Retrieval Evaluation

- Reference Collections

Berlin Chen 2005

References:

1. *Modern Information Retrieval*. Chapter 3
2. *Text REtrieval Conference*. http://trec.nist.gov/

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
  - 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

Text REtrieval Conference (TREC)

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

Text REtrieval Conference (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
- 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

Training/Development collections
Evaluation

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)

Sample Topic

<top>

<num> Number: 451

<title> What is a Bengals cat?

<desc> Description:
Provide information on the Bengal cat breed.

<narr> Narrative:
Item should include any information on the Bengal cat breed, including description, origin, characteristics, breeding program, names of breeders and catteries carrying bengals. References which discuss bengal clubs only are not relevant. Discussion of bengal tigers are not relevant.
</top>

The request is a description of an information need in natural language

Text REtrieval Conference (TREC)

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

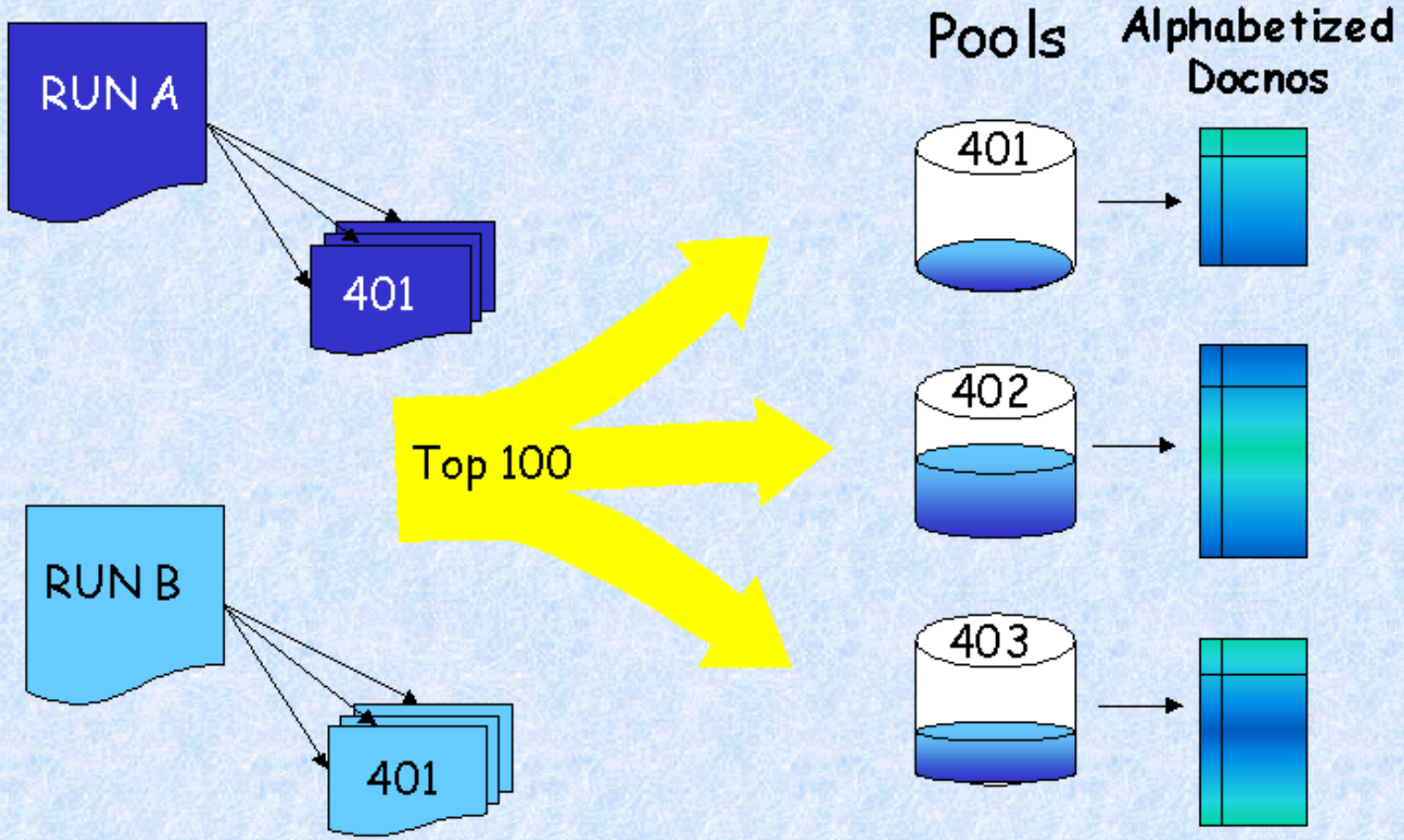
Text REtrieval Conference (TREC)

TREC - Creating Relevance Judgments

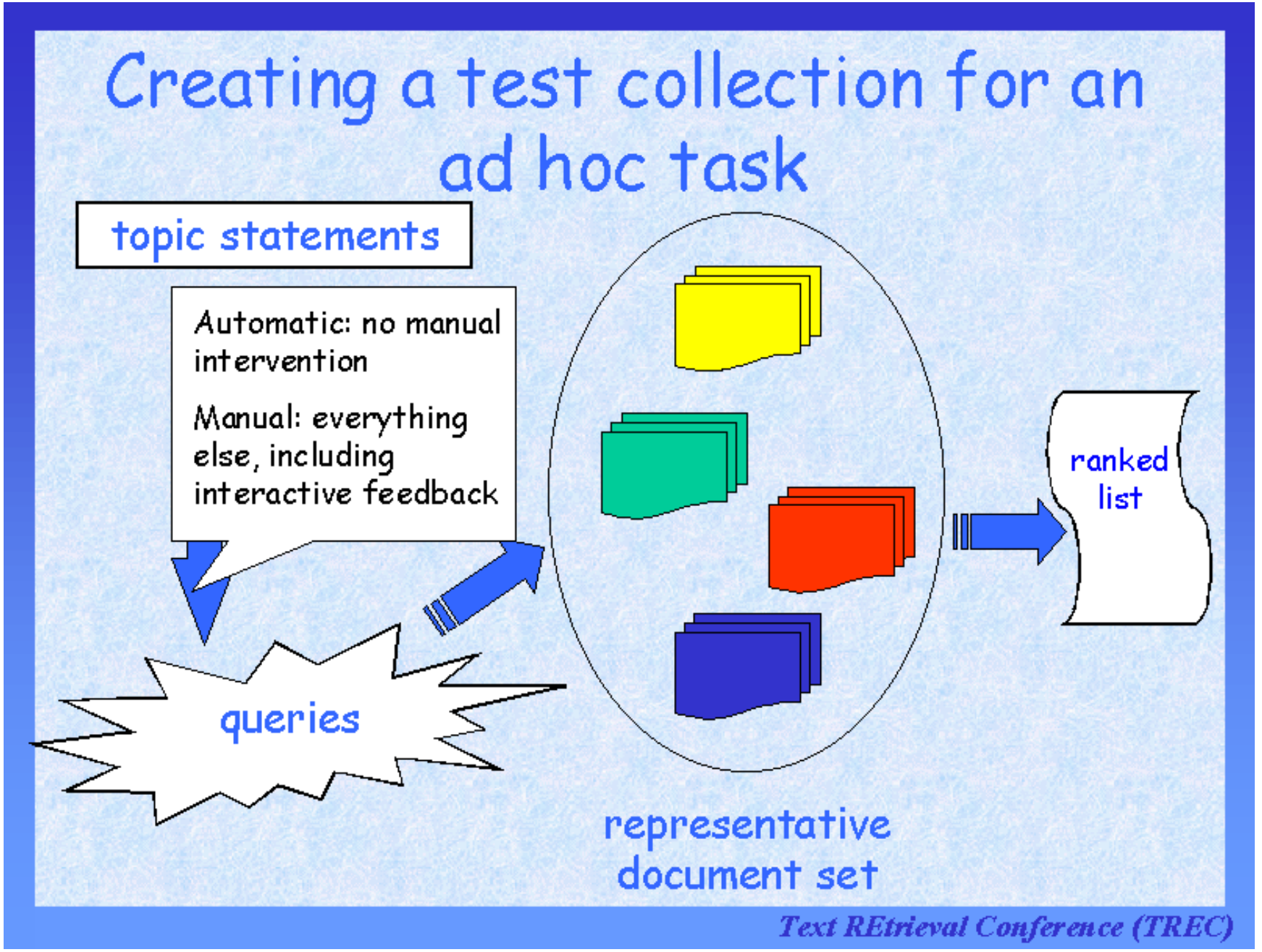
- For each topic (example information request)
  - Each participating systems created top *K* (e.g. *K*=100) docs and put in a pool
  - Human "assessors" decide on the relevance of each doc

- The so-called "**pooling method**"
  - Two assumptions
    - Vast majority of relevant docs is collected in the assembled pool
    - Docs not in the pool were considered to be irrelevant
  - Such assumptions have been verified to be accurate !

Creating Relevance Judgments




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?

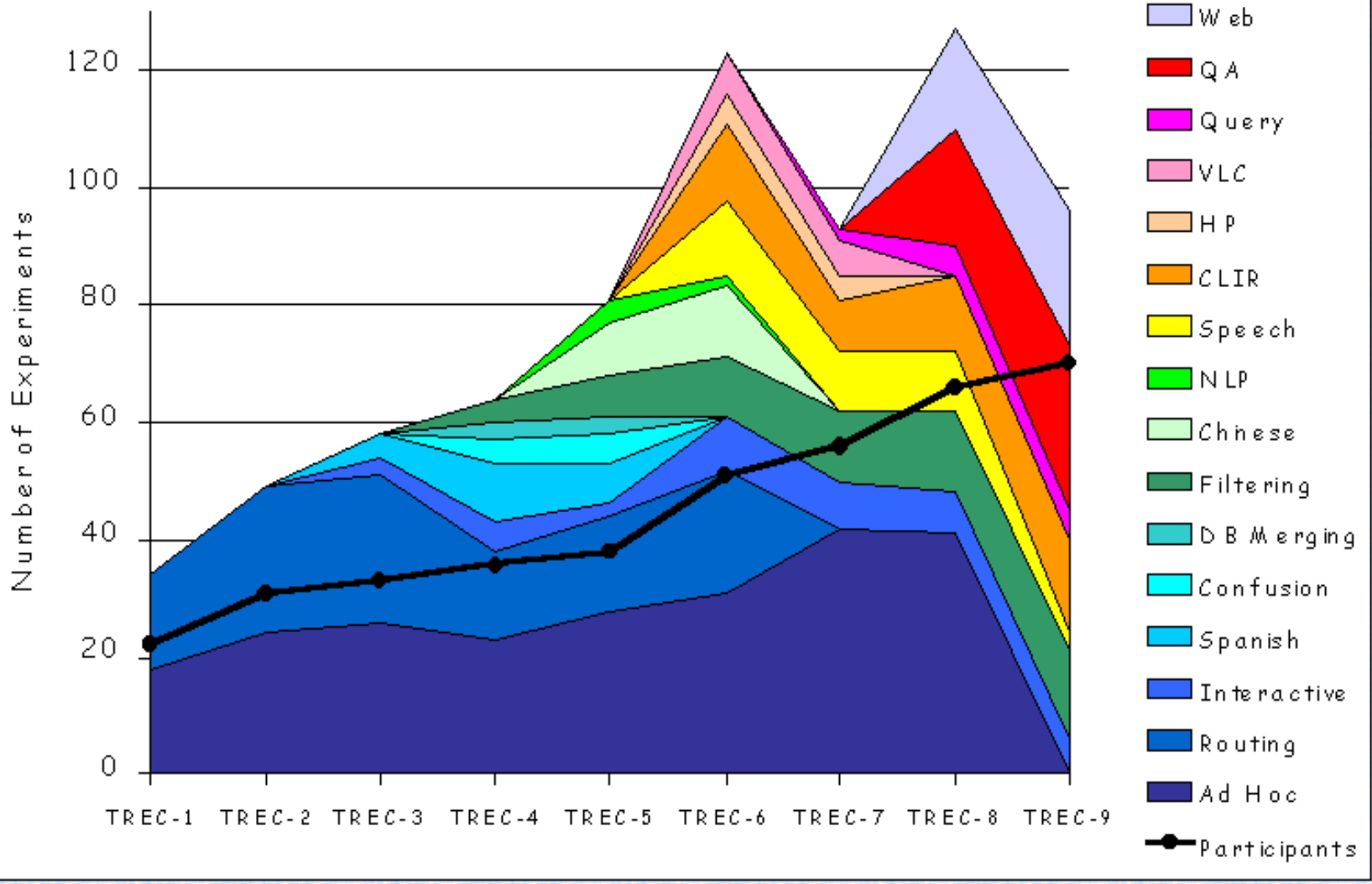
Text REtrieval Conference (TREC)

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

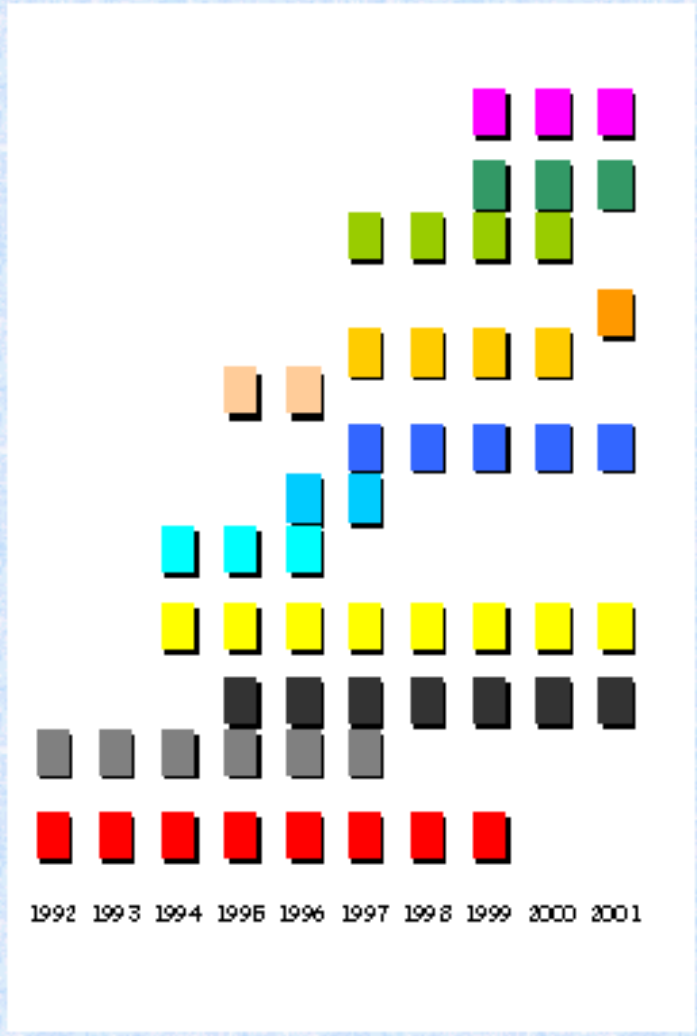
Text REtrieval Conference (TREC)

TREC Tasks



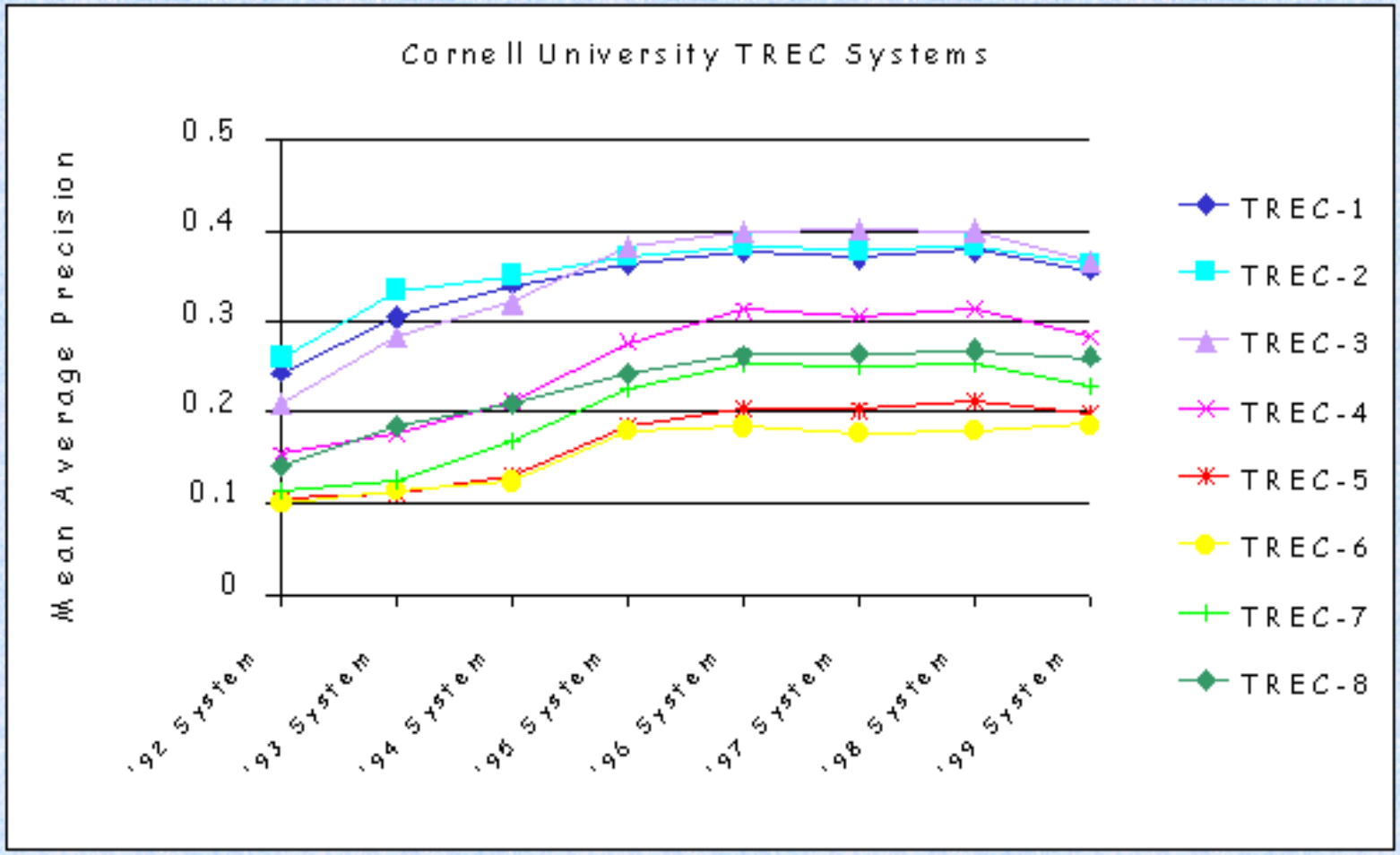
Text REtrieval Conference (TREC)

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
  - 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

Other Collections

- The CACM Collection
  - 3204 articles published in the *Communications of the ACM* from 1958 to 1979
  - Topics cover computer science literatures
- 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