Information Retrieval

Slides are adapted from

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze,

Raymond J. Mooney

Simone Teufel and Ronan Cummins

Theo Huibers, Dolf Trieschnigg and Djoerd Hiemstra

What is Information Retrieval

Information Retrieval

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
  - These days we frequently think first of web search, but there are many other cases:
    - E-mail search
    - Searching your laptop
    - Corporate knowledge bases
    - Legal information retrieval

Taken from: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze

Information retrieval (IR) is finding material . . . of an unstructured nature . . . that satisfies an information need from within large collections . . . .

Taken from: Simone Teufel and Ronan Cummins

Document Collections

MS 3391
Library catalogue. Babylonia, 2000-1600 BC

Taken from: Simone Teufel and Ronan Cummins

Document Collections




IR in the 17th century: Samuel Pepys, the famous English diarist, subject-indexed his treasured 1000+ books library with key words.

Taken from: Simone Teufel and Ronan Cummins

Document Collections

Taken from: Simone Teufel and Ronan Cummins

Information retrieval (IR) is finding material (usually documents) of an unstructured nature . . . that satisfies an information need from within large collections (usually stored on computers).

- Document Collection: text units we have built an IR system over.
- Usually documents
- But could be
  - memos
  - book chapters
  - paragraphs
  - scenes of a movie
  - turns in a conversation...
- Lots of them

Taken from: Simone Teufel and Ronan Cummins

Information retrieval (IR) is finding material (usually documents) of an unstructured nature . . . that satisfies an information need from within large collections (usually stored on computers).

Taken from: Simone Teufel and Ronan Cummins

Structured vs Unstructured Data

Unstructured data means that a formal, semantically overt, easy-for-computer structure is missing.

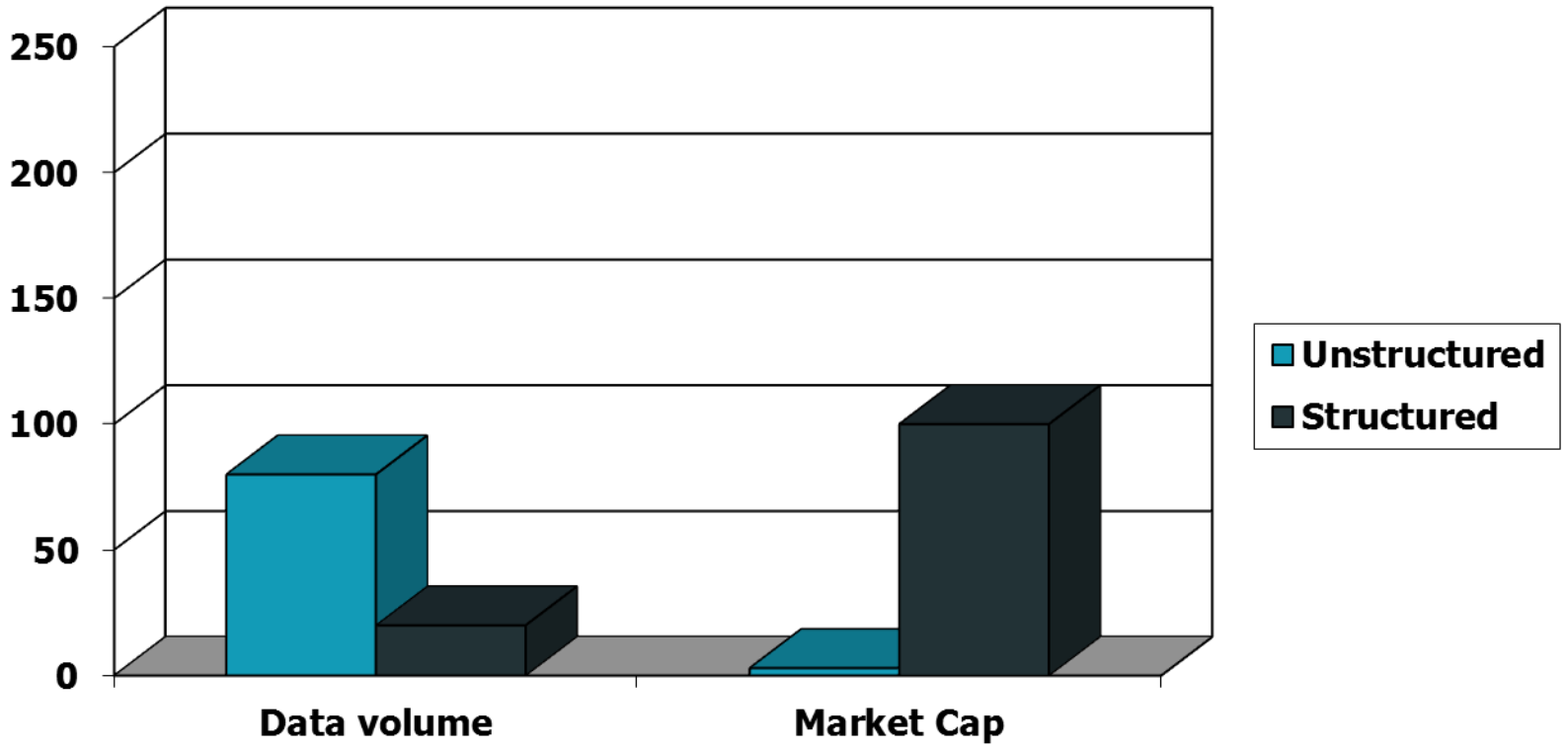
- In contrast to the rigidly structured data used in DB style searching (e.g. product inventories, personnel records)


```
SELECT *
FROM business_catalogue
WHERE category = 'florist'
AND city_zip = 'cb1'
```

- This does not mean that there is no structure in the data
  - Document structure (headings, paragraphs, lists. . .)
  - Explicit markup formatting (e.g. in HTML, XML. . .)
  - Linguistic structure (latent, hidden)

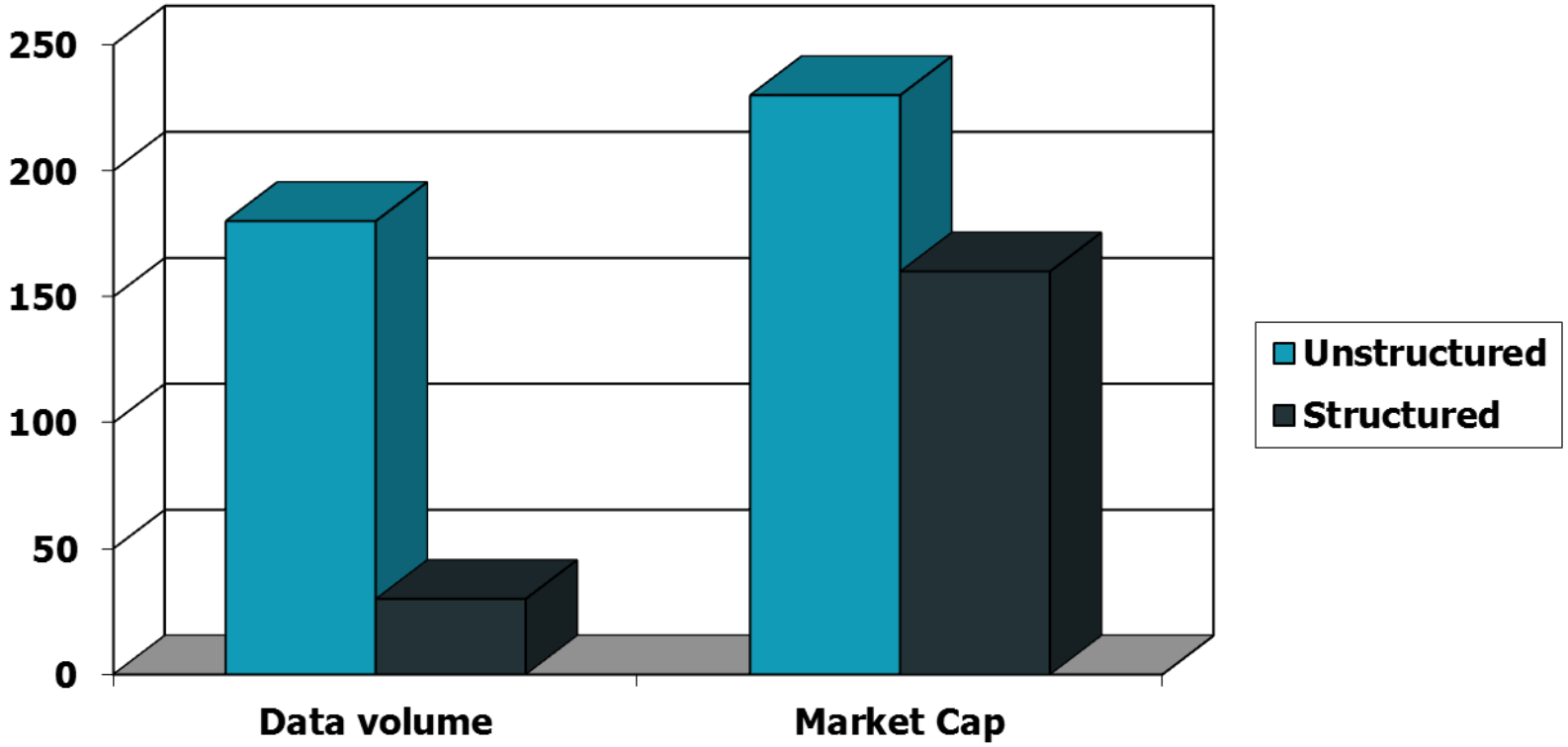
Taken from: Simone Teufel and Ronan Cummins

Unstructured (text) vs. structured (database) data in the mid-nineties



Taken from: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze

Unstructured (text) vs. structured (database) data today



Taken from: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze

Information Needs and Relevance

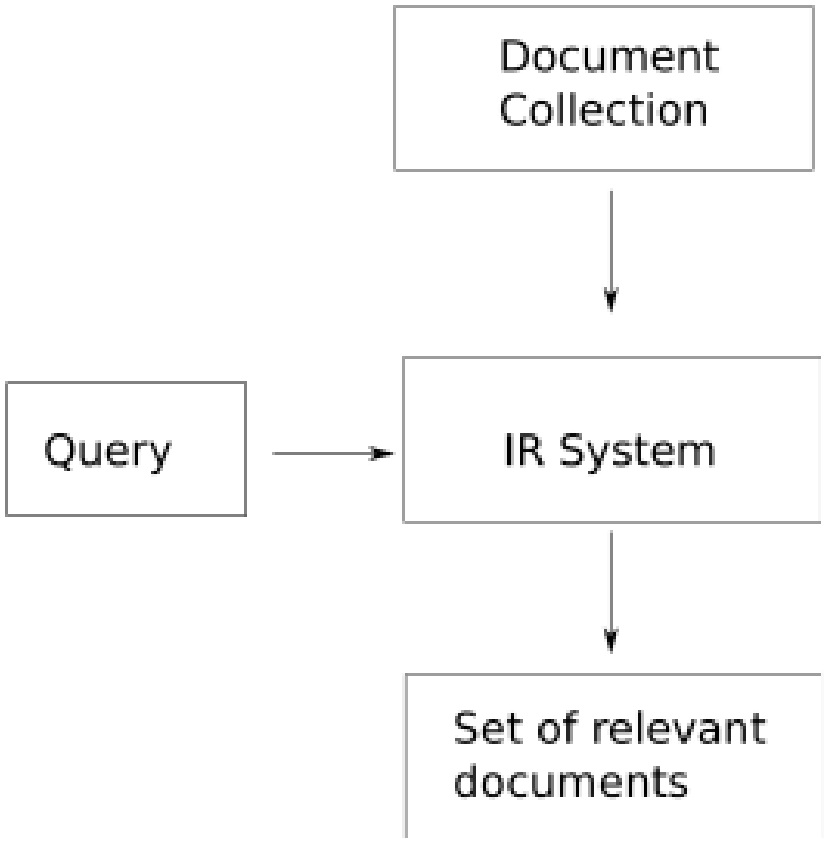
Manning et al, 2008:

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- An information need is the topic about which the user desires to know more about.
- A query is what the user conveys to the computer in an attempt to communicate the information need.
- A document is relevant if the user perceives that it contains information of value with respect to their personal information need.

Taken from: Simone Teufel and Ronan Cummins

IR Basics



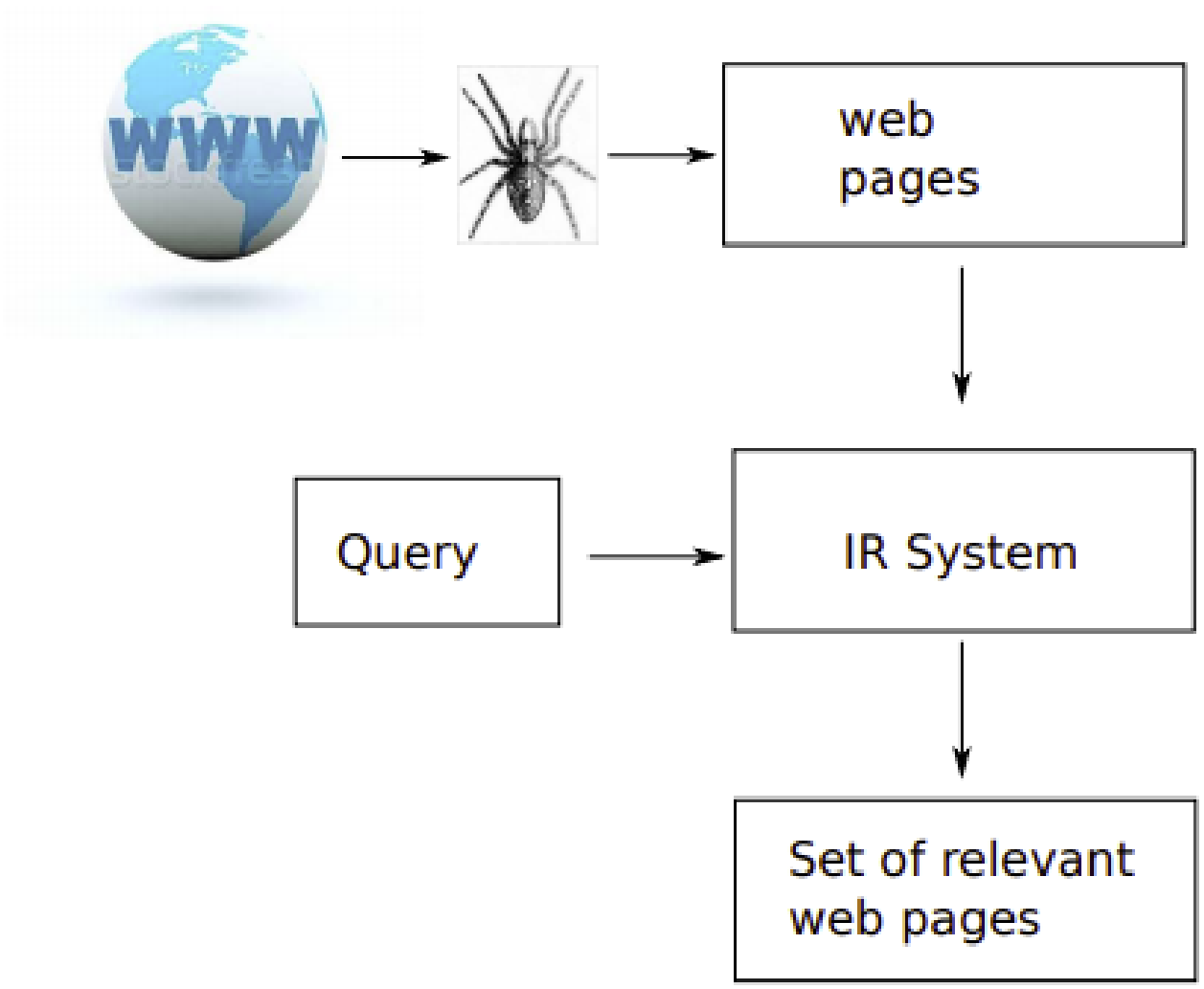
Taken from: Simone Teufel and Ronan Cummins

WHAT IS INFORMATION RETRIEVAL?

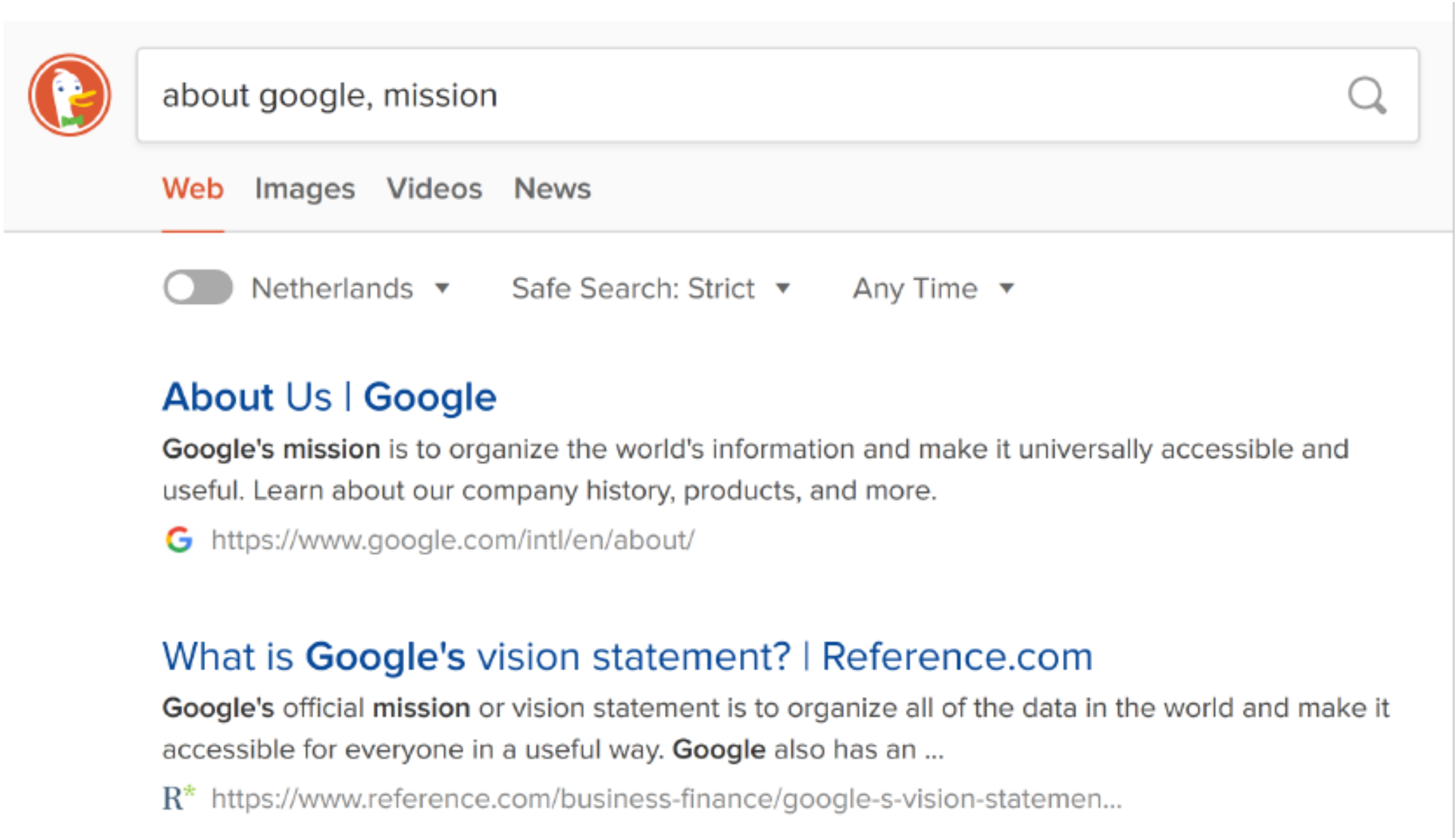
= web search !!!

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

IR Basics



Taken from: Simone Teufel and Ronan Cummins



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

What other organisations have this mission?

- Libraries ?
- Scopus, Web of Science, … ?
- Twitter / Facebook ?
- Netflix ?
- Amazon ?
- iTunes / Spotify ?
- Medium ?

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Types of information needs

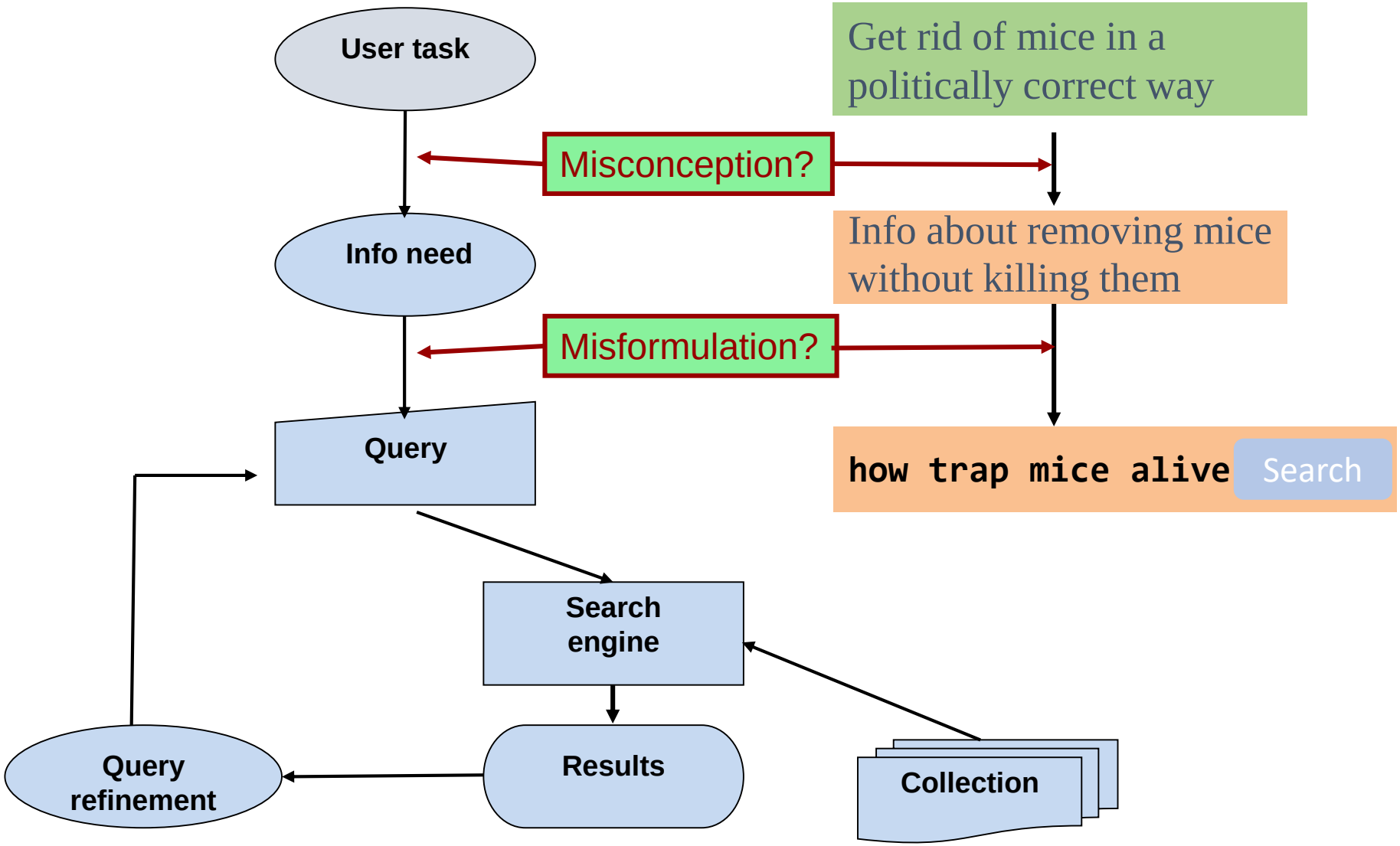
Manning et al, 2008:

> Information retrieval (IR) is finding material . . . of an unstructured nature . . . that satisfies an information need from within large collections . . . .

- Known-item search
- Precise information seeking search
- Open-ended search ("topical search")

Taken from: Simone Teufel and Ronan Cummins

The classic search model



Taken from: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze

Information scarcity vs. information abundance

- Information scarcity problem (or needle-in-haystack problem): hard to find rare information
  - Lord Byron's first words? 3 years old? Long sentence to the nurse in perfect English?

> . . . when a servant had spilled an urn of hot coffee over his legs, he replied to the distressed inquiries of the lady of the house, 'Thank you, madam, the agony is somewhat abated.' [not Lord Byron, but Lord Macaulay]

- Information abundance problem (for more clear-cut information needs): redundancy of obvious information
  - What is toxoplasmosis?

Taken from: Simone Teufel and Ronan Cummins

Relevance

Manning et al, 2008:

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- Are the retrieved documents
  - about the target subject
  - up-to-date?
  - from a trusted source?
  - satisfying the user's needs?
- How should we rank documents in terms of these factors?
- More on this in a lecture soon

Taken from: Simone Teufel and Ronan Cummins

How well has the system performed?

The effectiveness of an IR system (i.e., the quality of its search results) is determined by two key statistics about the system's returned results for a query:

- Precision: What fraction of the returned results are relevant to the information need?
- Recall: What fraction of the relevant documents in the collection were returned by the system?
- What is the best balance between the two?
  - Easy to get perfect recall: just retrieve everything
  - Easy to get good precision: retrieve only the most relevant

There is much more to say about this – lecture 6

Taken from: Simone Teufel and Ronan Cummins

IR today

- Web search (Google bing)
  - Search ground are billions of documents on millions of computers
  - issues: spidering; efficient indexing and search; malicious manipulation to boost search engine rankings
  - Link analysis covered in Lecture 8
- Enterprise and institutional search (PubMed LexisNexis)
  - e.g company's documentation, patents, research articles
  - often domain-specific
  - Centralised storage; dedicated machines for search.
  - Most prevalent IR evaluation scenario: US intelligence analyst's searches
- Personal information retrieval (email, pers. documents;)
  - e.g., Mac OS X Spotlight; Windows' Instant Search
  - Issues: different file types; maintenance-free, lightweight to run in background

Taken from: Simone Teufel and Ronan Cummins

History of IR

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

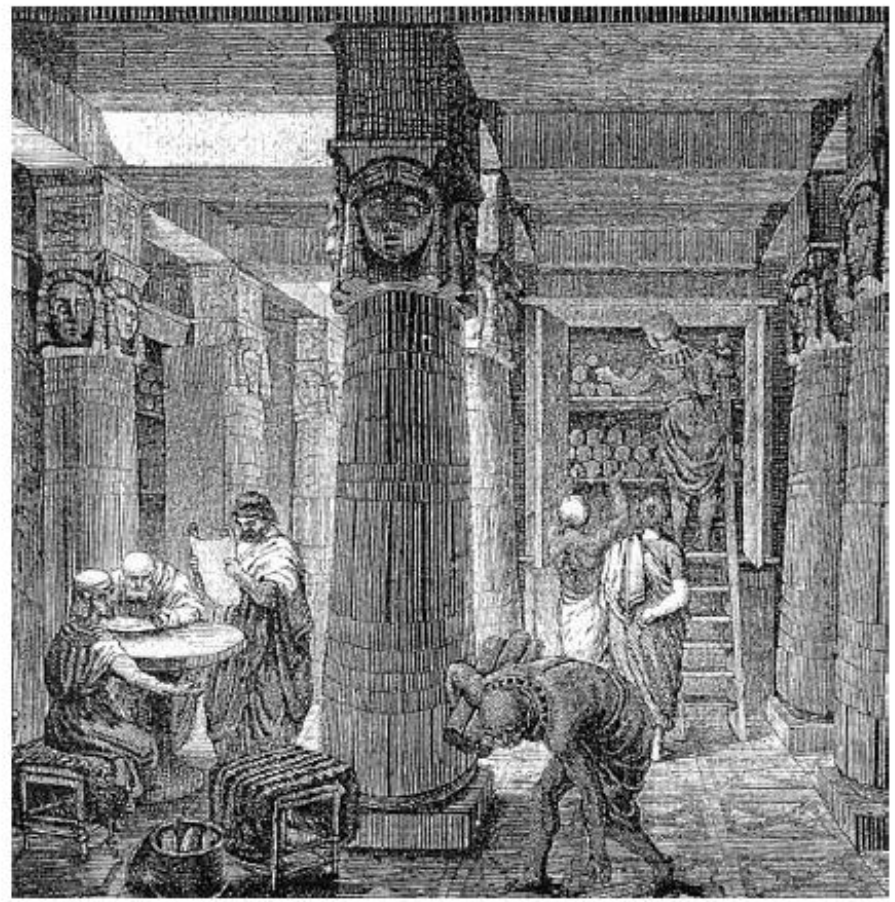


Image from Wikipedia

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

THE LIBRARY OF ALEXANDRIA

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

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

THE LIBRARY OF ALEXANDRIA

- **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.

Image: allpostersimages.com

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

PRE-HISTORY: STANDARDS

- Melvil Dewey's Decimal Classification (1876)

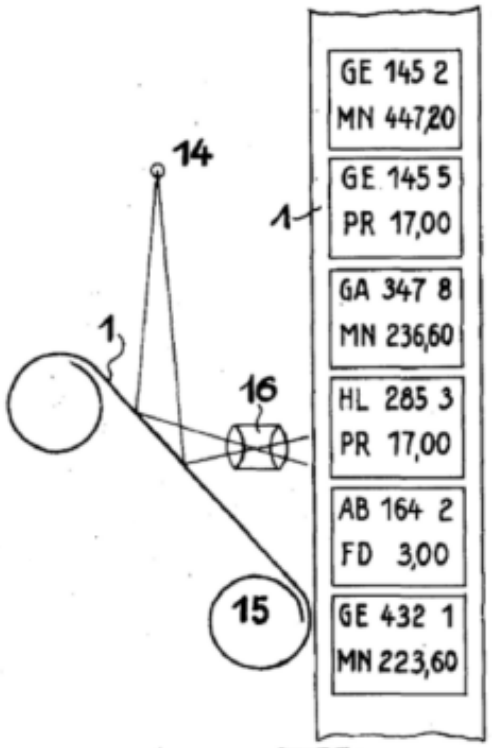
  Hierarchical numbering scheme made up of ten classes, each divided into ten divisions, each having ten sections.
  Decimals create further divisions:

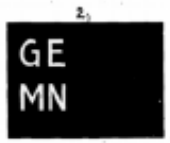
  500 Natural sciences and mathematics
  510 Mathematics
  516 Geometry
  516.3 Analytic geometries
  516.37 Metric differential geometries
  516.375 Finsler Geometry

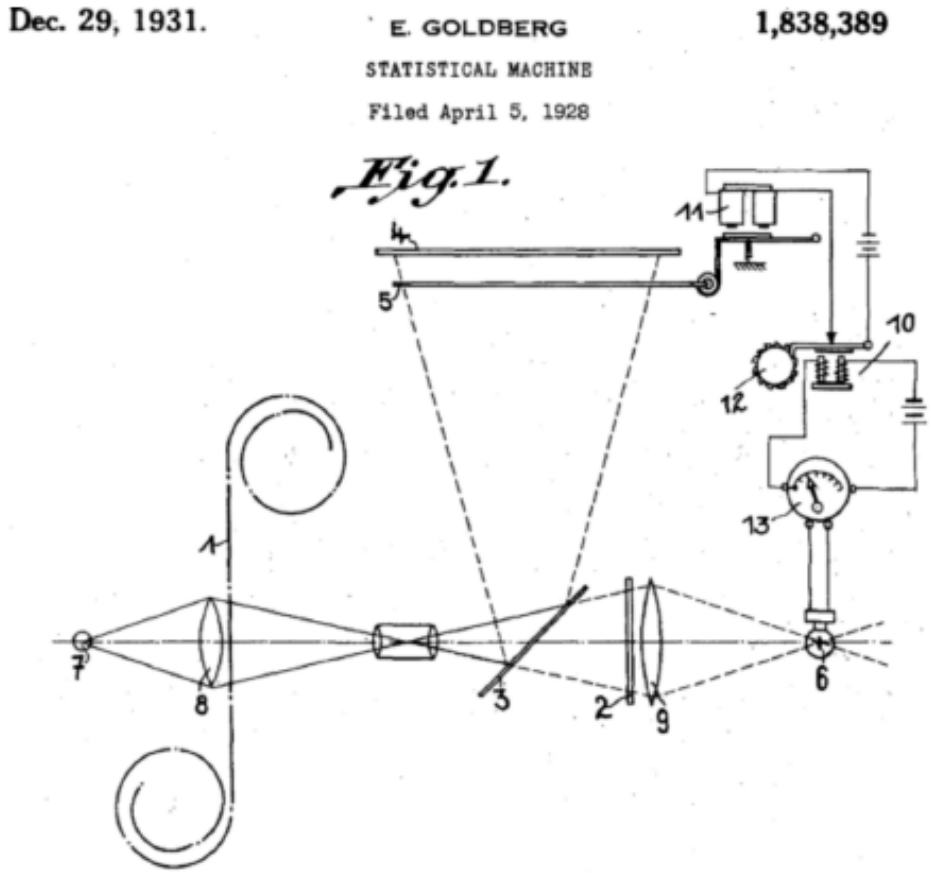
Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

PRE-HISTORY: FIRST MACHINES

- Emanuel Goldberg's Microfilm Search "Statistical Machine" (patent 1931)







Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

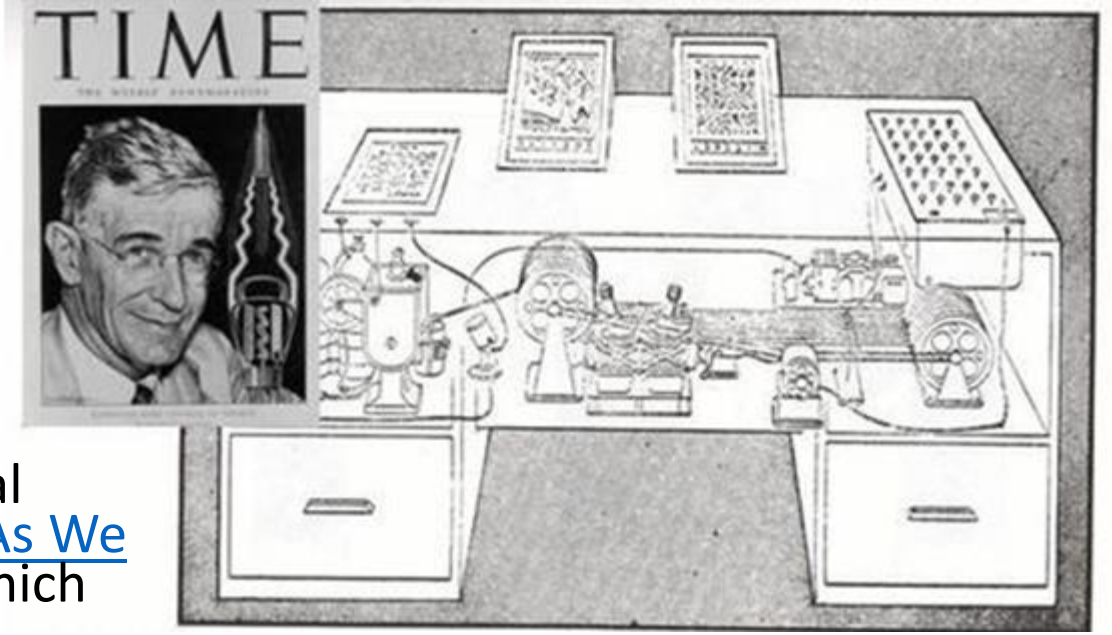
PRE-HISTORY: FIRST MACHINES

- Emanuel Goldberg's Microfilm Search
  "Statistical Machine" (patent 1931)

  "Here it can be seen that catalogue entries were stored on a roll of film (No. 1 of the figure). A query (2) was also on film showing a negative image of the part of the catalogue being searched for; in this case the 1 st and 6 th entries on the roll. A light source (7) was shone through the catalogue roll and query film, focused onto a photocell (6). If an exact match was found, all light was blocked to the cell causing a relay to move a counter forward (12) and for an image of the match to be shown via a half silvered mirror (3), reflecting the match onto a screen or photographic plate (4 & 5)."

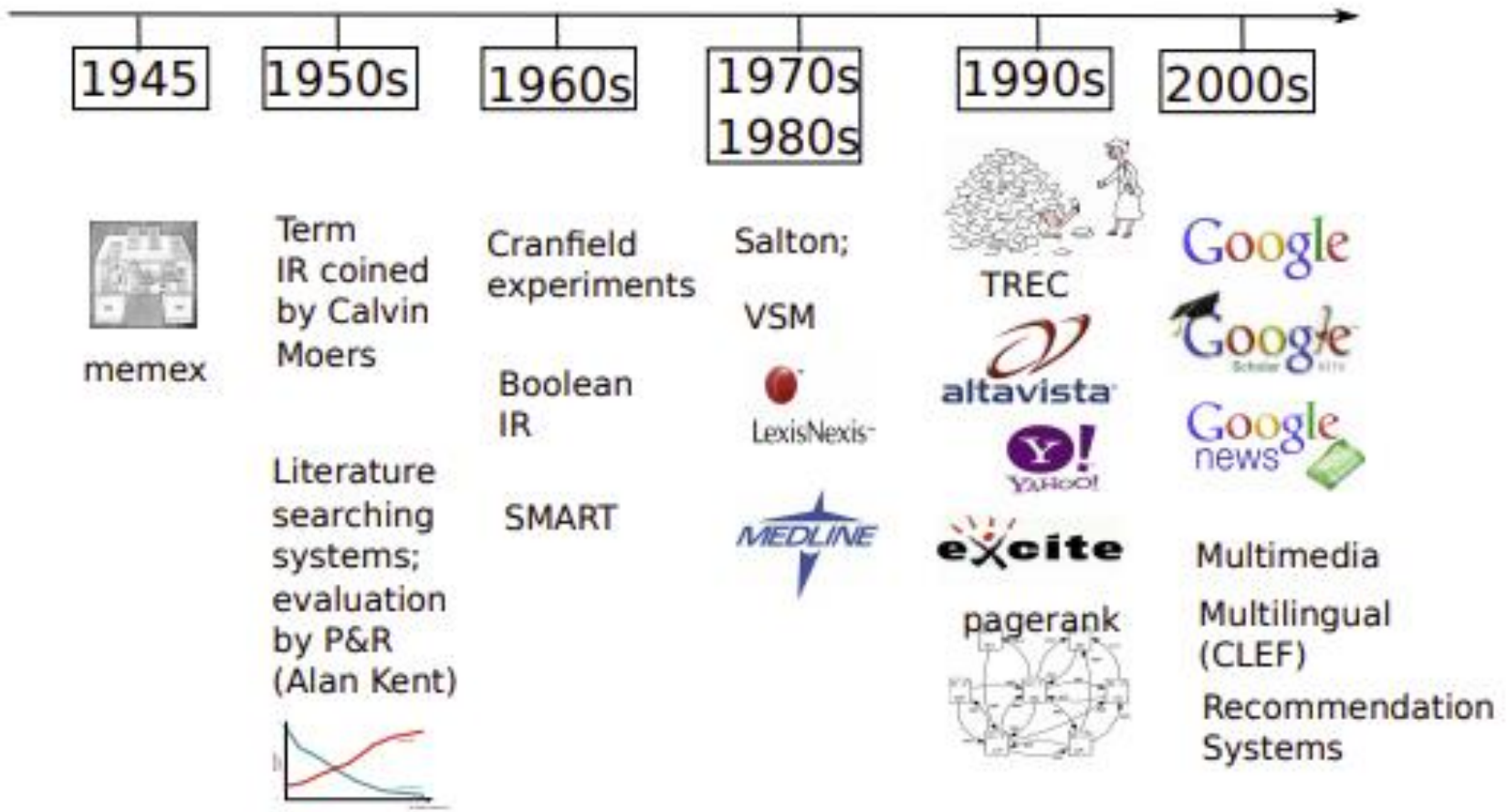
Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Memex (Wikipedia)



- **Memex** is the name of the hypothetical electromechanical device that Vannevar Bush described in his 1945 article "As We May Think". Bush envisioned the memex as a device in which individuals would compress and store all of their books, records, and communications, "mechanized so that it may be consulted with exceeding speed and flexibility". The individual was supposed to use the memex as automatic personal filing system, making the memex "an enlarged intimate supplement to his memory".[1]
- The concept of the memex influenced the development of early hypertext systems, eventually leading to the creation of the World Wide Web, and personal knowledge base software.[2] The hypothetical implementation depicted by Bush for the purpose of concrete illustration was based upon a document bookmark list of static microfilm pages and lacked a true hypertext system, where parts of pages would have internal structure beyond the common textual format.

A short history of IR



Taken from: Simone Teufel and Ronan Cummins

History of IR

- 1960-70’s:
  - Initial exploration of text retrieval systems for “small” corpora of scientific abstracts, and law and business documents.
  - Development of the basic Boolean and vector-space models of retrieval.
  - Prof. Salton and his students at Cornell University are the leading researchers in the area.

Taken from: Raymond J. Mooney

IR History Continued

- 1980’s:
  - Large document database systems, many run by companies:
    - Lexis-Nexis
    - Dialog
    - MEDLINE

Taken from: Raymond J. Mooney

IR History Continued

- 1990’s:
  - Searching FTPable documents on the Internet
    - Archie
    - WAIS
  - Searching the World Wide Web
    - Lycos
    - Yahoo
    - Altavista

Taken from: Raymond J. Mooney

IR History Continued

- 1990’s continued:
  - Organized Competitions
    - NIST TREC
  - Recommender Systems
    - Ringo
    - Amazon
    - NetPerceptions
  - Automated Text Categorization & Clustering

Taken from: Raymond J. Mooney

IR History Continued

- 2000’s
  - Link analysis for Web Search
    - Google
  - Automated Information Extraction
  - Parallel Processing
    - Map/Reduce
  - Question Answering
    - TREC Q/A track

Taken from: Raymond J. Mooney

IR History Continued

- 2000's continued:
  - Multimedia IR
    - Image
    - Video
    - Audio and music
  - Cross-Language IR
    - DARPA Tides
  - Document Summarization
  - Learning to Rank

Taken from: Raymond J. Mooney

Recent IR History

- 2010’s
  - Intelligent Personal Assistants
    - Siri
    - Cortana
    - Google Now
    - Alexa
  - Complex Question Answering
    - IBM Watson
  - Distributional Semantics
  - Deep Learning

Taken from: Raymond J. Mooney

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.”

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

HISTORY: STANDARDS

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

1910 - 1965

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

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 (!)

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

HISTORY: RANKING

- Many researchers argued that *ranking* is essential

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

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

HISTORY: TEXT RETRIEVAL CONFERENCE (TREC)

- Development of standard reusable test collections based on Cleverdon's work (1992)
- Organized by Donna Harman and later Ellen Voorhees

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

HISTORY: EFFICIENCY & COMPRESSION

- Ian Witten, Alistair Moffat, and Timothy Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, 1994



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

HISTORY: RANKING & MODELS

- Modern ranking models

Stephen Robertson (1994)
BM25 (with Steve Walker)

Bruce Croft (1998)
Language Models (with Jay Ponte)
(independently discovered by Djoerd Hiemstra and Miller, Leek & Schwartz)

Larry Page (1998)
Google PageRank (with Sergey Brin)

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

HISTORY: RANKING & MODELS

- Recent developments

  Machine Learning for IR:
  - "learning to rank"
  - "(deep) neural IR"

  Question answering
  - "conversational search"

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

FURTHER READING

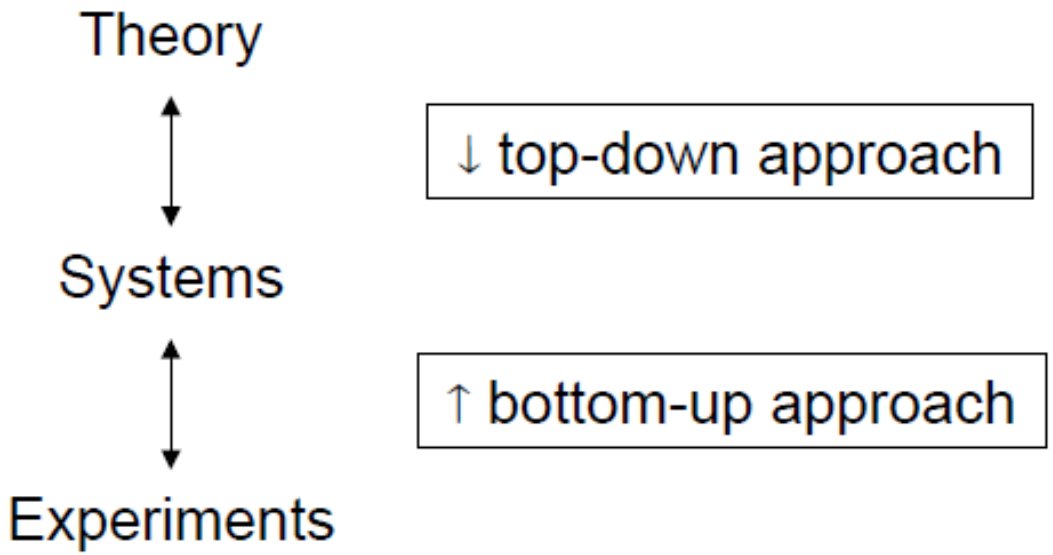
Mark Sanderson and Bruce Croft,
The History of Information Retrieval Research,
Proceedings of the IEEE, Volume 100, 2012
http://marksanderson.org/publications/my_papers/IEEE2012.pdf

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

IR System

IR RESEARCH

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



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Information Retrieval (IR)

- The indexing and retrieval of textual documents.
- Searching for pages on the World Wide Web
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.

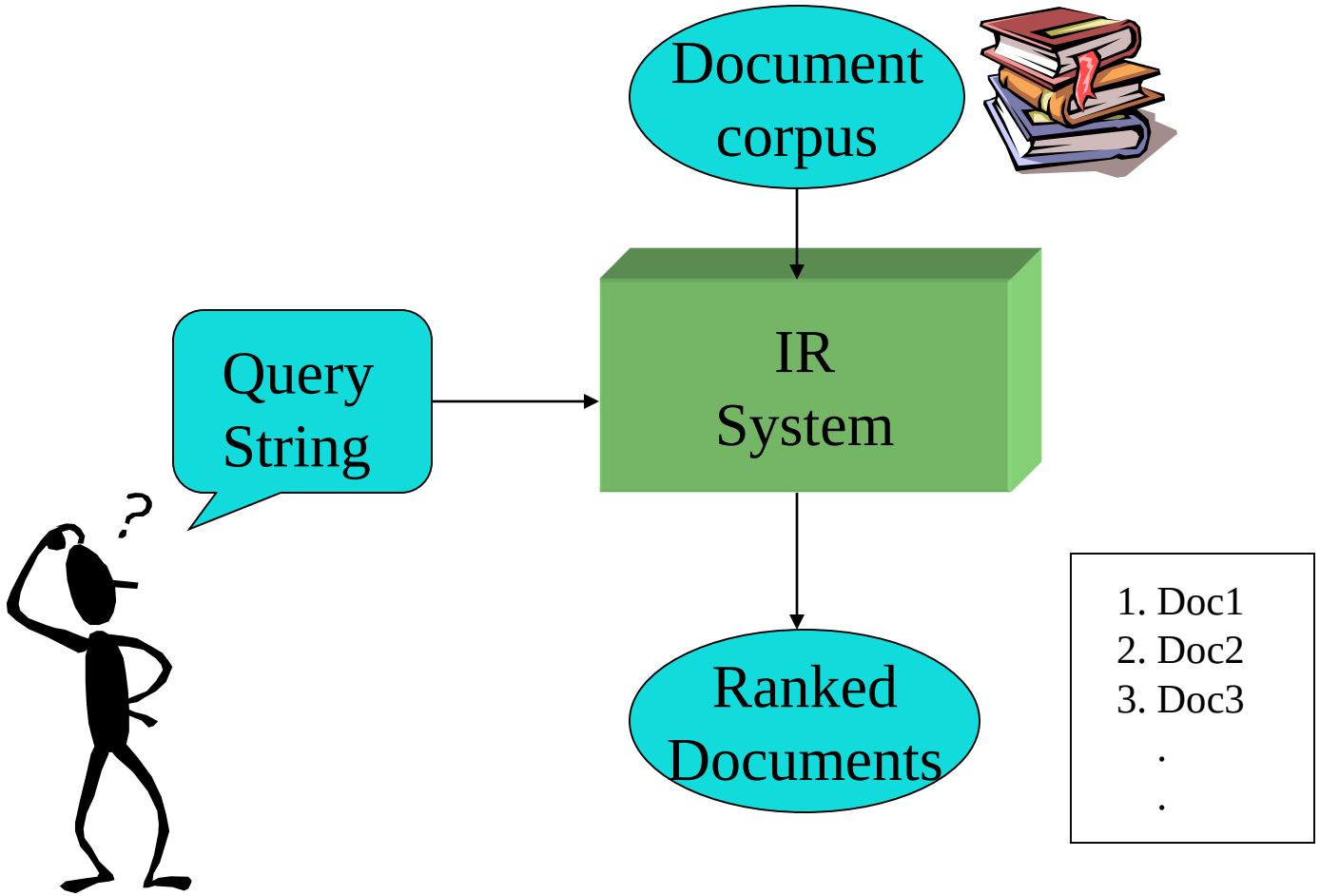
Taken from: Raymond J. Mooney

Typical IR Task

- Given:
  - A corpus of textual natural-language documents.
  - A user query in the form of a textual string.
- Find:
  - A ranked set of documents that are relevant to the query.

Taken from: Raymond J. Mooney

IR System



Taken from: Raymond J. Mooney

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

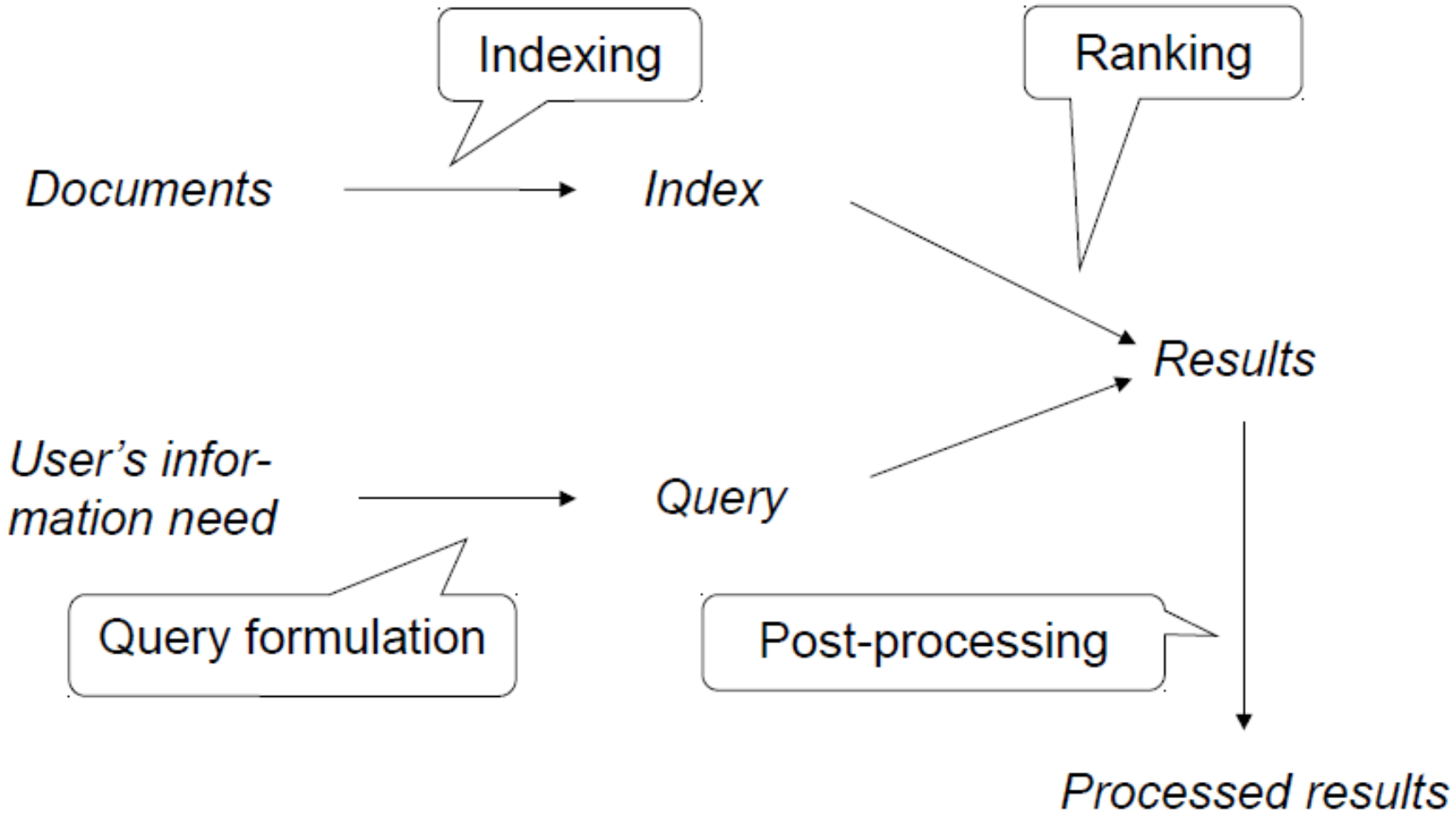
Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Relevance

- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (*information need*).

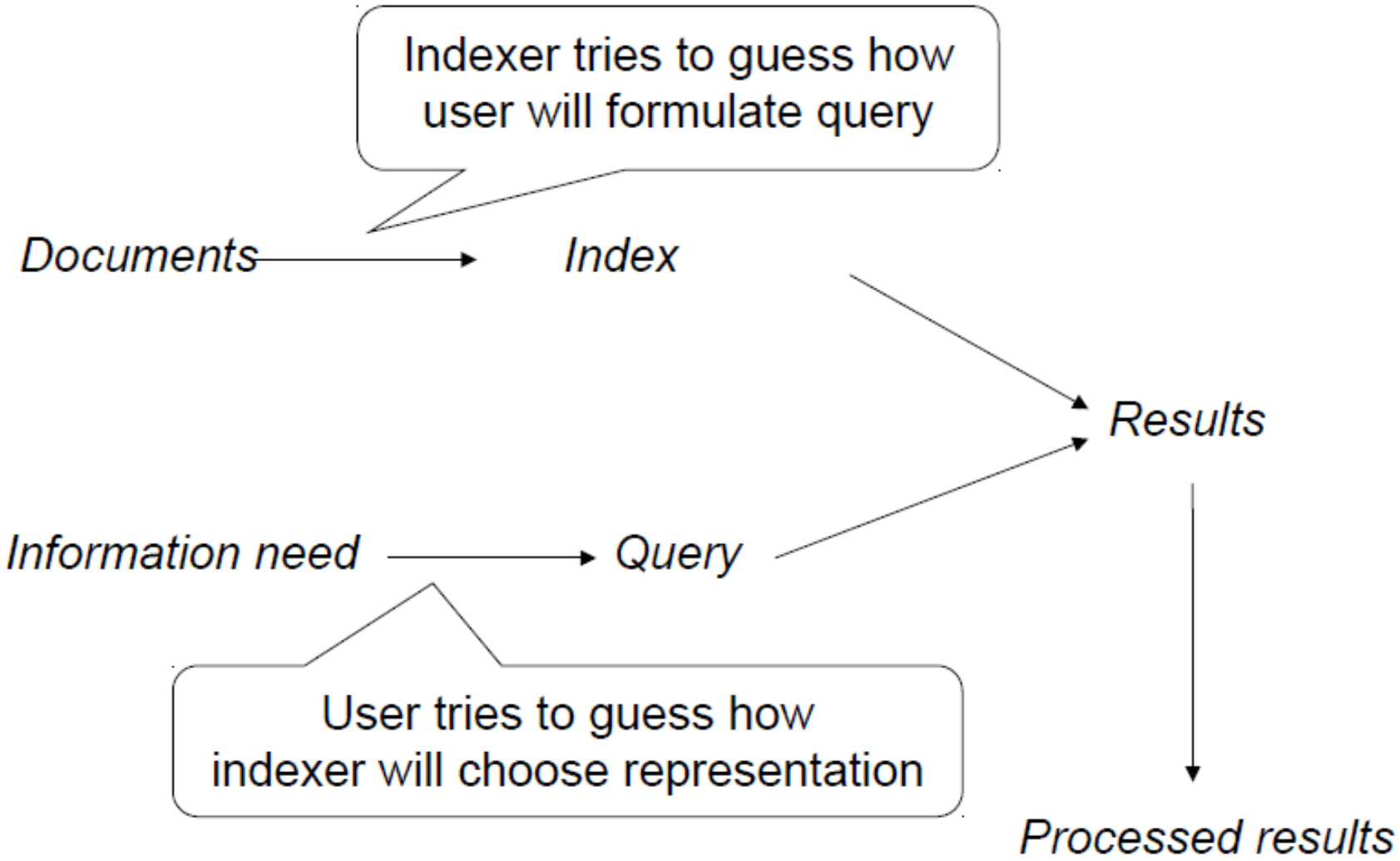
Taken from: Raymond J. Mooney

GRAPHICAL REPRESENTATION OF IR



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

THE PREDICTION GAME



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

ANOTHER VIEW

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

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

VARIANTS

- Pull: ad-hoc requests, like WWW-searches
  - collection static, query dynamic
- Push: filtering, like personalised news service or spam filter
  - collection dynamic, query static

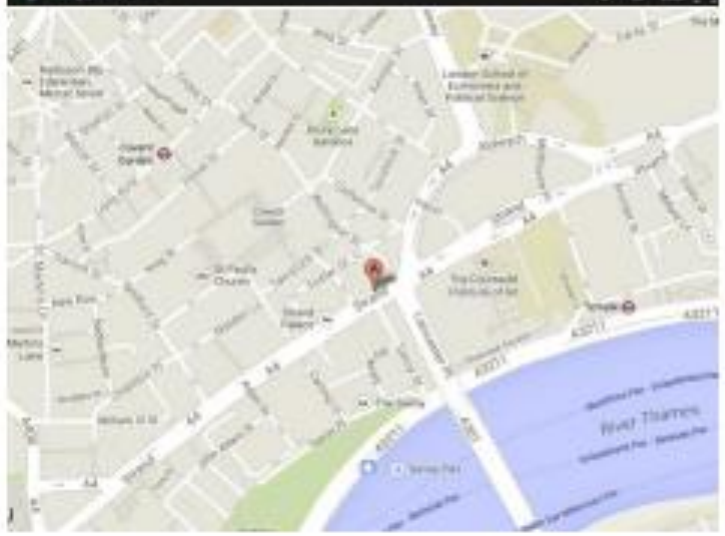
Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

MORE THAN TEXT

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

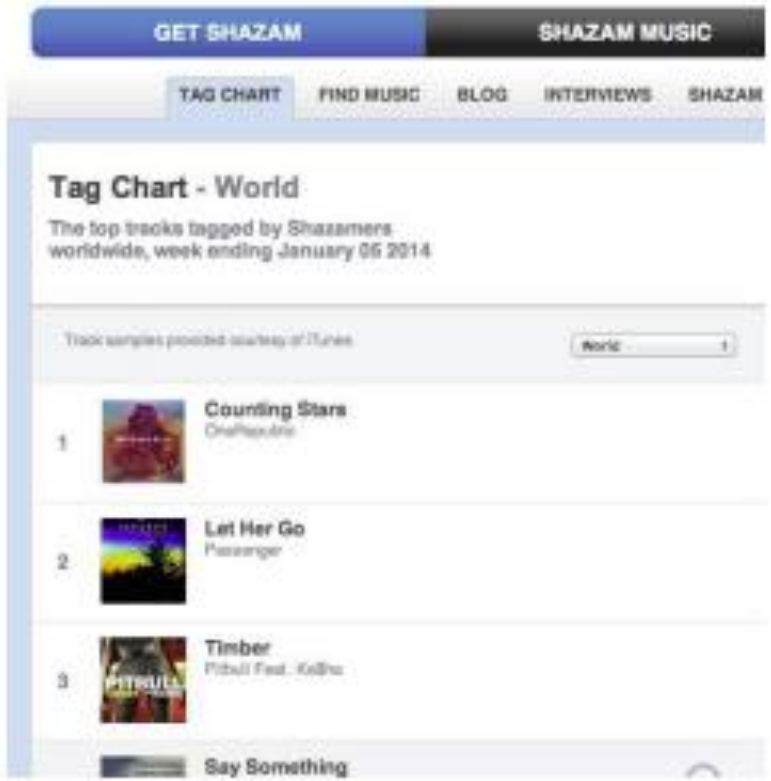
Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

IR for non-textual media



Taken from: Simone Teufel and Ronan Cummins

Similarity Searches



Taken from: Simone Teufel and Ronan Cummins

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that the words in the query appear frequently in the document, in any order (*bag of words*).

Taken from: Raymond J. Mooney

Problems with Keywords

- May not retrieve relevant documents that include synonymous terms.
  - “restaurant” vs. “café”
  - “PRC” vs. “China”
- May retrieve irrelevant documents that include ambiguous terms.
  - “bat” (baseball vs. mammal)
  - “Apple” (company vs. fruit)
  - “bit” (unit of data vs. act of eating)

Taken from: Raymond J. Mooney

Beyond Keywords

- We will cover the basics of keyword-based IR, but…
- We will focus on extensions and recent developments that go beyond keywords.
- We will cover the basics of building an *efficient* IR system, but…
- We will focus on basic capabilities and algorithms rather than systems issues that allow scaling to industrial size databases.

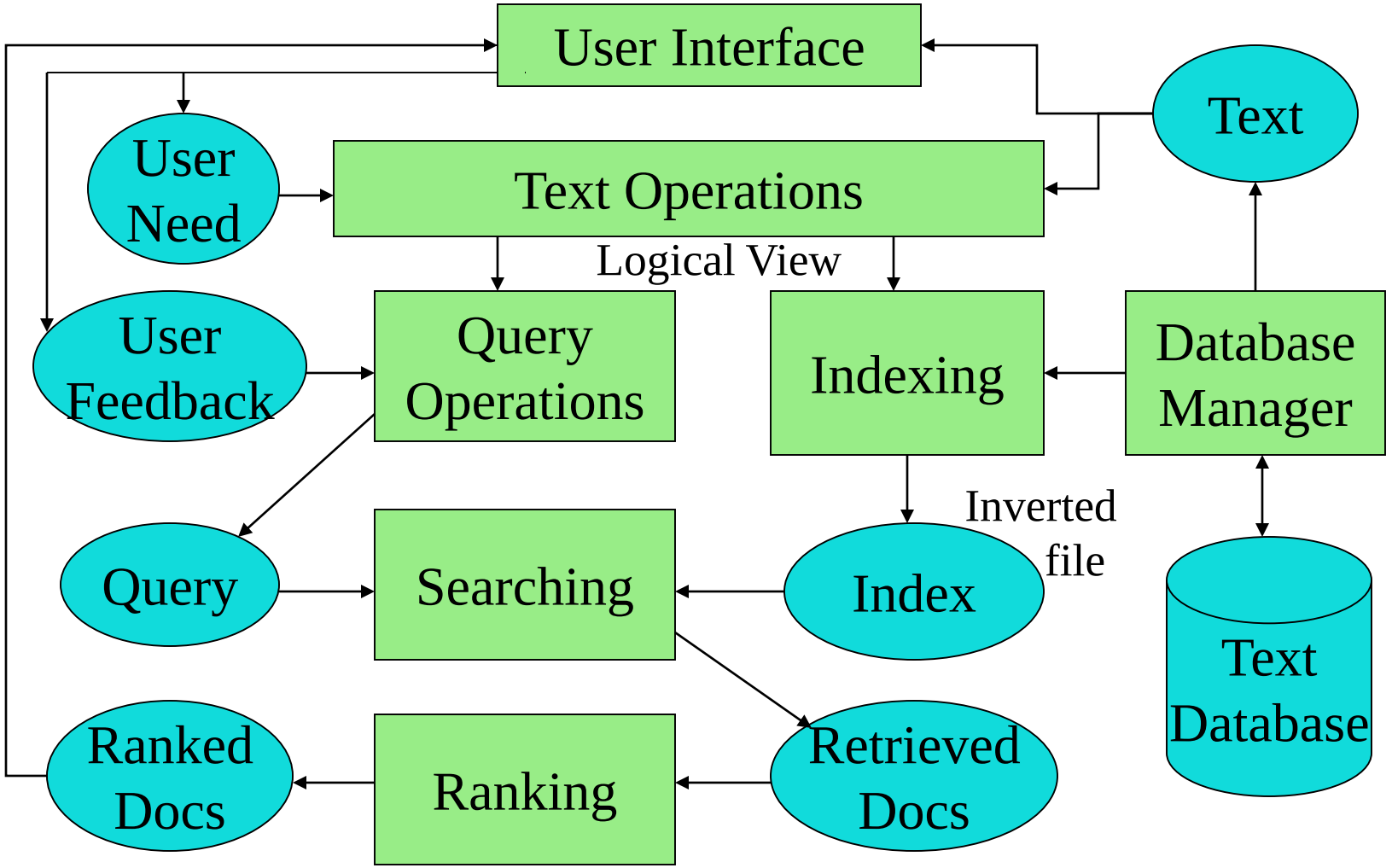
Taken from: Raymond J. Mooney

Intelligent IR

- Taking into account the *meaning* of the words used.
- Taking into account the *order* of words in the query.
- Adapting to the user based on direct or indirect feedback.
- Taking into account the *authority* of the source.

Taken from: Raymond J. Mooney

IR System Architecture



Taken from: Raymond J. Mooney

IR System Components

- Text Operations forms index words (tokens).
  - Stopword removal
  - Stemming
- Indexing constructs an *inverted index* of word to document pointers.
- Searching retrieves documents that contain a given query token from the inverted index.
- Ranking scores all retrieved documents according to a relevance metric.

Taken from: Raymond J. Mooney

IR System Components (continued)

- User Interface manages interaction with the user:
  - Query input and document output.
  - Relevance feedback.
  - Visualization of results.
- Query Operations transform the query to improve retrieval:
  - Query expansion using a thesaurus.
    - Query transformation using relevance feedback.

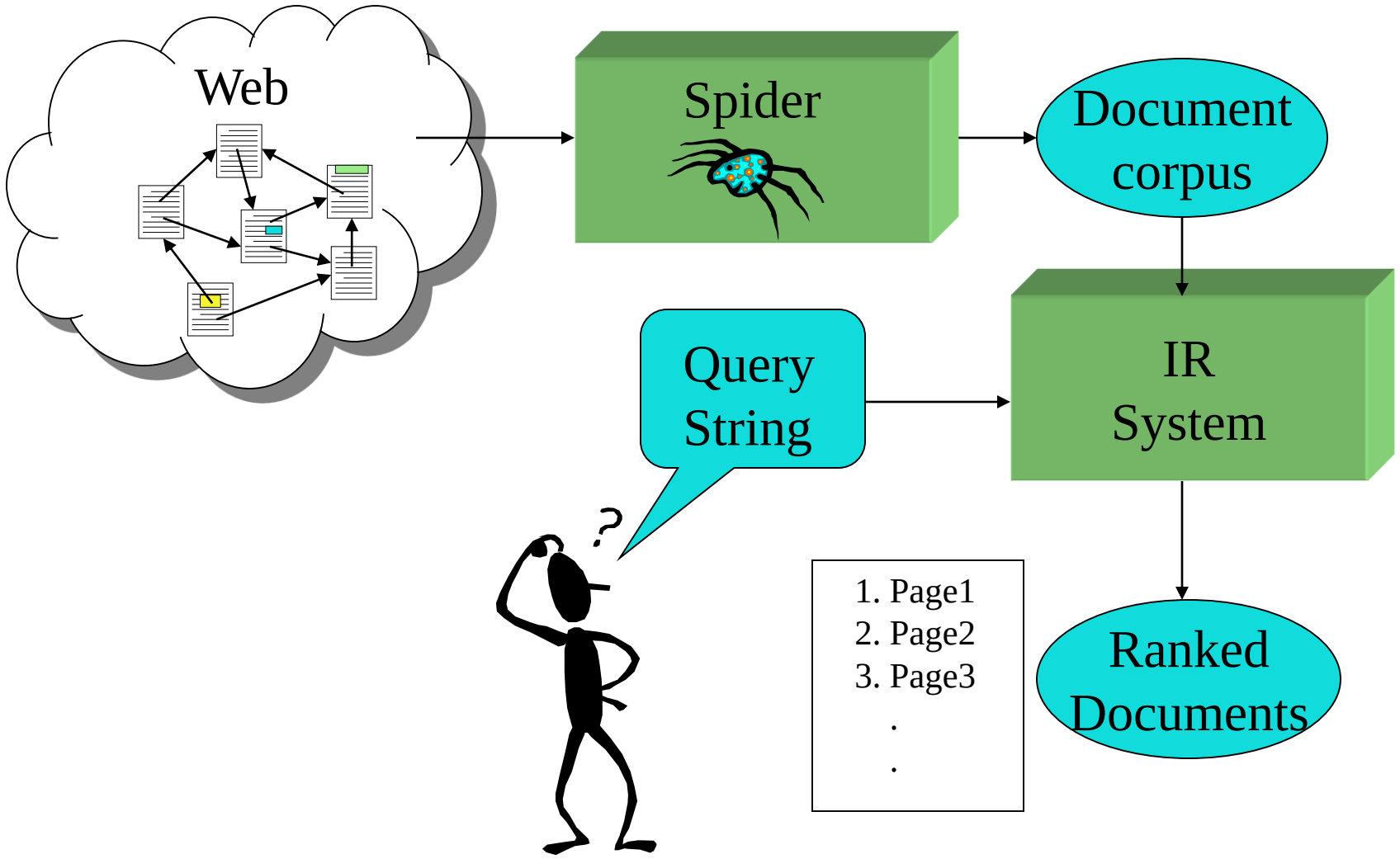
Taken from: Raymond J. Mooney

Web Search

- Application of IR to HTML documents on the World Wide Web.
- Differences:
  - Must assemble document corpus by spidering the web.
  - Can exploit the structural layout information in HTML (XML).
  - Documents change uncontrollably.
  - Can exploit the link structure of the web.

Taken from: Raymond J. Mooney

Web Search System



Taken from: Raymond J. Mooney

Other IR-Related Tasks

- Automated document categorization
- Information filtering (spam filtering)
- Information routing
- Automated document clustering
- Recommending information or products
- Information extraction
- Information integration
- Question answering

Taken from: Raymond J. Mooney

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?

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

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
  - personalisation
  - recommender systems

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

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

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Approaches: ordering engine

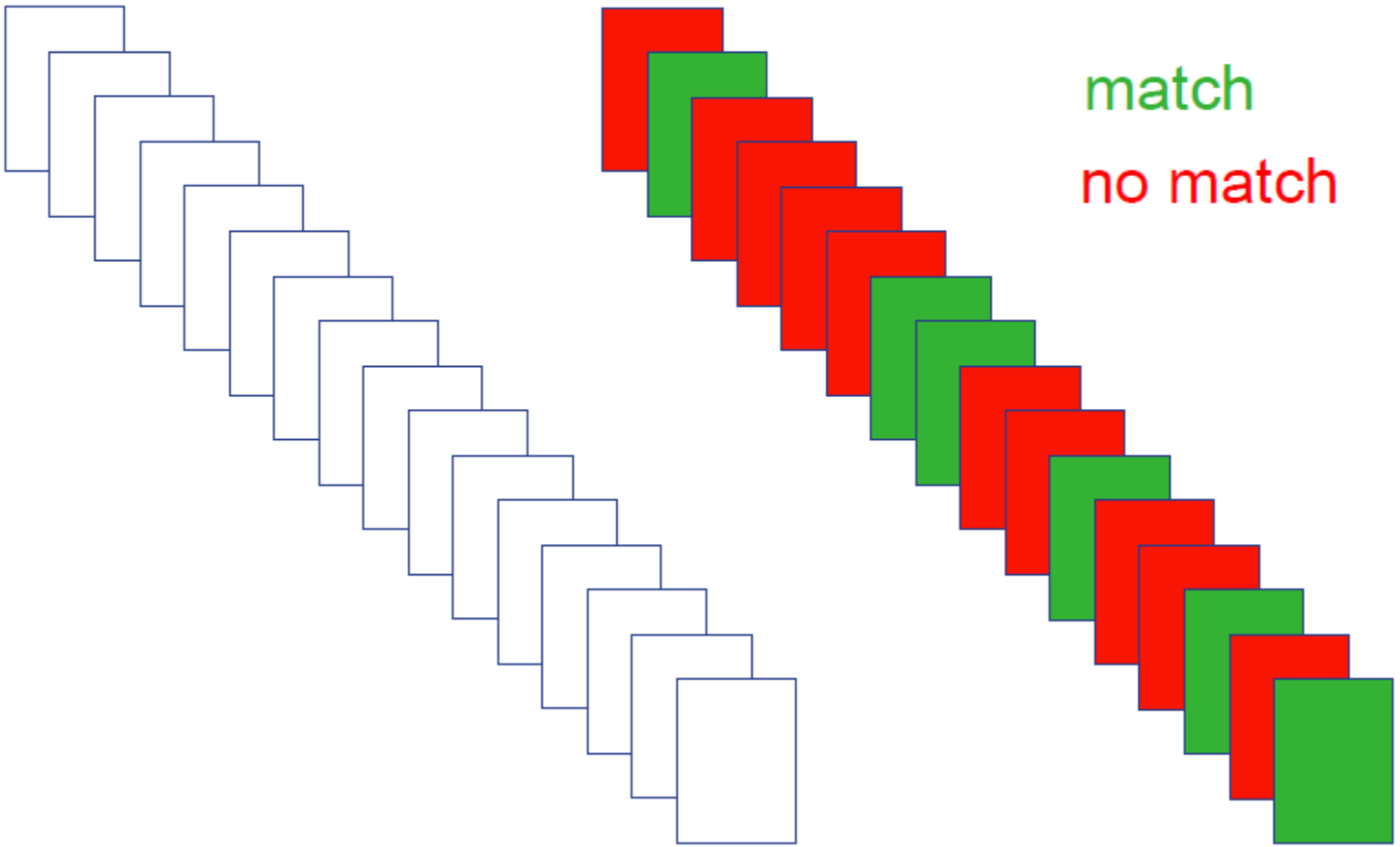
Two basic approaches:

- *Matching* imposes a dichotomy on the collection
- *Ranking* rank-orders the entire collection

- N.B. The set $\{A, B\}$ is a dichotomy of set C iff $A \cap B = \varnothing$ and $A \cup B = C$

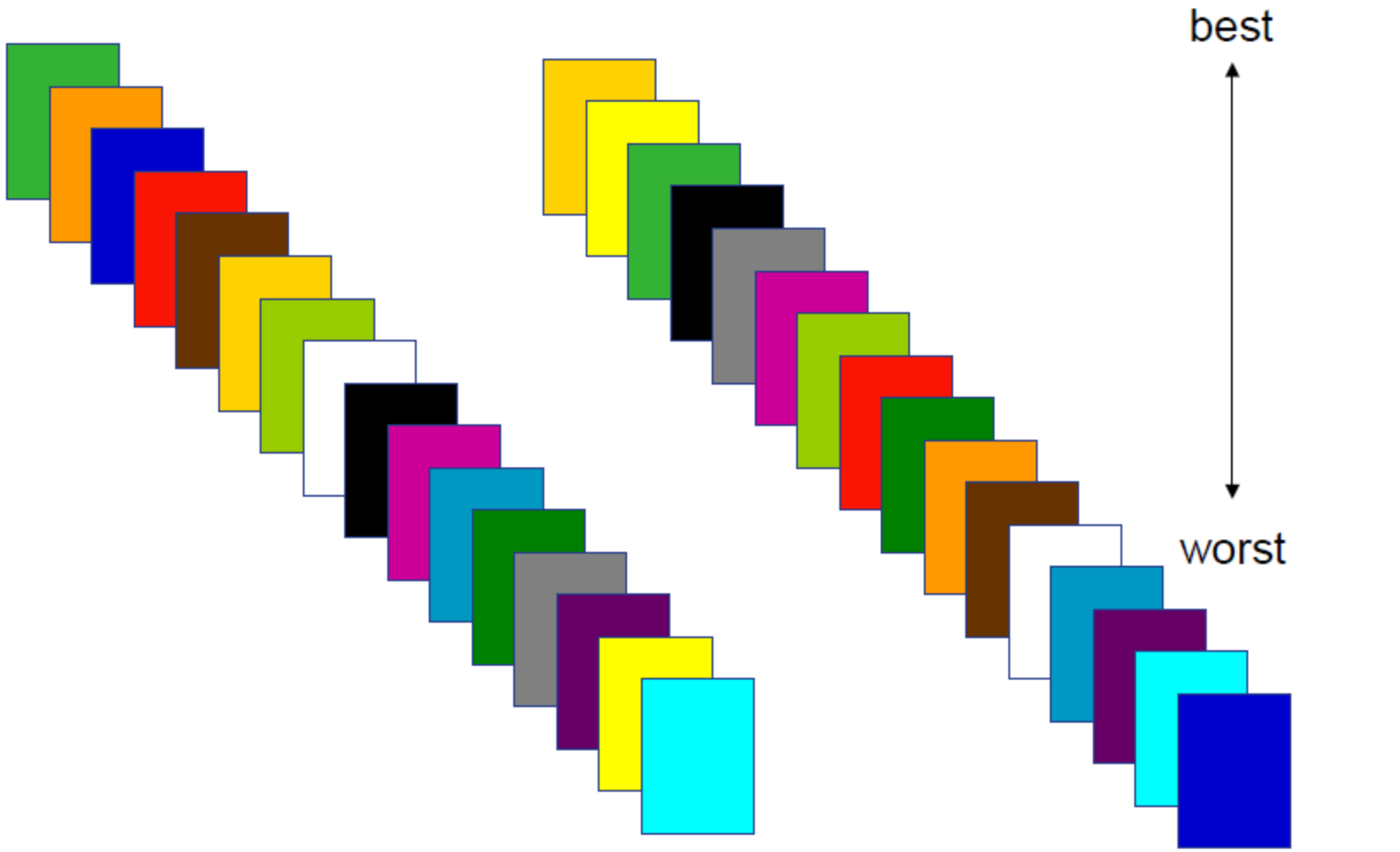
Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Matching



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Ranking



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

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)

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Performance

- Important decision: which system is better?
- Has large economic impact
- Compare Google's market value
- A good IR system can make the difference between winning or losing e.g.
  - a contract
  - a legal case

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Measuring performance

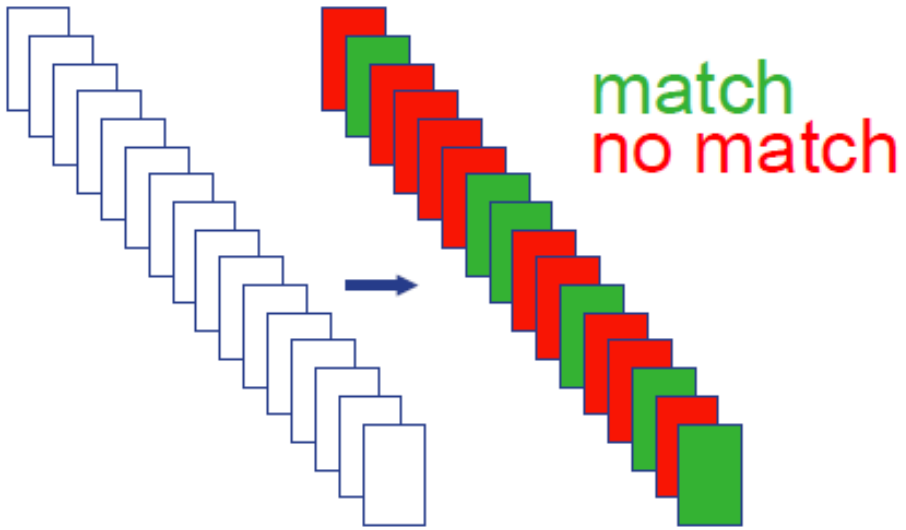
Theory of measurement in IR is difficult, for example:

- 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?

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

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)



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Performance for matching

System says:

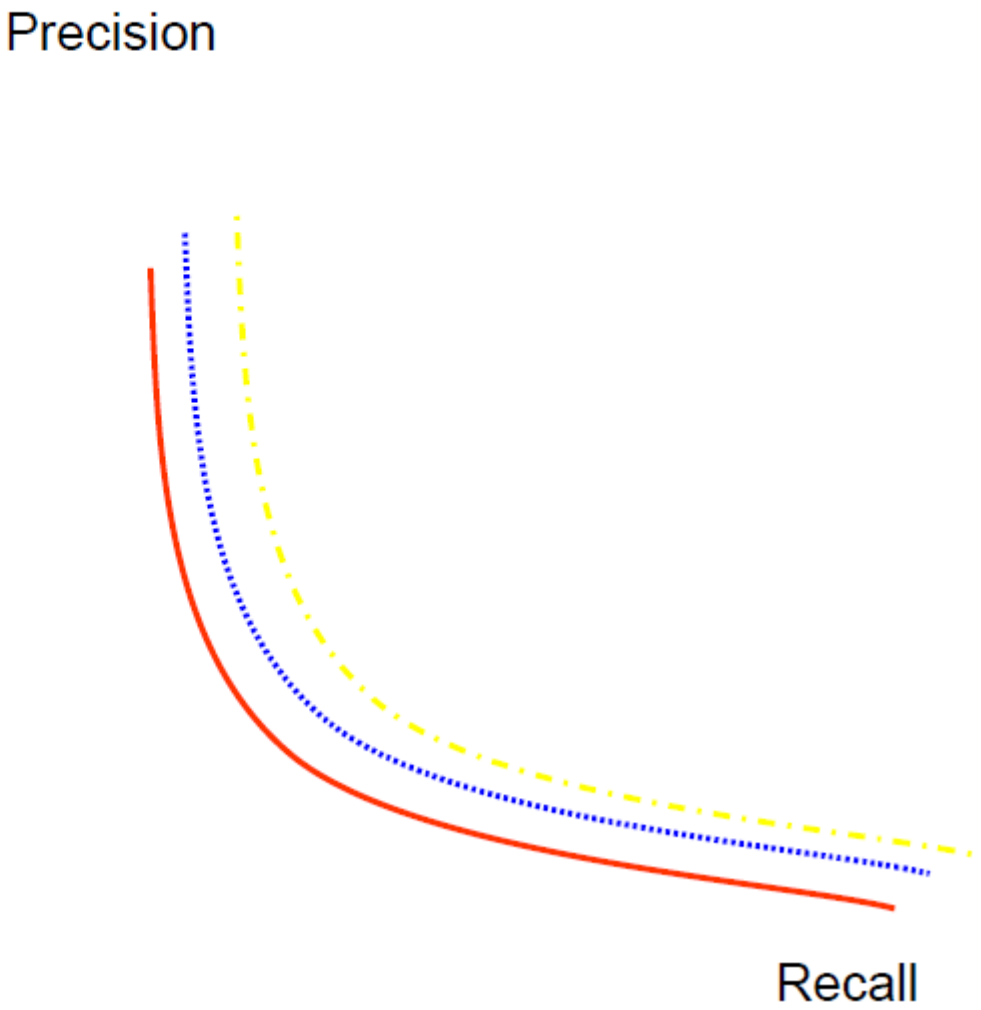
User says:	Match	No match
Relevant	True positives *(#TP)*	False negatives *(#FN)*
Not relevant	False positives *(#FP)*	True negatives *(#TN)*

$$Recall = \frac{\#TP}{\#TP + \#FN} \qquad Precision = \frac{\#TP}{\#TP + \#FP}$$

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Performance for matching

- "Fact of life": improving recall typically decreases precision.



Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

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”
- Pro:
  - State of the art known
  - Competition incentive for improvement
  - Forum for exchange of ideas
- Con:
  - Test environment sets constraints on what can be done and what not

Taken from: Theo Huibers, Djoerd Hiemstra, Dolf Trieschnigg, University of Twente

Relation to Other Areas

Related Areas

- Database Management
- Library and Information Science
- Artificial Intelligence
- Natural Language Processing
- Machine Learning

Taken from: Raymond J. Mooney

Database Management

- Focused on *structured* data stored in relational tables rather than free-form text.
- Focused on efficient processing of well-defined queries in a formal language (SQL).
- Clearer semantics for both data and queries.
- Recent move towards *semi-structured* data (XML) brings it closer to IR.

Taken from: Raymond J. Mooney

Library and Information Science

- Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization).
- Concerned with effective categorization of human knowledge.
- Concerned with citation analysis and *bibliometrics* (structure of information).
- Recent work on *digital libraries* brings it closer to CS & IR.

Taken from: Raymond J. Mooney

Artificial Intelligence

- Focused on the representation of knowledge, reasoning, and intelligent action.
- Formalisms for representing knowledge and queries:
  - First-order Predicate Logic
  - Bayesian Networks
- Recent work on web ontologies and intelligent information agents brings it closer to IR.

Taken from: Raymond J. Mooney

Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse.
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords.

Taken from: Raymond J. Mooney

Natural Language Processing: IR Directions

- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*).
- Methods for identifying specific pieces of information in a document (*information extraction*).
- Methods for answering specific NL questions from document corpora or structured data like FreeBase or Google's Knowledge Graph.

Taken from: Raymond J. Mooney

Machine Learning

- Focused on the development of computational systems that improve their performance with experience.
- Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*).
- Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*).

Taken from: Raymond J. Mooney

Machine Learning: IR Directions

- Text Categorization
  - Automatic hierarchical classification (Yahoo).
  - Adaptive filtering/routing/recommending.
  - Automated spam filtering.
- Text Clustering
  - Clustering of IR query results.
  - Automatic formation of hierarchies (Yahoo).
- Learning for Information Extraction
- Text Mining
- Learning to Rank

Taken from: Raymond J. Mooney