Shared Data and Community Evaluations

Ian Soboroff, NIST

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

A brief history of datasets in IR

- Information Retrieval – the science of search engines.
- In the 1960s, Cyril Cleverdon built what may have been the first dataset for measuring search: “Cranfield”
  - 1400 abstracts from aerospace engineering journal articles.
  - 200 queries taken from random documents.
  - Goal: figure out the best way to index the articles for search. (several styles of fixed vocabularies vs. using the words in the abstract)
- This dataset was a sensation, because now researchers could use a common benchmark.

History, cont.

- From the 60s through the 80s, a number of other datasets were built by different organizations and shared with the research community.
- Each one had its own quirks, and its own bugs.
- And, to a one, they were all quite small, on the order of 10k documents maximum.
- The cost of obtaining documents, labeling them, and processing them made it prohibitive to grow datasets much larger than this.
  - Part of this is 1970s computers, part is labor costs, and part is each group reinventing the process of building the dataset.

Birth of TREC

- In 1991, DARPA asked NIST to build a dataset with around a million documents.

- NIST proposed an open-participation workshop series:
  - NIST would collect and label the data.
  - Participant contributions would help create the dataset by identifying which parts of the data to label.
  - The resulting community could explore the quality of the dataset.
  - In the next year, those lessons would inform the **next** dataset.

What problems did TREC solve?

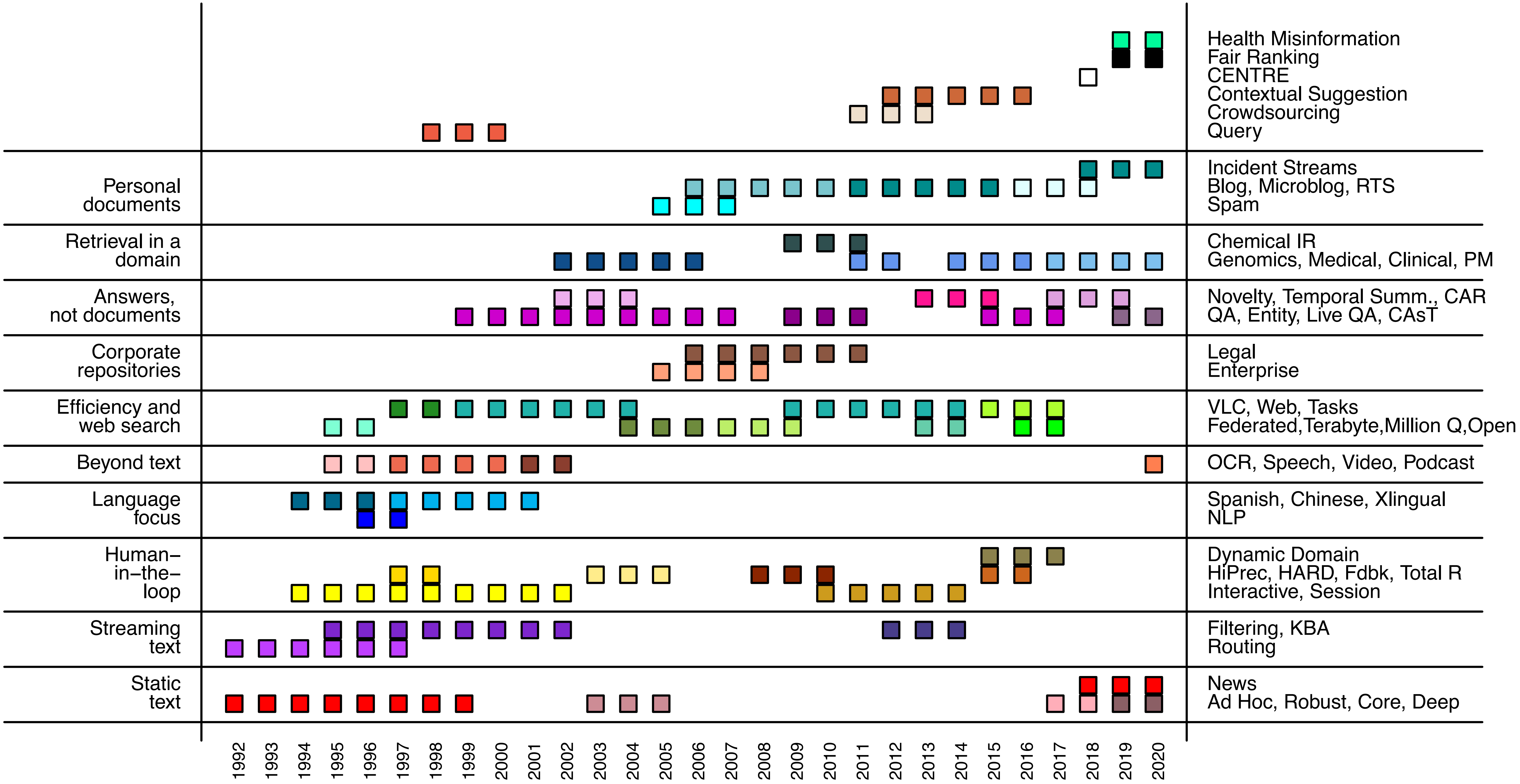
- NIST was able to absorb a cost that was beyond individual research teams, and to make the benefit available to everyone.

- Many eyes made bugs shallow.

- Shared datasets **and** a cycle for improving them.

- The community became involved in the process of creating the datasets, which meant that they were more strongly informed by the needs of the community.

- Together, NIST and the research community were able to standardize methods for building datasets, the experimental methods for using them, and how results would be reported.



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