INFORMATION RETRIEVAL

Dr. Reda M. Hussien

What is information retrieval?

Information Retrieval (IR) is the scientific discipline that studies computer-based search tools.

What is information retrieval?

Mission

- “Organize the world’s information and make it universally accessible and useful.”

What other organizations have this mission?

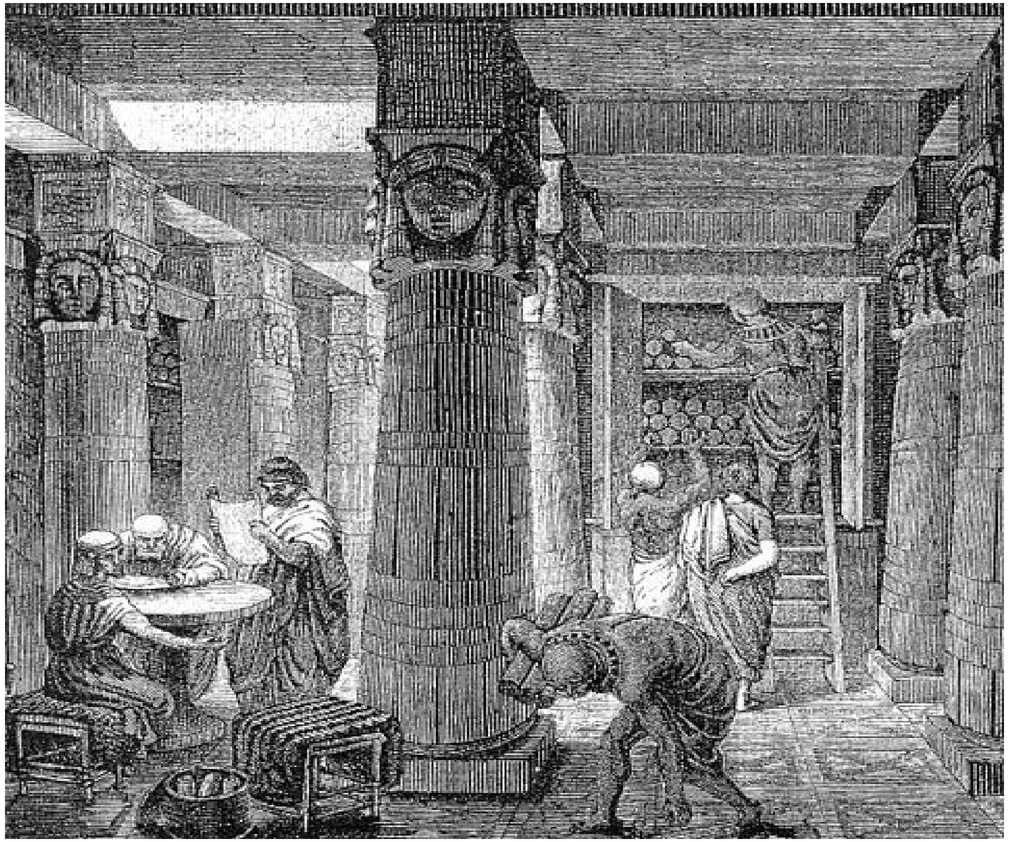
- Libraries
- Scopus,
- Web of Science
- Twitter
- Facebook ?
- Netflix ?
- Amazon ?
- iTunes
- Spotify
- Medium
- U. Twente Search

Google

- (Google books)
- (Google Scholar)
- (Google Plus)
- (Google’s YouTube)
- (Google shopping)
- (Google Play Music)
- (Google Blogger)
- (Google Custom search)

A history of “organizing the world’s info”

- **pre-history of IR**
  - The Library of Alexandria
    - Built: 3rd century BC by Ptolemy I
    - Over 400,000 Papyrus scrolls
    - Visited by a.o. Euclid, Archimedes, …
    - Burned down as Romans conquested Greeks/Egypt



A history of “organizing the world’s info”

- How did Archimedes find the right (relevant) scroll among 400,000 Papyrus scrolls ?

Look at all scrolls?
Randomly look…?

Ask the librarian?

A history of “organizing the world’s info”

- Callimachus: poet, critic and scholar at the Library of Alexandria
- Made the Pinakes: considered to be the first library catalog.
  - It divided works in:
    - genres & categories:
      - rhetoric, law, epic, tragedy, comedy, lyric poetry, history, medicine, mathematics, natural science, miscellanies, …
    - each category was alphabetized by author.

Pre-history: standards

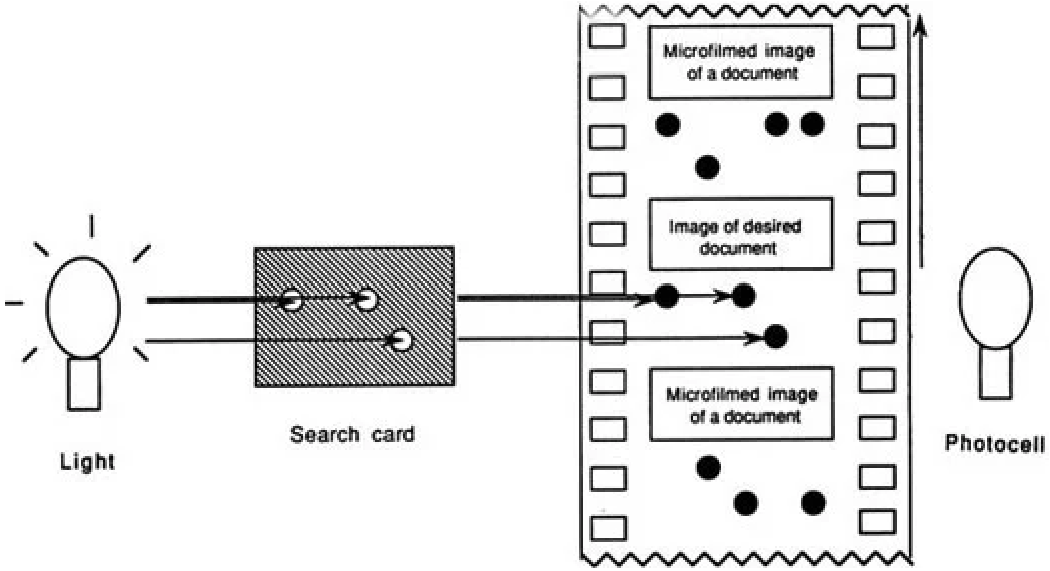
- Melvil Dewey's Decimal Classification (1876)
  - Hierarchical numbering scheme made up of ten classes, each divided into ten divisions, each having ten sections.
  - [List of Dewey Decimal classes]

MELVIL DEWEY.

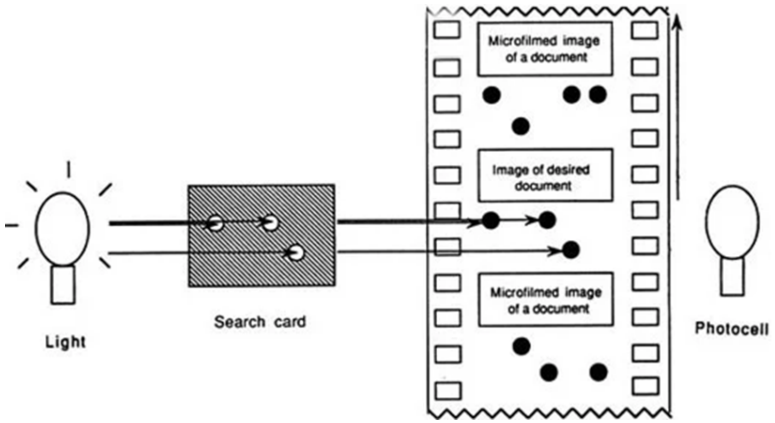
Pre-history: first machines

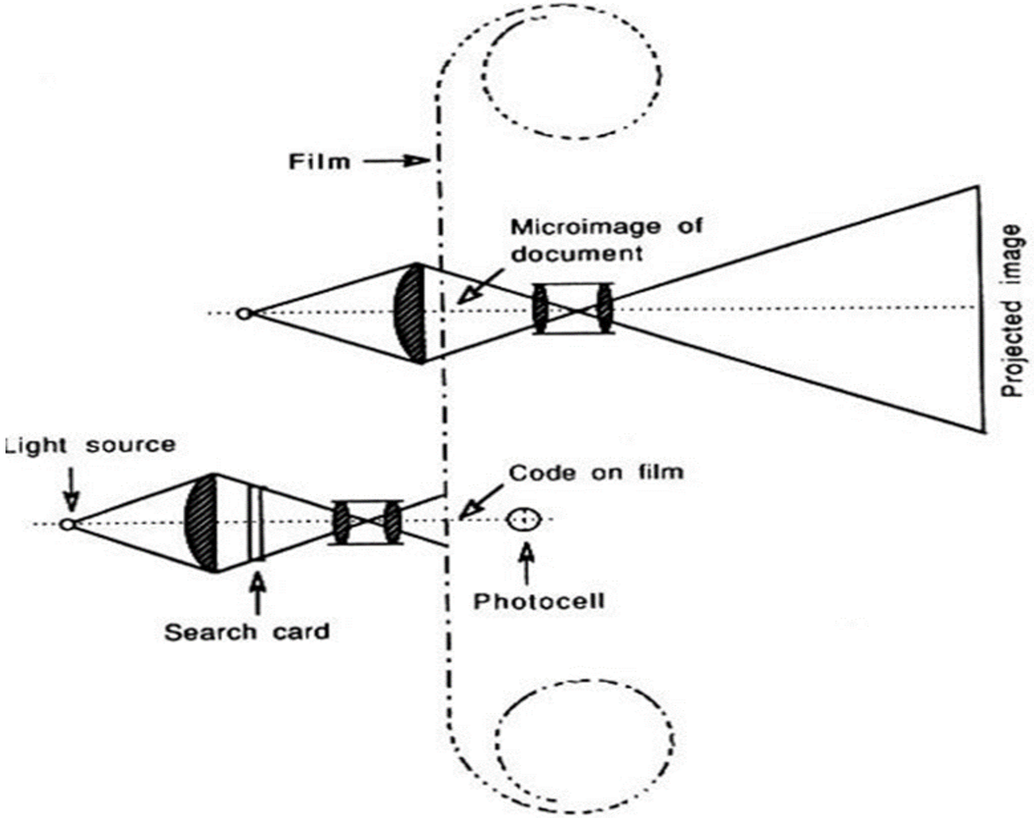
- Emanuel Goldberg Microfilm Search

  “Statistical Machine” (patent 1931)



Pre-history: first machines


Microfilmed image of a document
Image of desired document
Microfilmed image of a document
Light
Search card
Photocell



History: first machines

- Calvin Mooers coined the name “Information Retrieval” (1950)
- “The problem under discussion here is machine searching and retrieval of information from storage according to a specification by subject... It should not be necessary to dwell upon the importance of information retrieval before a scientific group such as this for all of us have known frustration from the operation of our libraries – all libraries, without exception.”

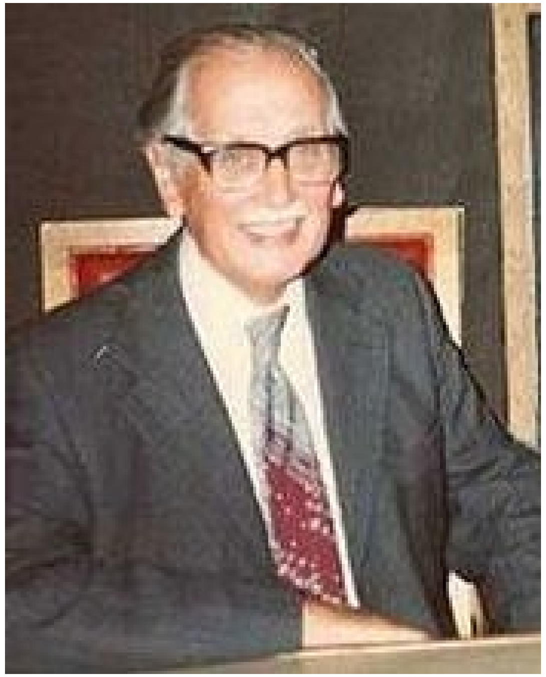
History: standards

- Mortimer Taube (1952)
- “Unit terms”: a proposal to index items by a list of keywords.

1910-1965

History: evaluation

- Cyril Cleverdon (1960s)
- First empirical evaluation of information retrieval systems
  - Measures: Precision & Recall
  - Showed that using all keywords from abstract outperform manual indexing



History: ranking

Hans Peter Luhn (1957)
Similarity based in term frequencies (tf)

Karen Sparck-Jones (1972)
Specificity based on inverse document frequency (idf)

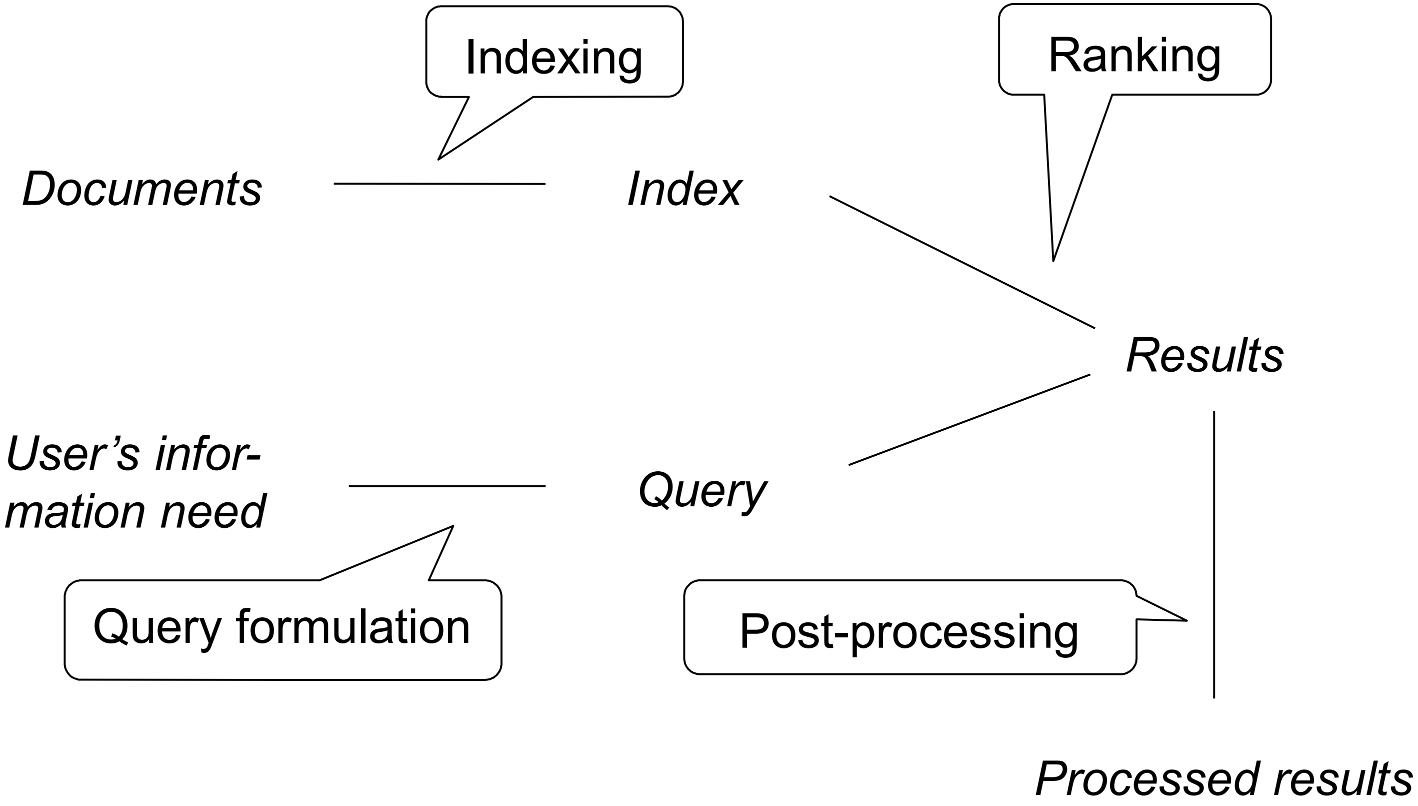
Gerard Salton (1975)
based on tf x idf

Keith van Rijsbergen (1975)
Information Retrieval: first popular scholarly book

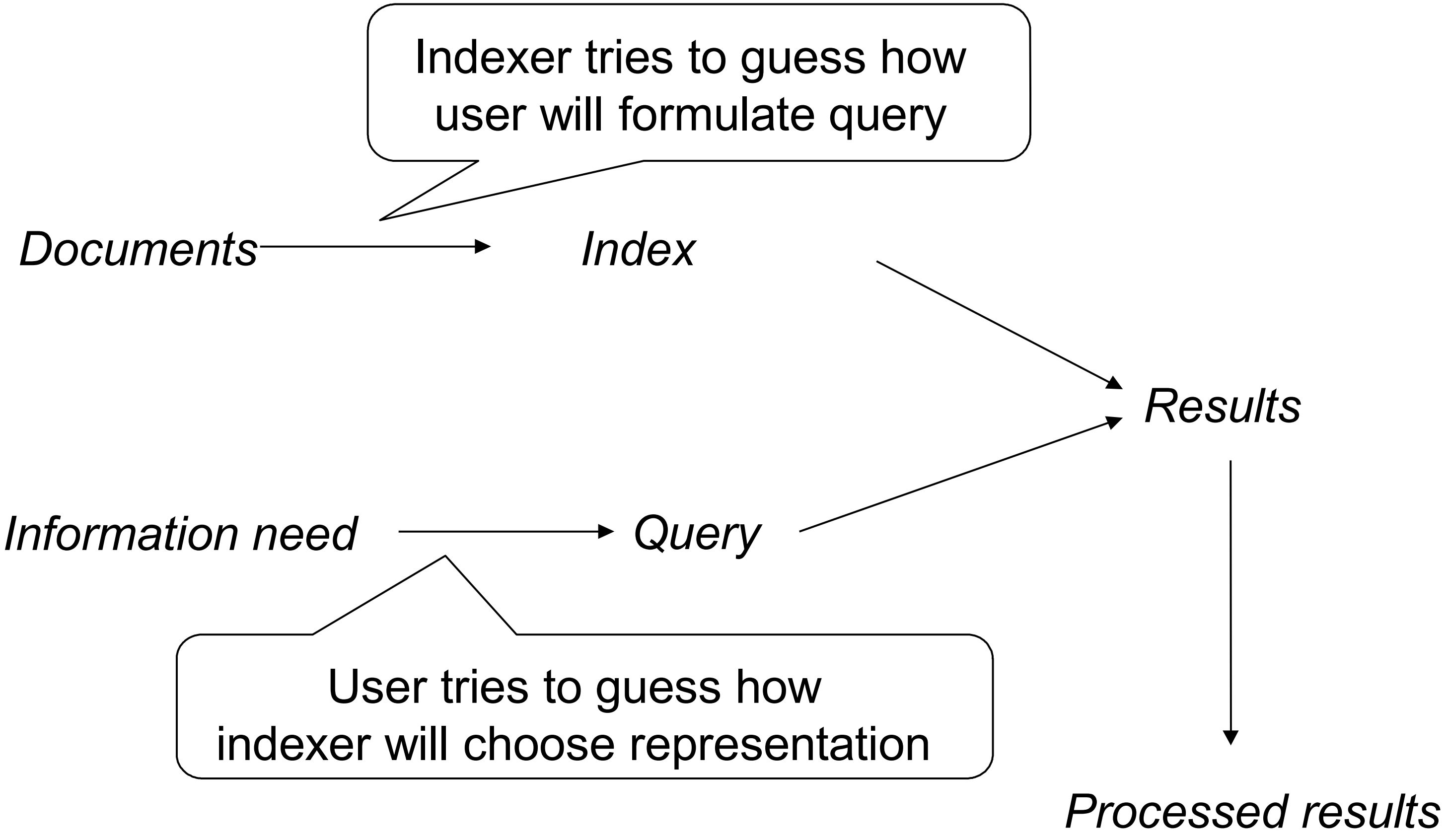
What is information retrieval?

- General characteristics:
  - Users with an information need
  - Documents
    - provide information, and (units part of bigger sources: sections, videos, scenes)
  - A connection between the two

Graphical representation of IR



The prediction game



Indexer tries to guess how user will formulate query

Documents → *Index*

Results

Information need → *Query*

User tries to guess how indexer will choose representation

Processed results

Another view

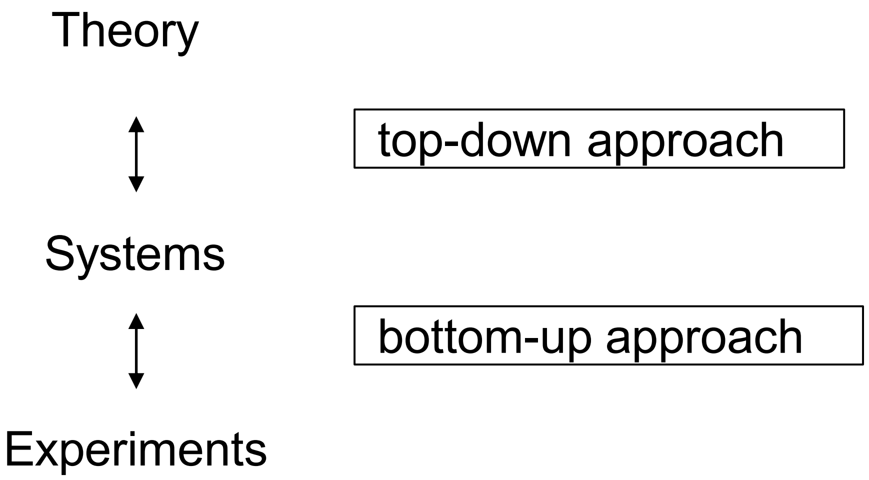
- Information retrieval is search for similarity:
  - between a document and a query
  - between documents in a collection (clustering)
  - between users (collaborative filtering)

More than text

- Texts
  - journal articles, press releases, WWW pages, …
- Pictures
- Audio
  - music, speeches, sounds for medical or engineering purposes, …
- Video
- Any combination

IR Research

- Research in IR is concerned with the design of better IR systems



Theory

top-down approach

Systems

bottom-up approach

Experiments

Approaches: indexing

- Traditionally, two styles:
  - Manually by trained indexers, taking terms from pre-defined list (thesaurus)
  - Automatically by deriving features like
    - words, word stems, phrases from texts
    - graphical features (colour distribution, texture etc.) from images how about sounds, how about videos, how about smells?

Approaches: query formulation

- Traditionally by hand
- Formulating a good query is difficult!
- Increasing attention to automated aids for query formulation
  - natural-language queries
  - relevance feedback
  - personalization
  - recommender systems

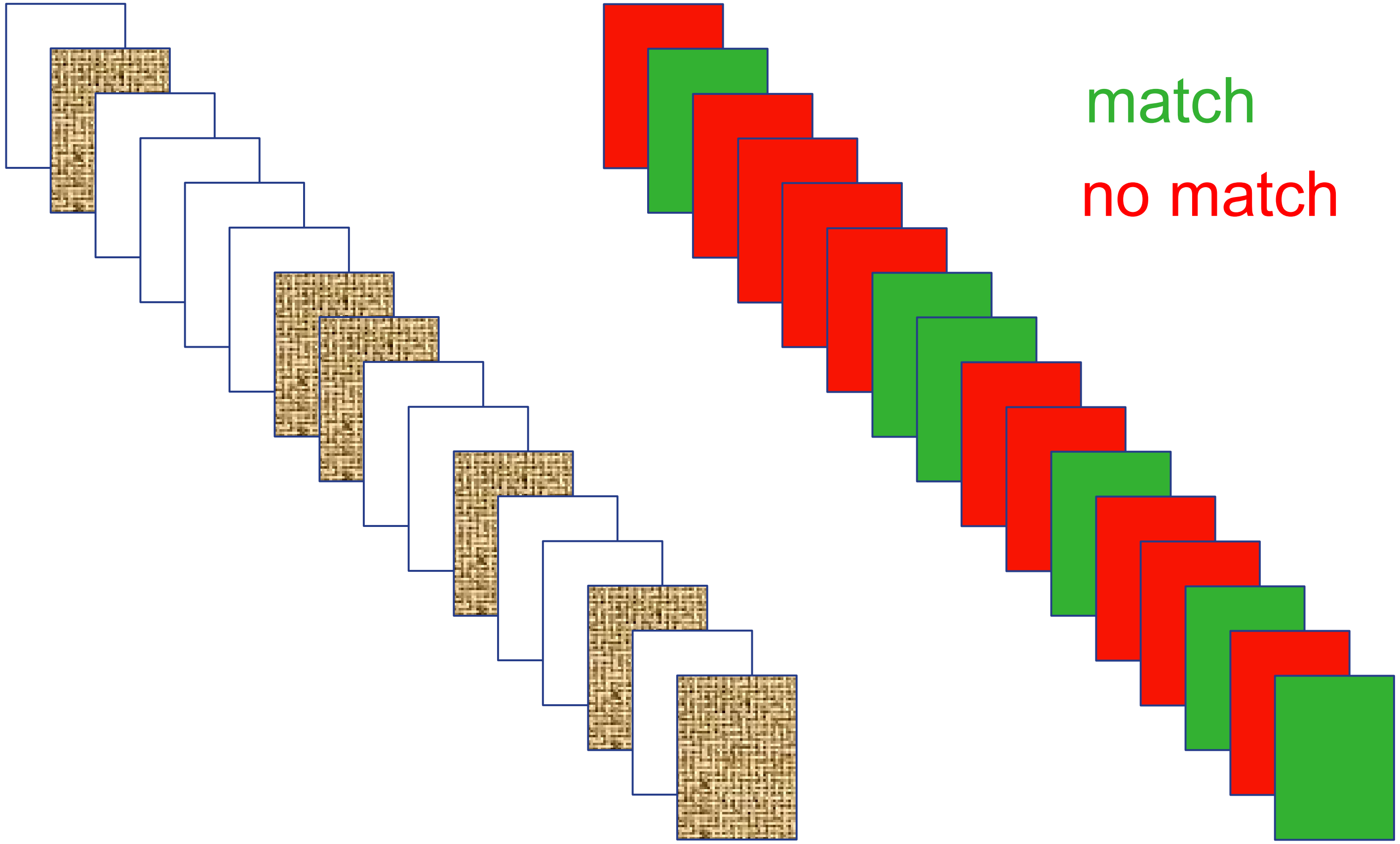
Approaches: query formulation

- Other dimensions:
  - Query in Italian, answer in Dutch
  - Query by example: natural-language fragment, part of a picture
  - Spoken query
  - More expressive query languages (e.g., a description logic)
  - Conversational systems

Approaches: ordering engine

- Two basic approaches:
  - *Matching*: imposes a dichotomy on the collection
  - *Ranking* rank-orders the entire collection
  - The set $\{A, B\}$ is a dichotomy of set $C\ iff\ A \cap B = \emptyset$ and $A \cup B = C$

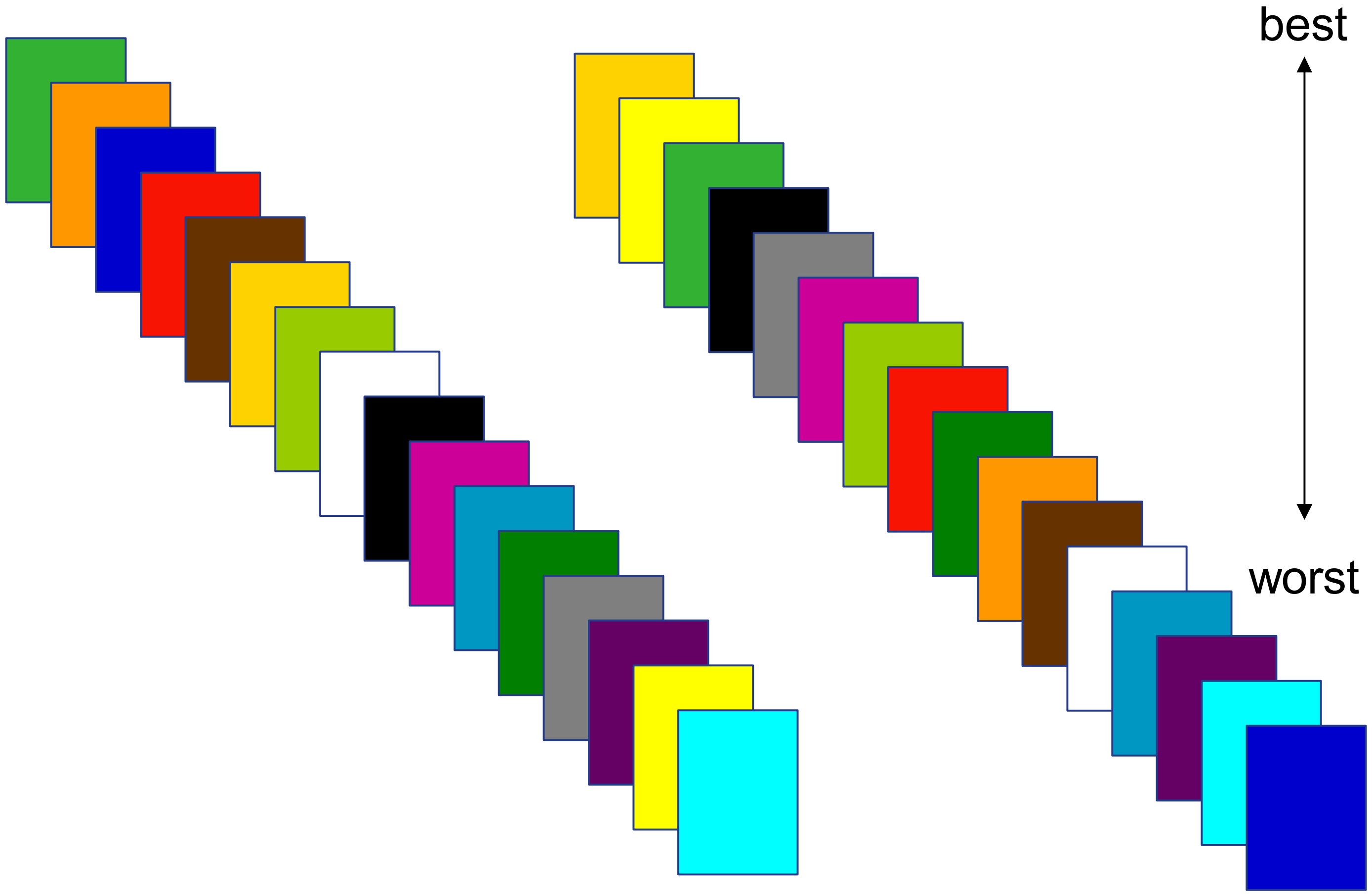
Matching



match

no match

Ranking



Approaches: presentation

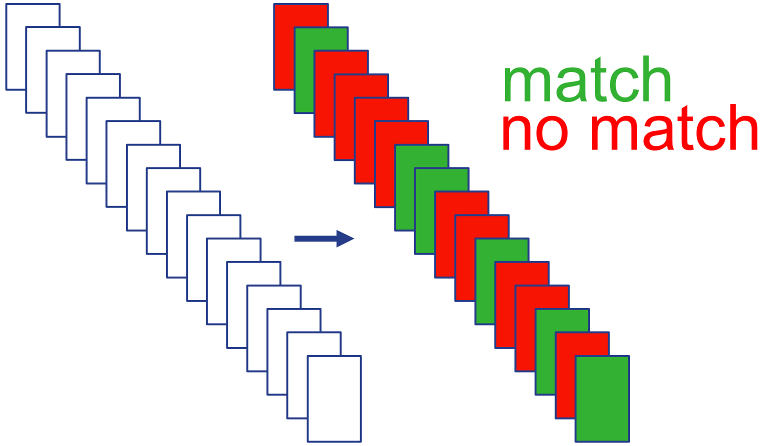
- The item as it is found in the collection
- Part of the document: a section, a paragraph, audio fragment A summary
- An answer to the question you posed (question-answering systems)

Measuring performance

- Theory of measurement in IR is difficult, for example:
  - Which queries are a representative sample of the population of all queries?
  - Does a good measurement mean that the user is satisfied?
  - What about queries that can only be answered by combinations of items?

Performance: matching as example

- Match / no match is a system decision
- Relevant / not relevant is a user decision
- Gives rise to familiar quadrant (compare medical tests)



Performance for matching

System says:

User says:

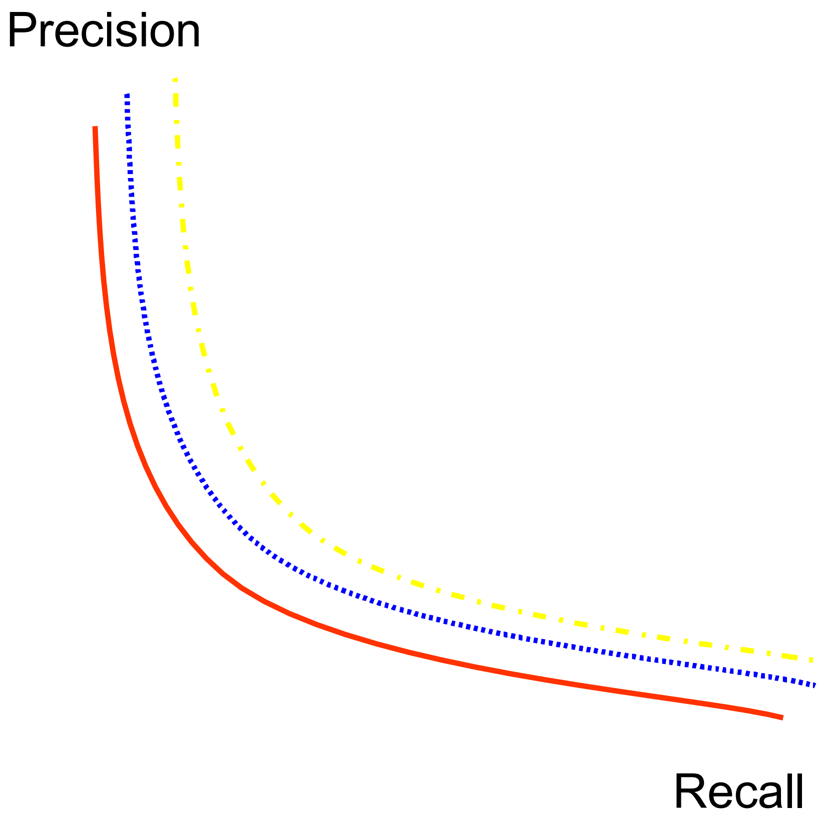
	Match	No match
Relevant	True positives *(#TP)*	False negatives *(#FN)*
Not relevant	False positives *(#FP)*	True negatives *(#TN)*

$$\mathit{Recall} = \frac{\#TP}{\#TP + \#FN}$$

$$\mathit{Precision} = \frac{\#TP}{\#TP + \#FP}$$

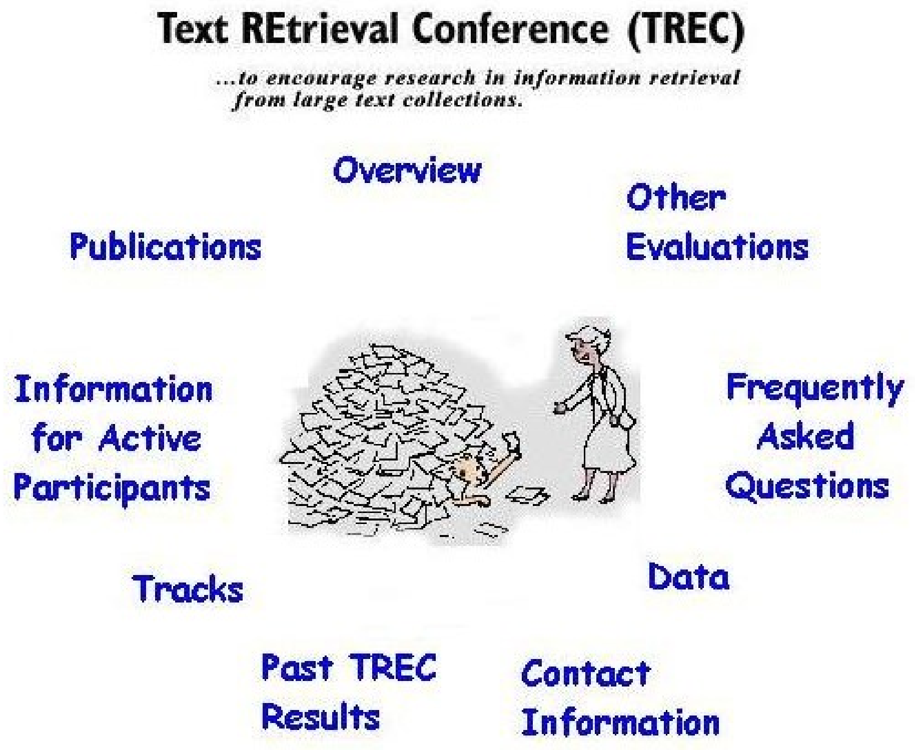
Performance for matching

- “Fact of life”:
  - improving recall typically decreases precision.



Measuring performance: TREC

- Yearly competition, held in November
- Idea: demonstrate your system on unknown queries for a known, very large collection
- System with the best recall-precision performance "wins"



Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

Labs

- Instructor: Amany M. Draz
- Python
- Jupiter Notebook
- PyTerrier:
  - A Python Framework for Information Retrieval

GitHub

PyTerrier