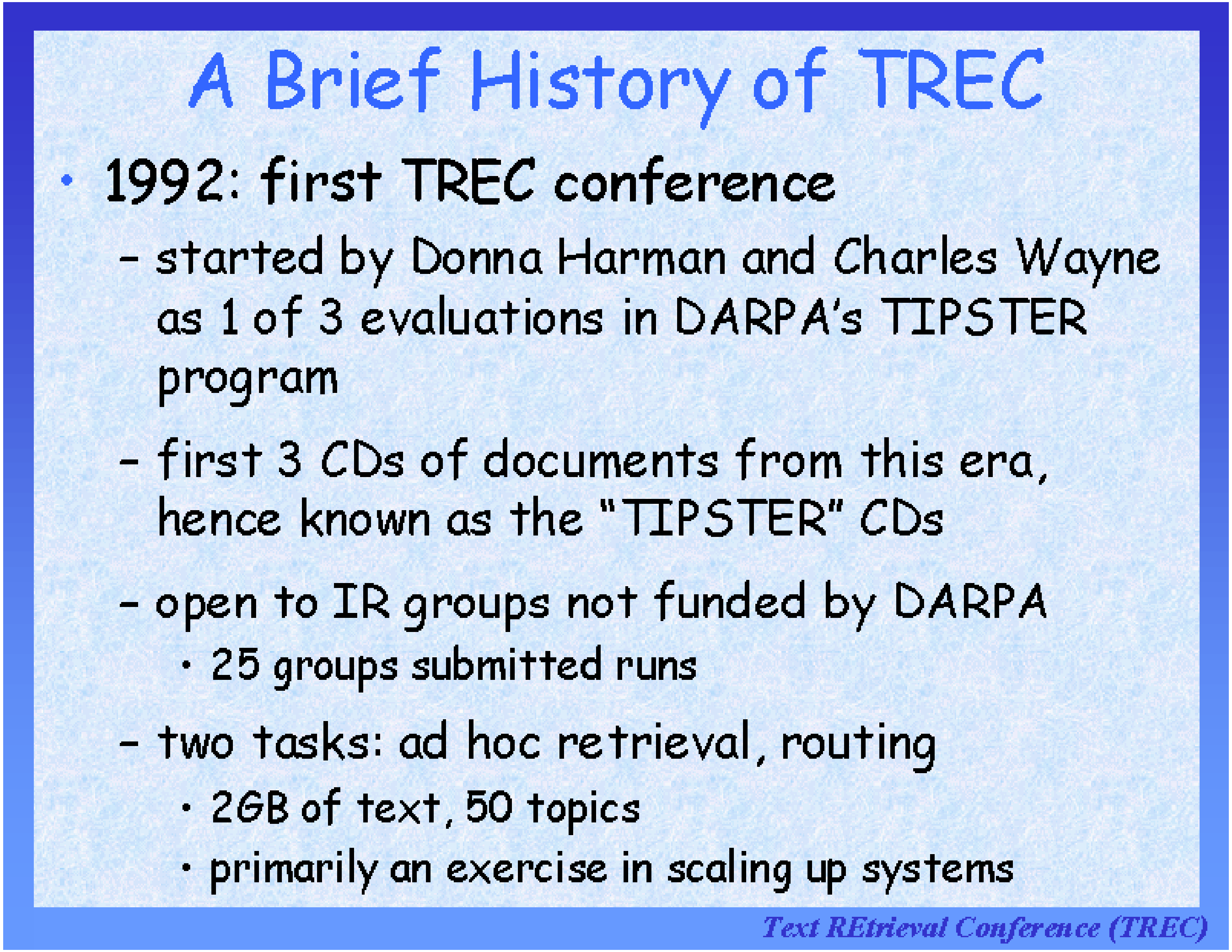
The TREC Collection

- Text REtrieval Conference (TREC)
  - Established in 1991, co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA)
    - Evaluation of large scale IR problems
  - Premier Annual conferences held since 1992
    - Most well known IR evaluation setting

http://trec.nist.gov/overview.html

TREC Goals

- To increase research in information retrieval based on large-scale collections
- To provide an open forum for exchange of research ideas to increase communication among academia, industry, and government
- To facilitate technology transfer between research labs and commercial products
- To improve evaluation methodologies and measures for text retrieval
- To create a series of text collections covering different aspects of text retrieval



A Brief History of TREC

- 1992: first TREC conference
  - started by Donna Harman and Charles Wayne as 1 of 3 evaluations in DARPA's TIPSTER program
  - first 3 CDs of documents from this era, hence known as the "TIPSTER" CDs
  - open to IR groups not funded by DARPA
    - 25 groups submitted runs
  - two tasks: ad hoc retrieval, routing
    - 2GB of text, 50 topics
    - primarily an exercise in scaling up systems

Text REtrieval Conference (TREC)

A Brief History of TREC

- 1993 (TREC-2)
  - true baseline performance for main tasks
- 1994 (TREC-3)
  - initial exploration of additional tasks in TREC
- 1995 (TREC-4)
  - official beginning of TREC track structure
- 1998 (TREC-7)
  - routing dropped as a main task, though incorporated into filtering track
- 2000 (TREC-9)
  - ad hoc main task dropped; first all-track TREC

TREC - Test Collection and Benchmarks

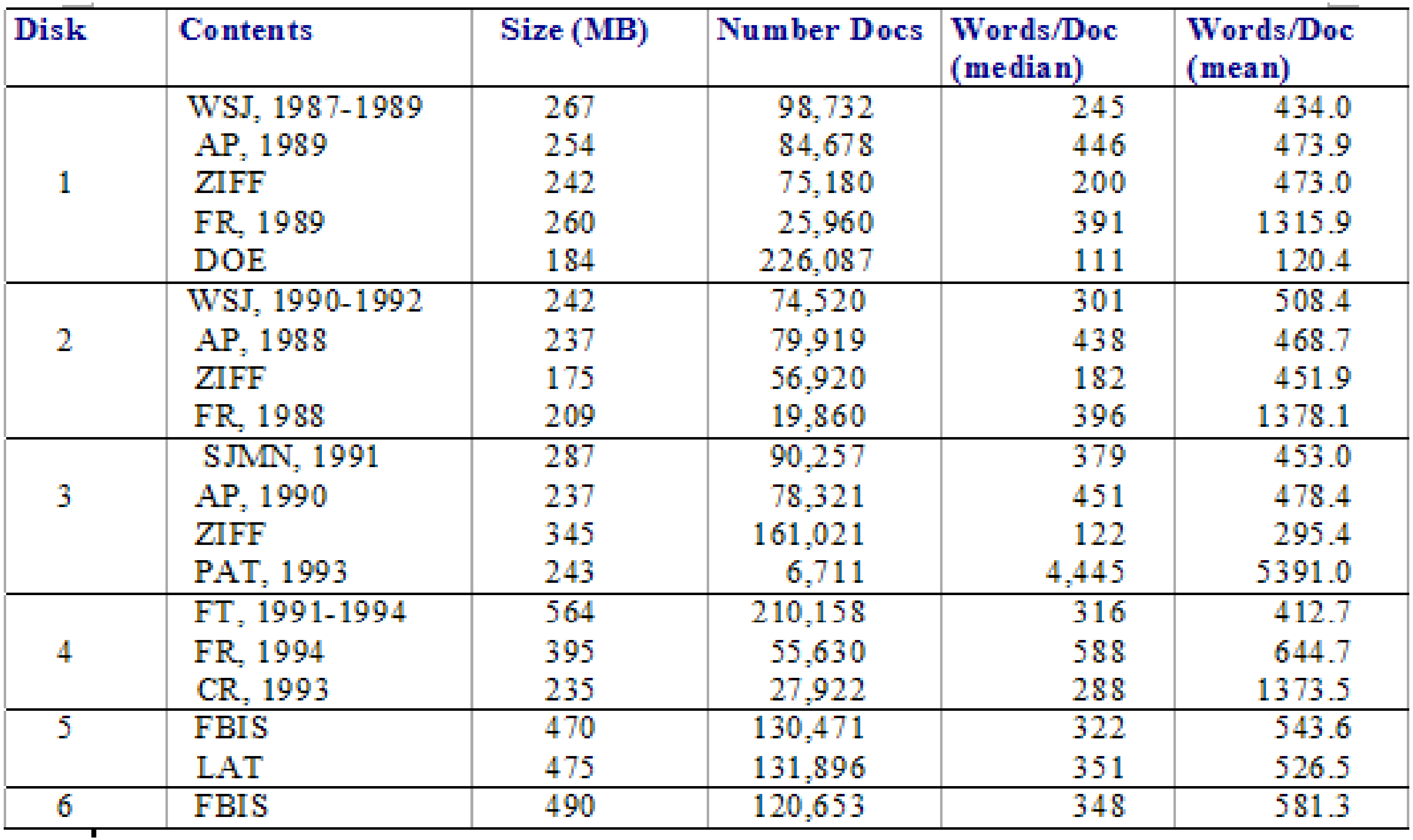
- TREC test collection consists
  - The documents
  - The example information requests/needs (called **topics** in the TREC nomenclature)
  - A set of relevant documents for each example information request

Training/Development
Evaluation
collections

- Benchmark Tasks
  - Ad hoc task
    - New queries against a set of static docs
  - Routing task
    - Fixed queries against continuously changing doc
    - The retrieved docs must be ranked
  - Other tasks started from TREC-4

TREC - Document Collection

- Example: TREC-6



Disk	Contents	Size (MB)	Number Docs	Words/Doc (median)	Words/Doc (mean)
1	WSJ, 1987-1989	267	98,732	245	434.0
	AP, 1989	254	84,678	446	473.9
	ZIFF	242	75,180	200	473.0
	FR, 1989	260	25,960	391	1315.9
	DOE	184	226,087	111	120.4
2	WSJ, 1990-1992	242	74,520	301	508.4
	AP, 1988	237	79,919	438	468.7
	ZIFF	175	56,920	182	451.9
	FR, 1988	209	19,860	396	1378.1
3	SJMN, 1991	287	90,257	379	453.0
	AP, 1990	237	78,321	451	478.4
	ZIFF	345	161,021	122	295.4
	PAT, 1993	243	6,711	4,445	5391.0
4	FT, 1991-1994	564	210,158	316	412.7
	FR, 1994	395	55,630	588	644.7
	CR, 1993	235	27,922	288	1373.5
5	FBIS	470	130,471	322	543.6
	LAT	475	131,896	351	526.5
6	FBIS	490	120,653	348	581.3

TREC - Document Collection

- TREC document example: WSJ880406-0090

```
<doc>
<docno> WSJ880406-0090 </docno>
< hl > AT&T Unveils Services to Upgrade Phone Networks Under Global Plan </hl>
<author> Janet Guyon (WSJ staff) </author>
<dateline> New York </dateline>

<text>
 American Telephone & Telegraph Co. introduced the first of a new generation of
 phone services with broad …
</ text >

</ doc >
```

- Docs are tagged with SGML (Standard Generalized Markup Languages)

TREC Topic Example

<top>

<num> Number: 794

taken as a short query, more typical of a web application

<title> pet therapy

taken as a long query, more typical of a web application

<desc> Description:

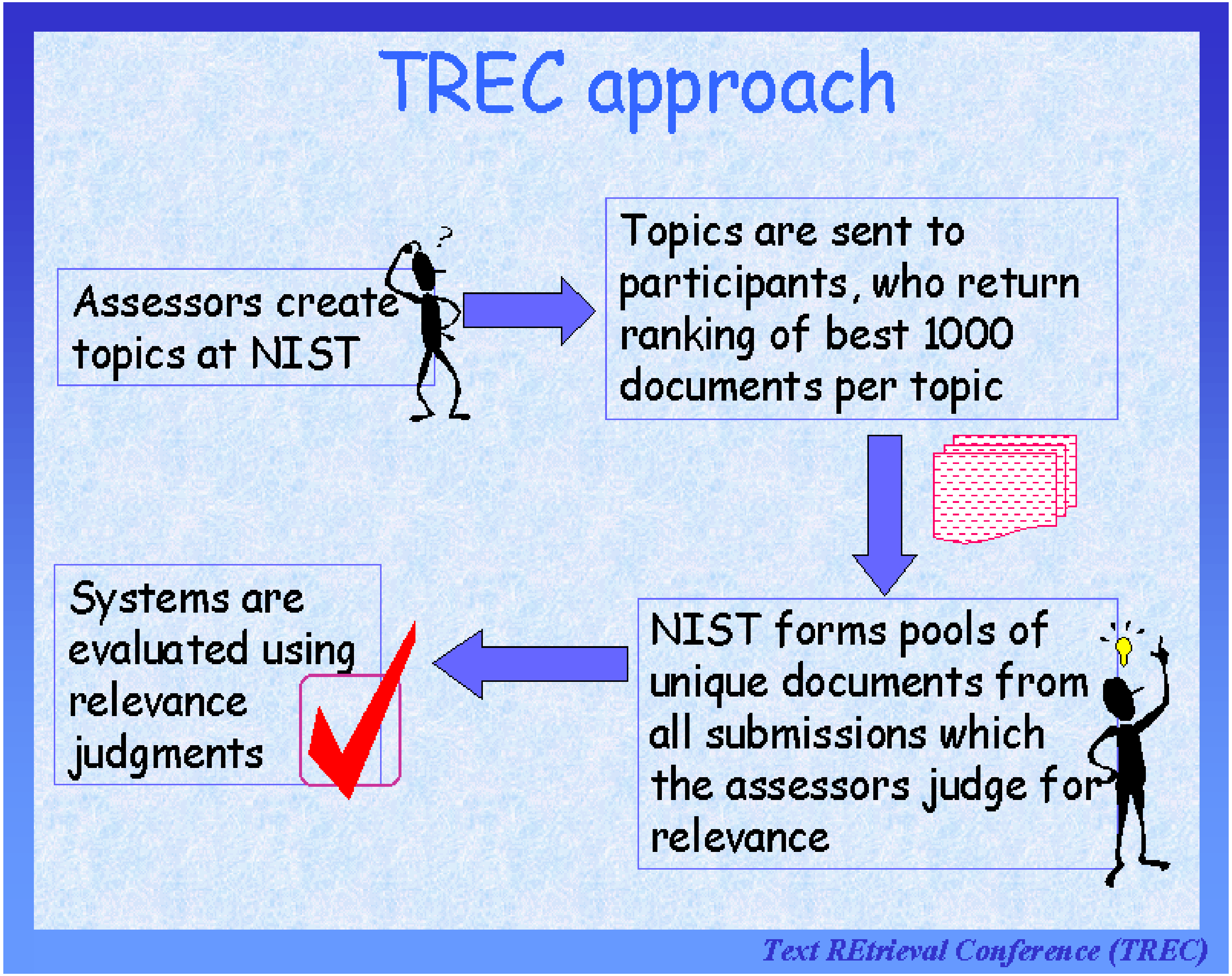
How are pets or animals used in therapy for humans and what are the benefits?

describe the criteria for relevance, used by the people doing relevance judgments, and not taken as a query

<narr> Narrative:

Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

</top>



TREC approach

Assessors create topics at NIST

Topics are sent to participants, who return ranking of best 1000 documents per topic

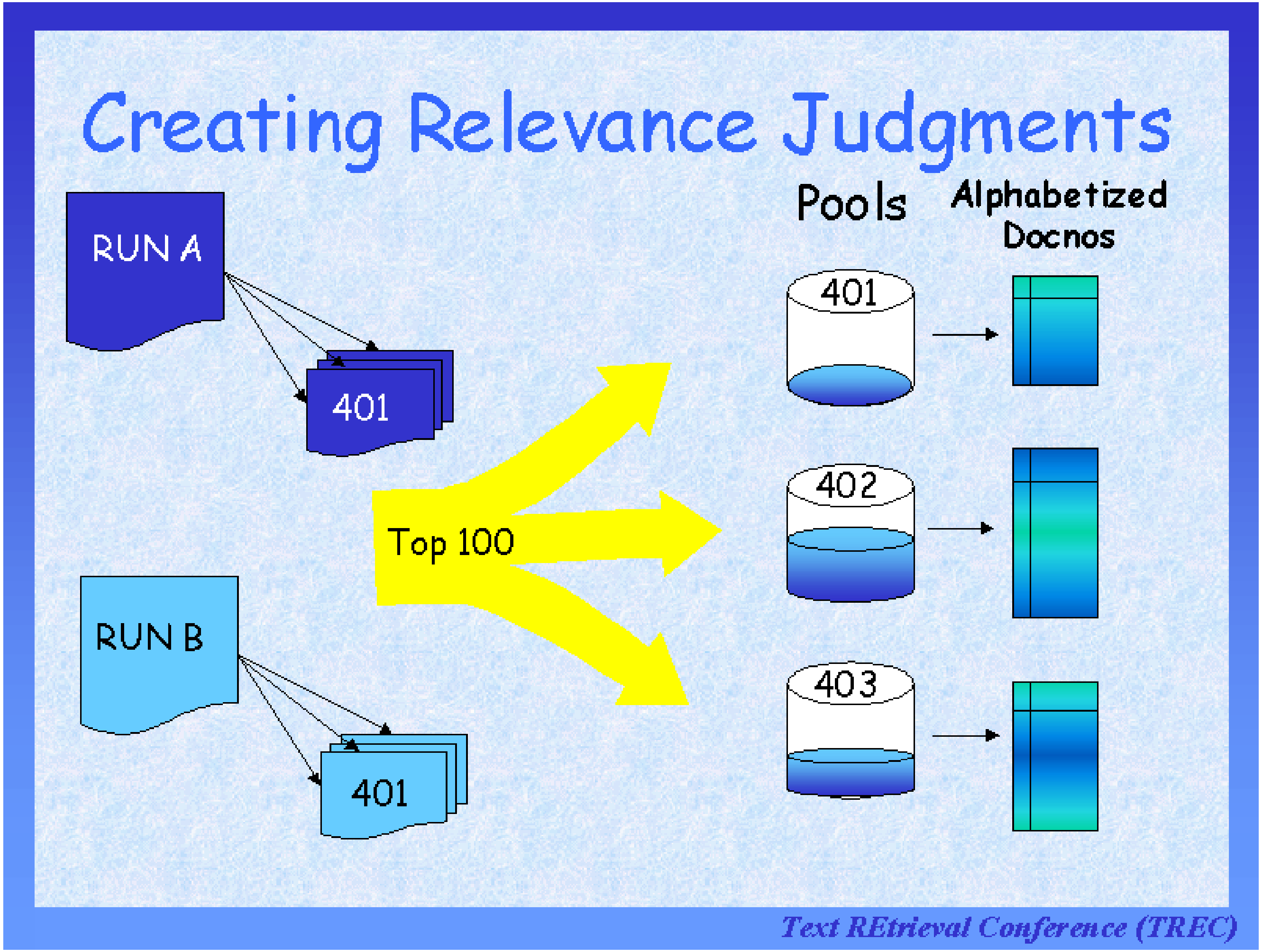
NIST forms pools of unique documents from all submissions which the assessors judge for relevance

Systems are evaluated using relevance judgments

Text REtrieval Conference (TREC)

TREC - Creating Relevance Judgments

- For each topic (example information request)
  - Each participating systems created top *K* (set between 50 and 200) documents and put in a pool
  - Duplicates are removed, while documents are presented in some random order to the relevance judges
  - Human “assessors” decide on the relevance of each document
    - Usually, an assessor judged a document as relevant (most are binary judgments) if it contained information that could be used to help write a report on the query topic
- The so-called “**pooling method**”
  - Two assumptions
    - Vast majority of relevant documents is collected in the assembled pool
    - Documents not in the pool were considered to be irrelevant
  - Such assumptions have been verified to be accurate!

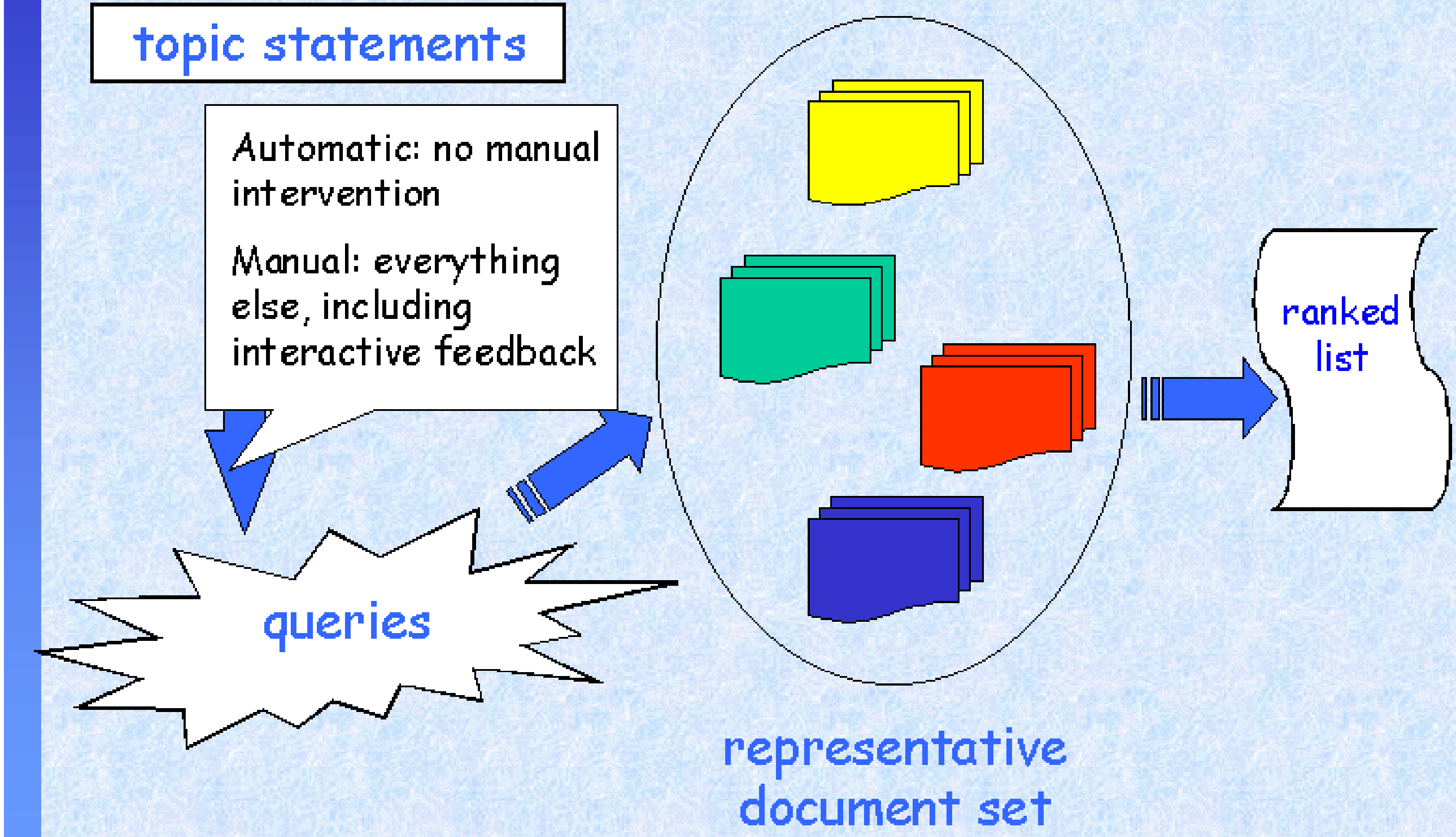


Creating Relevance Judgments

Text REtrieval Conference (TREC)

Text REtrieval Conference (TREC)

Creating a test collection for an ad hoc task



topic statements

Automatic: no manual intervention

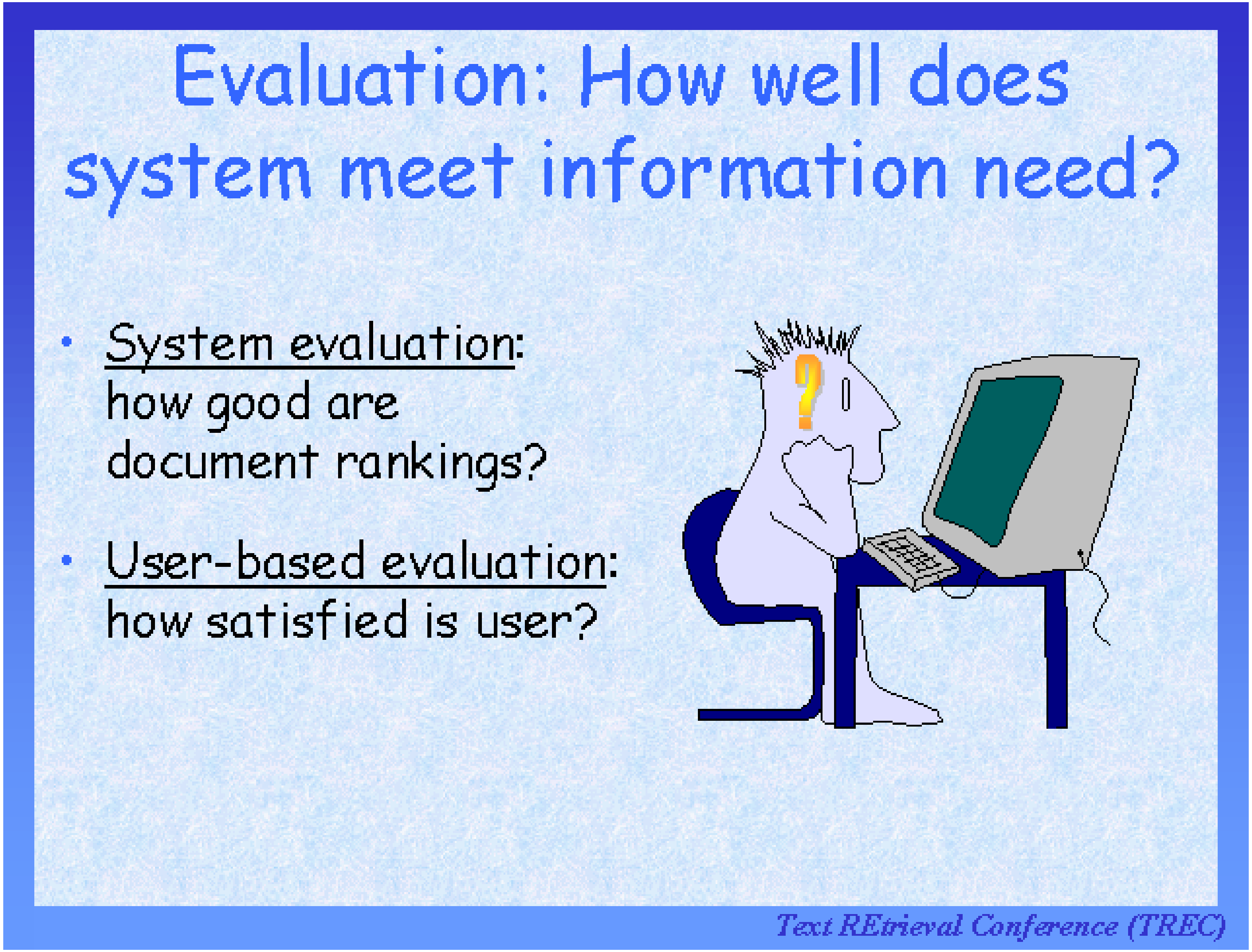
Manual: everything else, including interactive feedback

queries

representative document set

ranked list

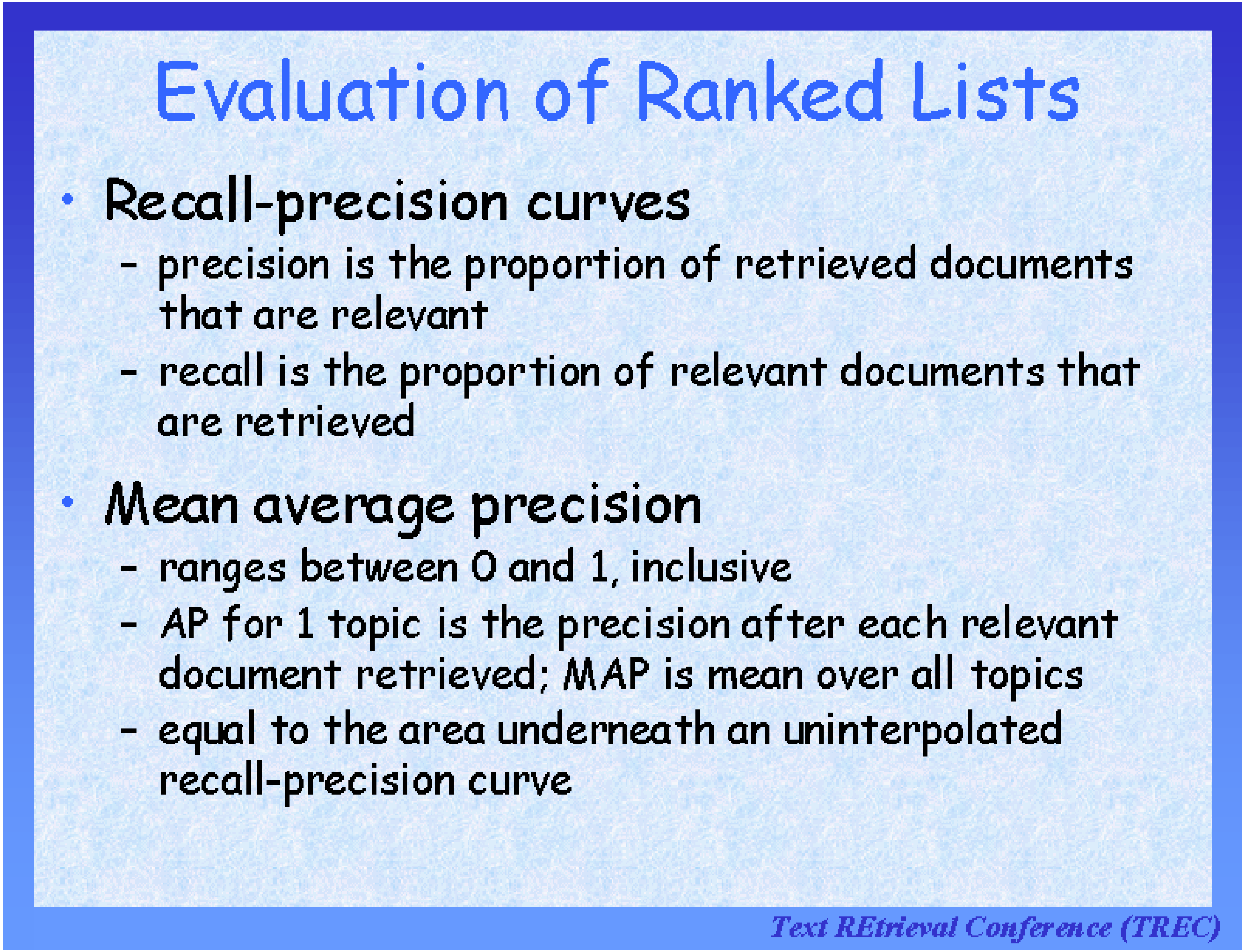
Text REtrieval Conference (TREC)



Evaluation: How well does system meet information need?

- System evaluation: how good are document rankings?
- User-based evaluation: how satisfied is user?

Text REtrieval Conference (TREC)

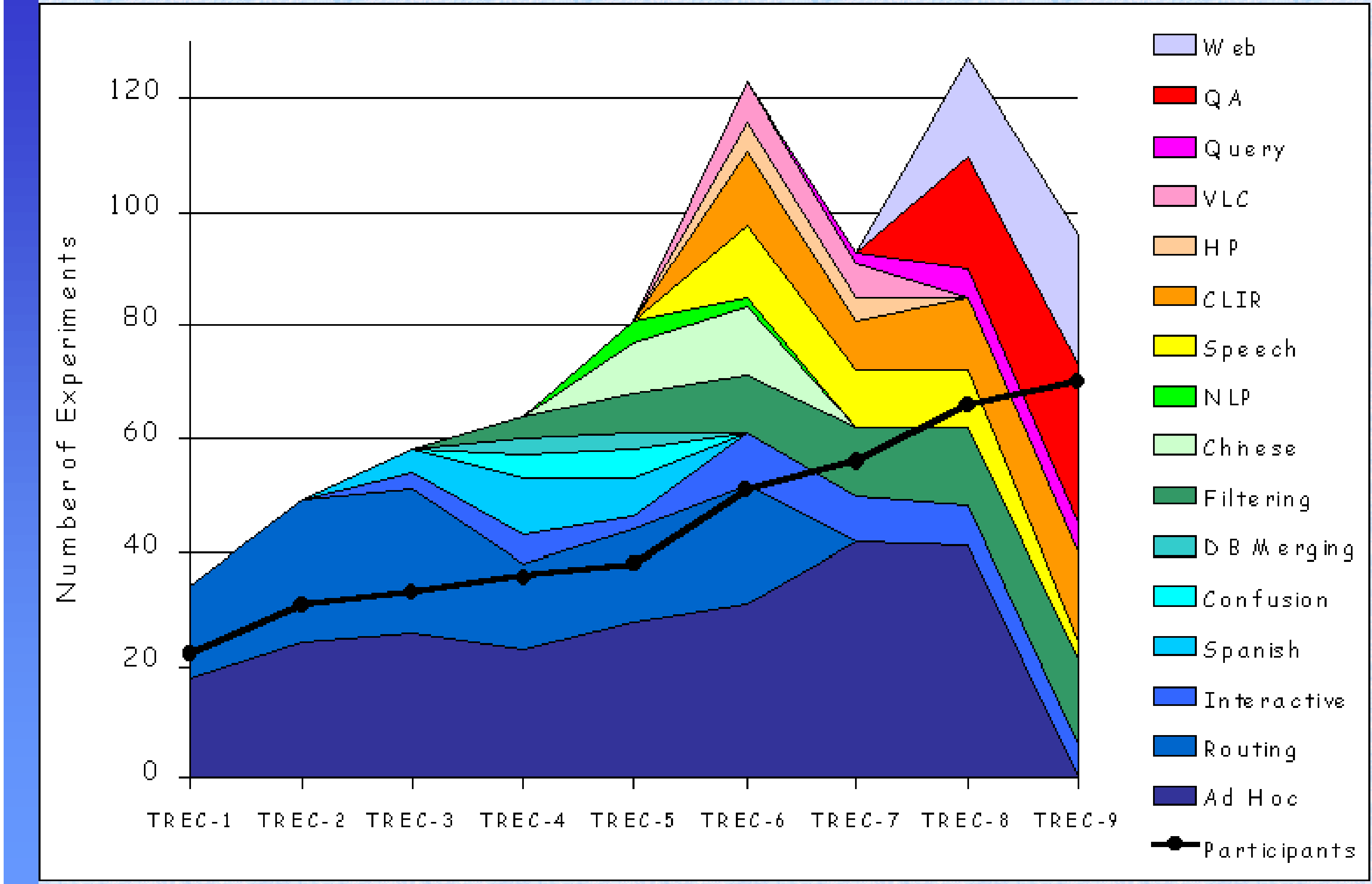


Evaluation of Ranked Lists

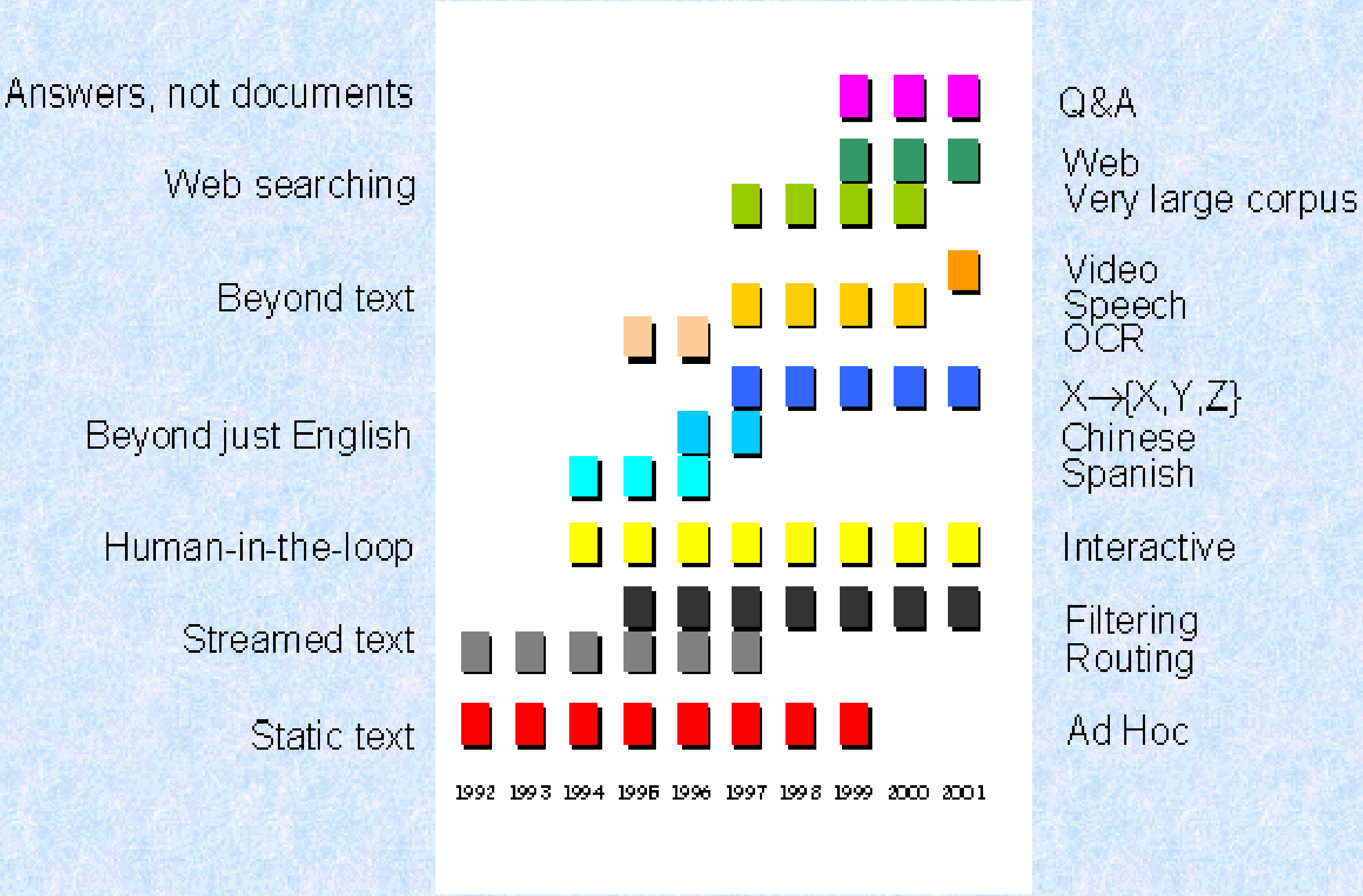
- Recall-precision curves
  - precision is the proportion of retrieved documents that are relevant
  - recall is the proportion of relevant documents that are retrieved
- Mean average precision
  - ranges between 0 and 1, inclusive
  - AP for 1 topic is the precision after each relevant document retrieved; MAP is mean over all topics
  - equal to the area underneath an uninterpolated recall-precision curve

Text REtrieval Conference (TREC)

TREC Tasks

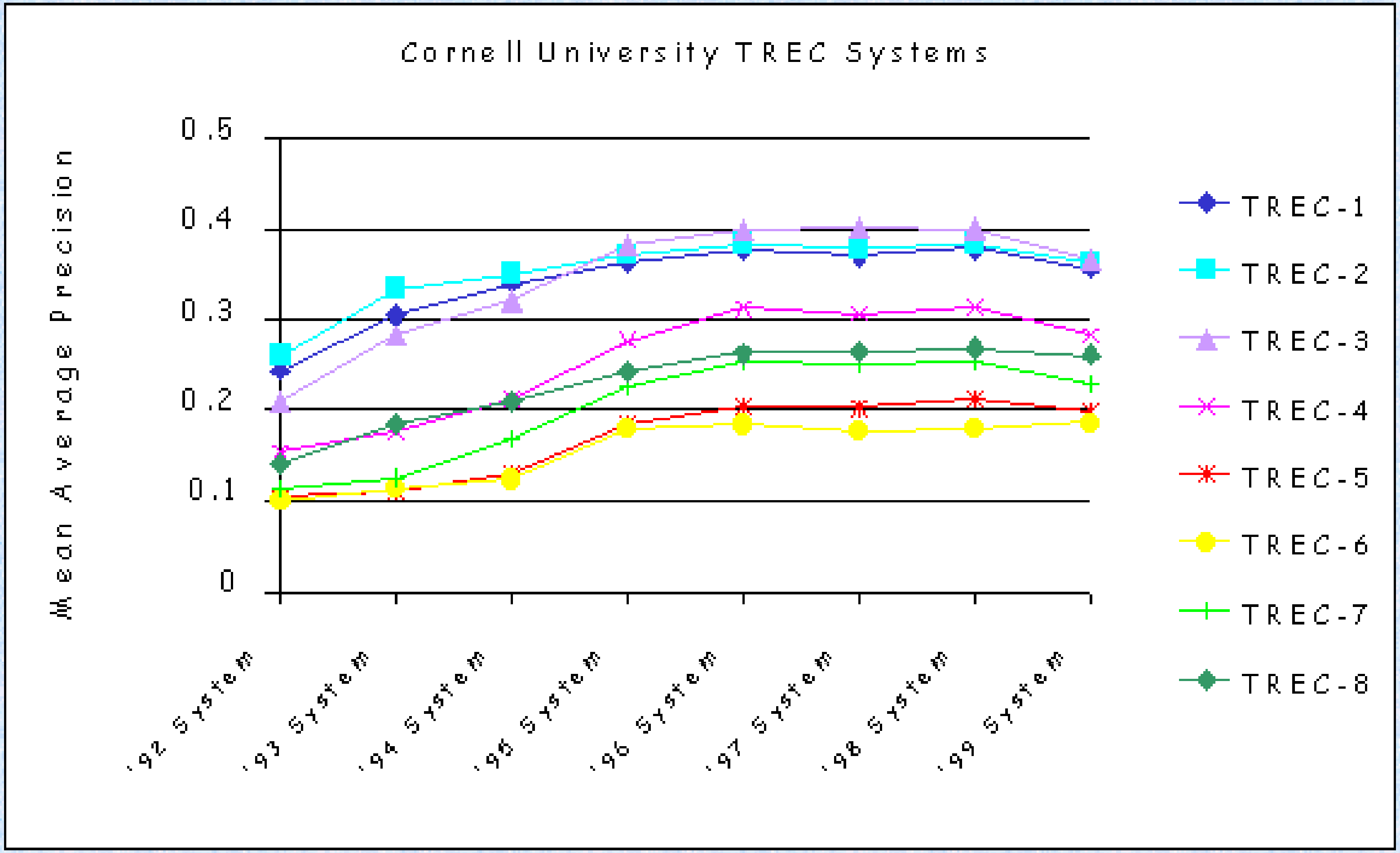

Number of Experiments
0
20
40
60
80
100
120
TREC-1
TREC-2
TREC-3
TREC-4
TREC-5
TREC-6
TREC-7
TREC-8
TREC-9
Web
QA
Query
VLC
HP
CLIR
Speech
NLP
Chinese
Filtering
DB Merging
Confusion
Spanish
Interactive
Routing
Ad Hoc
Participants

TREC Tracks



Text REtrieval Conference (TREC)

TREC Impacts



Text REtrieval Conference (TREC)

TREC – Pros and Cons

- Pros
  - Large-scale collections applied to common task
  - Allows for somewhat controlled comparisons

- Cons
  - Time-consuming in preparation and testing
  - Very long queries, also unrealistic
  - A high-recall search task and collections of news articles are sometimes inappropriate for other retrieval tasks
  - Comparisons still difficult to make, because systems are quite different on many dimensions
  - Also, topics used in every conference year present little overlap , which make the comparison difficult
  - Focus on batch ranking rather than interaction
    - There is an interactive track already

Some Experiences Learned from TREC

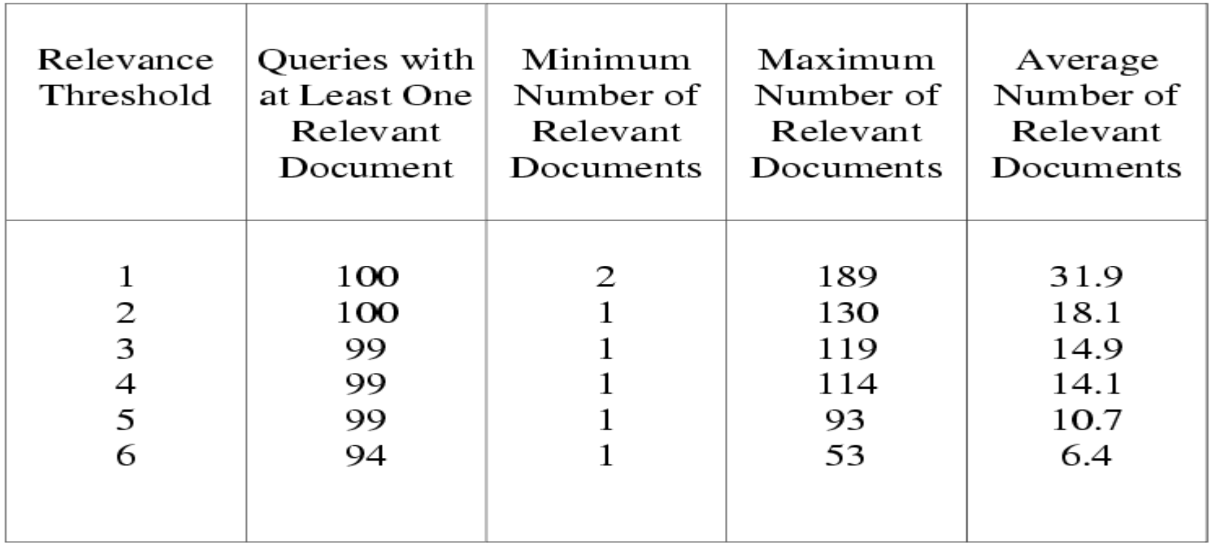
- An analysis of TREC experiments has shown that
  - With 25 queries, an absolute difference in the effectiveness measure *m*AP of 0.05 will results in the wrong conclusion about which system is better is about 13 % of the comparisons
  - With 50 queries, this error rate falls below 4% (which means an absolute difference of 0.05 in *m*AP is quite large)
  - If a significance test is used, a relative difference of 10 % in *m*AP is sufficient to guarantee a low error rate with 50 queries
- If more relevance judgments are made possible, it will be more productive to judge more queries rather than to judge more documents from existing queries
- Though relevance may be a very subjective concept
  - Differences in relevance judgments do not have a significant effect on the error rate for comparisons (*because of "narrative"* ?)

Other Collections

- The CACM Collection
  - 3204 articles (only containing the title and abstract parts) published in the *Communications of the ACM* from 1958 to 1979
  - Topics cover computer science literatures
  - Queries were generated students and faculty of computer science department (Relevance judgment were also done by the same people)
- The ISI Collection
  - 1460 documents selected from a collection assembled at Institute of Scientific Information (ISI)
- The Cystic Fibrosis (CF) Collection
  - 1239 documents indexed with the term “cystic fibrosis” in National Library of Medicine’s MEDLINE database

much human expertise involved

The Cystic Fibrosis (CF) Collection



Relevance Threshold	Queries with at Least One Relevant Document	Minimum Number of Relevant Documents	Maximum Number of Relevant Documents	Average Number of Relevant Documents
1	100	2	189	31.9
2	100	1	130	18.1
3	99	1	119	14.9
4	99	1	114	14.1
5	99	1	93	10.7
6	94	1	53	6.4

- 1,239 abstracts of articles
- 100 information requests in the form of complete questions
  - 4 separate relevance scores for each request
- Relevant docs determined and rated by 3 separate subject experts and one medial bibliographer on 0-2 scale
  - 0: Not relevant
  - 1: Marginally relevant
  - 2: Highly relevant

User Actions as Implicit Relevance Judgments

- Query logs that capture user interactions with a search engine have become an extremely important resource for web search engine development
- Many user actions can also be considered implicit relevance judgments
  - If these can be exploited, we can substantially reduce the effort of constructing a test collection
  - The following actions (i.e., clickthrough data) to some extent may indicate the relevance of a document to a query
    - Clicking on a document in a result list
    - Move a document to a folder
    - Send a document to a printer, etc.
- *But how to maintain the privacy of users ?*

More on Clickthrough Data

- May use clickthough data to predict preferences between pairs of documents (high correlation with relevance)
  - Appropriate for tasks with multiple levels of relevance (graded relevance), focused on user relevance (rather than purely topical relevance)
  - Clickthough data can also be aggregated to remove potential noise and individual differences
- *Skip Above and Skip Next*
  - click data

$$\begin{array}{l} d_1 \\ d_2 \\ d_3 \ \text{(clicked)} \\ d_4 \end{array}$$

  - generated preferences

$$\begin{array}{l} d_3 > d_2 \\ d_3 > d_1 \\ d_3 > d_4 \end{array}$$

Preference: documents with more relevance should ne ranked higher.