Chapter IR:I

I. Introduction

- Examples of Retrieval Tasks
- Terminology
- Delineation
- Historical Background
- Web Search

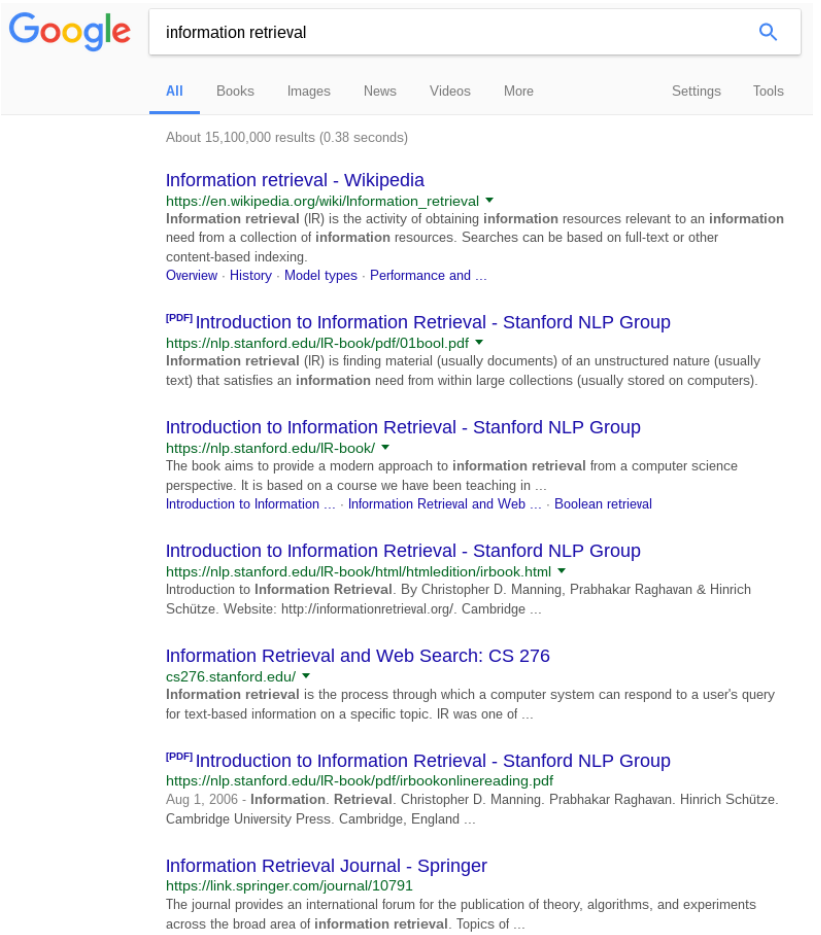
Examples of Retrieval Tasks

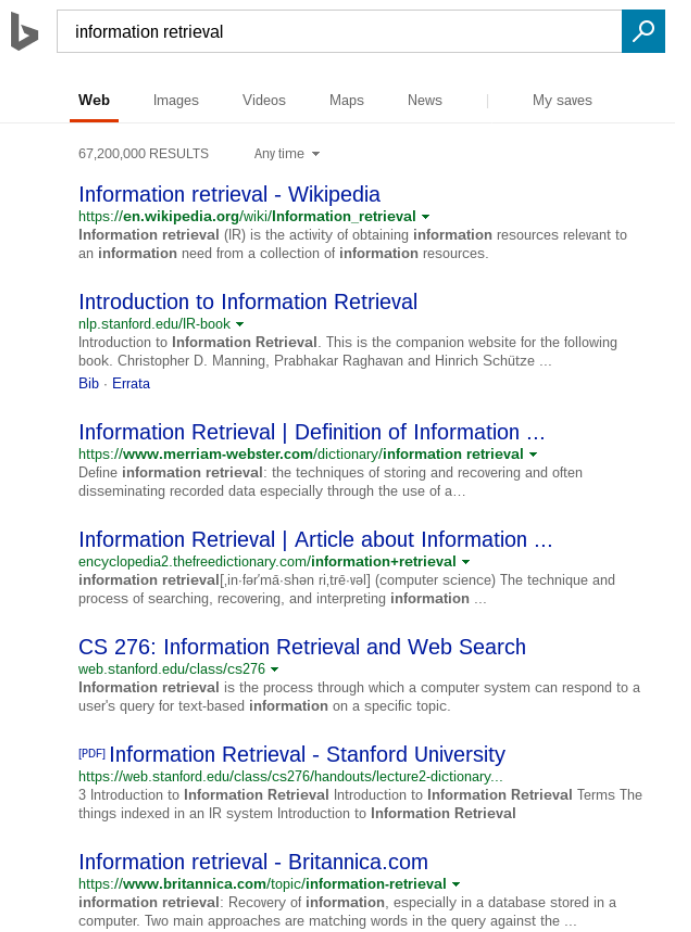
Task: *Learn everything there is to learn about information retrieval.*

Examples of Retrieval Tasks

Task: *Learn everything there is to learn about information retrieval.*

Search for texts that contain ‘information’ and ‘retrieval’.





Remarks:

- Here, the search engine is treated like a database of documents, looking up those that contain specific words a user expects to be contained in a relevant document. Unlike a database, the search engine ranks the retrieved documents with respect to its estimation which of them the user will most likely find useful.
- Compare the different total numbers of search results. The discrepancy may be due to different amounts of documents being indexed, indicating that Bing indexes many more documents than Google. However, search engines have long stopped sharing the numbers of documents they index for competitive reasons, and these numbers are only estimations based on a partial search. Usually, these estimations very much overestimate the actual number of documents that can be delivered.
- How can the number of search results be reduced without loosing many useful documents?

  Using the phrase search operator (i.e., enclosing “information retrieval” in quotes) ensures that the words are contained in order in all documents retrieved, severely reducing the number of results. Conceivably, all documents about information retrieval will contain this particular phrase at least once, whereas most documents talking about something else involving “information” and “retrieval” will not. Interestingly, only about 1.12% of all search engine users know about this or other search operators [White and Morris 2007].
- Results 2, 3, 4, and 6 of the Google result are from the same website. Similarly, results 3, 4, and 7 of the Bing result are dictionary sites. What’s wrong with that?
- The preview text (so-called snippet) of Bing’s result 6 is flawed.

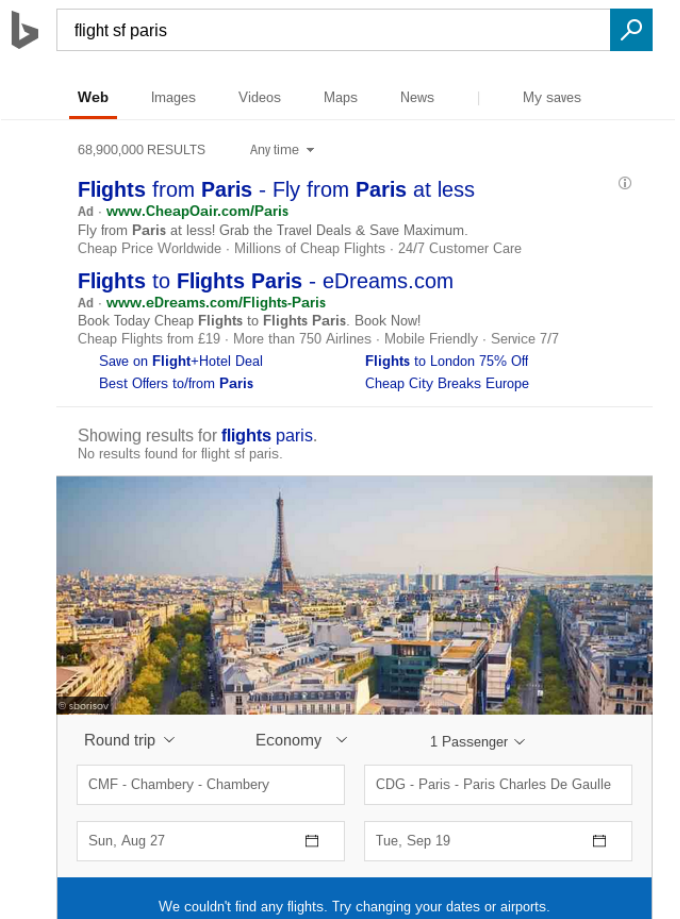
Examples of Retrieval Tasks

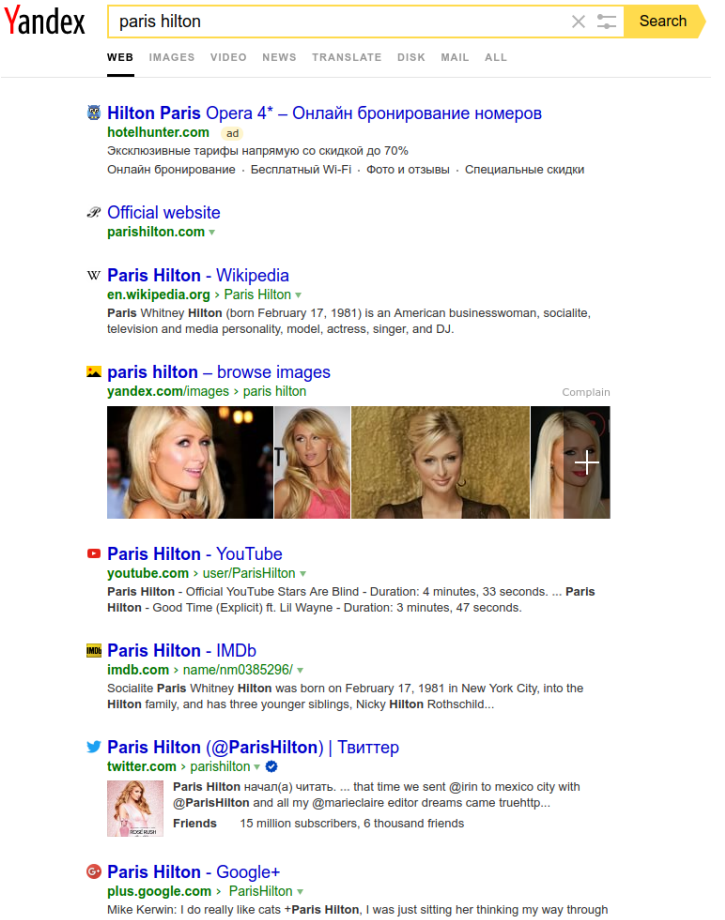
Task: *Plan a trip from San Francisco to Paris, France.*

Examples of Retrieval Tasks

Task: *Plan a trip from San Francisco to Paris, France.*

Search for flights from San Francisco to Paris, and for a hotel.





Remarks:

- Users use informal language to describe their goals. Here, Bing does not understand ‘sf’ as an abbreviation for San Francisco. ‘SFO’, the location identifier of San Francisco Airport, would have worked. Google at least offers a spell correction for ‘sf’ to ‘sfo’.
- Users use ambiguous queries whose interpretation depends on the user’s context. For example, when searching for the Hilton hotel in Paris, Yandex returns results about the celebrity Paris Hilton. Only the ad at the top of the search results hints at the intended interpretation. Searching for “hilton paris” instead yields better results.
- Search engines introduce shortcuts and additional information as so-called oneboxes into search results. The flight search box in Bing’s results is an example.
- Search engines follow a “universal search” paradigm, offering different kinds of results. The images in Yandex’ results are an example.

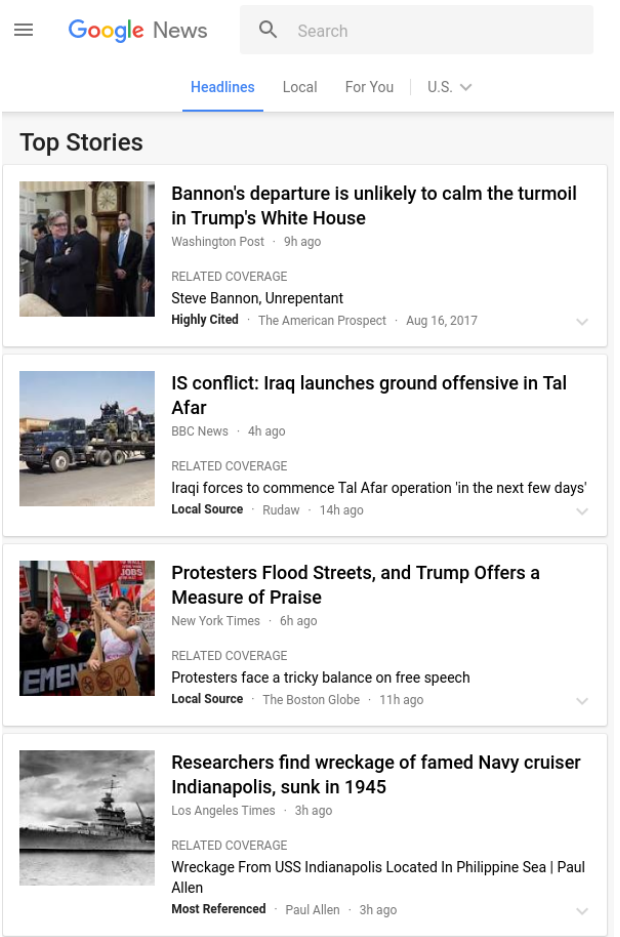
Examples of Retrieval Tasks

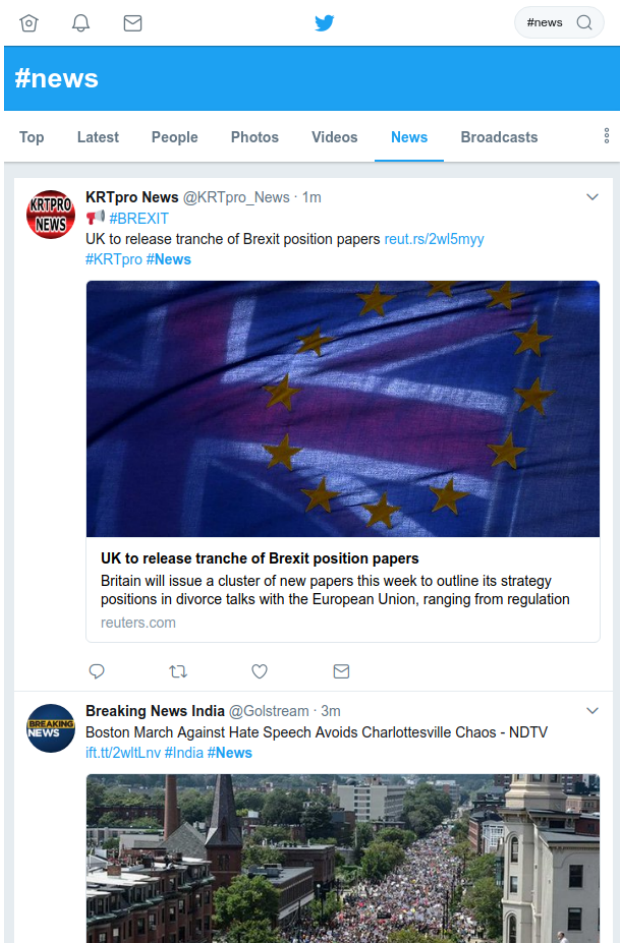
Task: *What were the news of the day?*

Examples of Retrieval Tasks

Task: *What were the news of the day?*

Hit the news feeds.

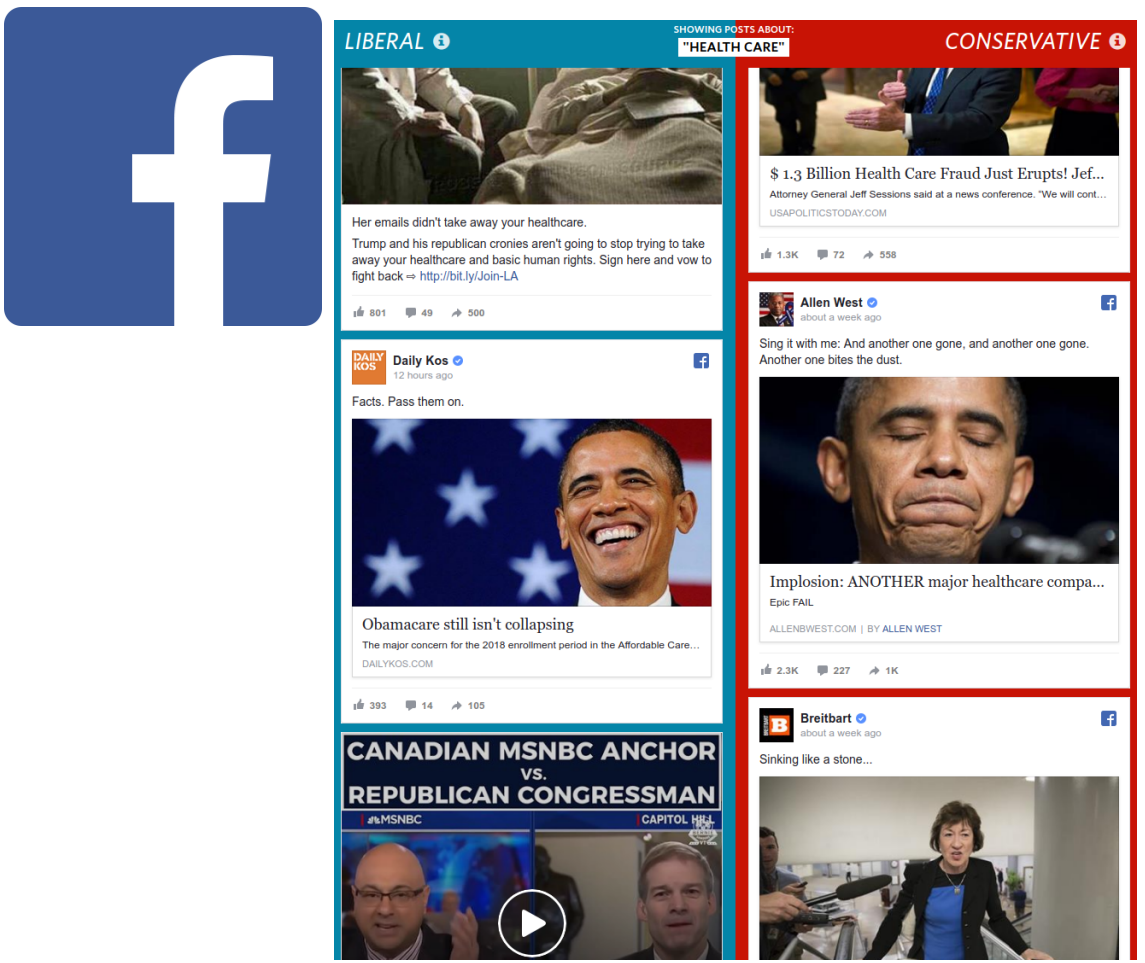




Examples of Retrieval Tasks

Task: *What were the news of the day?*

Hit the news feeds.



Remarks:

- One cannot search for things one does not know. Instead of searching for news, they are explored. Information systems for this purpose include news aggregators and social networks, but also simply the front page of a news outlet. The former recommend news based on user preferences.
- As of 2017, Google News does not show preview snippets for the news, anymore, but only the headlines. While claiming better usability, this change coincides with increasing pressure from news publishers as well as ancillary copyright laws being passed in various jurisdictions.
- Facebook's role in providing Americans with political news has never been stronger–or more controversial. Scholars worry that the social network can create "echo chambers," where users see posts only from like-minded friends and media sources. Facebook encourages users to "keep an open mind" by seeking out posts that don't appear in their feeds.

  To demonstrate how reality may differ for different Facebook users, The Wall Street Journal created two feeds, one "blue" and the other "red." If a source appears in the red feed, a majority of the articles shared from the source were classified as "very conservatively aligned" in a large 2015 Facebook study. For the blue feed, a majority of each source's articles aligned "very liberal." These aren't intended to resemble actual individual news feeds. Instead, they are rare side-by-side looks at real conversations from different perspectives.

  [Wall Street Journal]

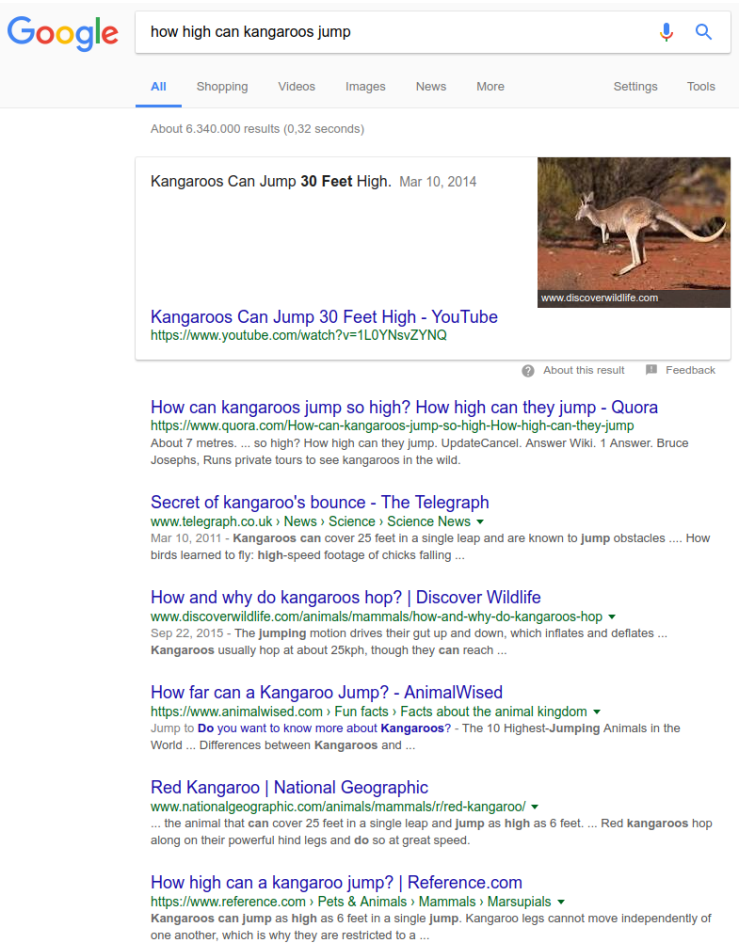
Examples of Retrieval Tasks

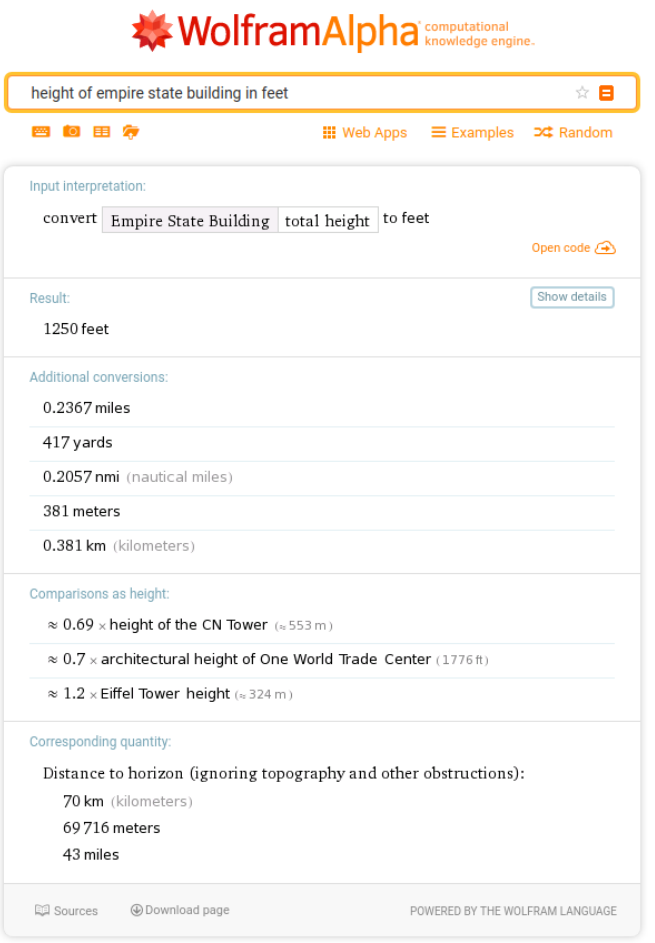
Task: *Answer “Can Kangaroos jump higher than the Empire State Building?”*

Examples of Retrieval Tasks

Task: *Answer "Can Kangaroos jump higher than the Empire State Building?"*

Search for facts

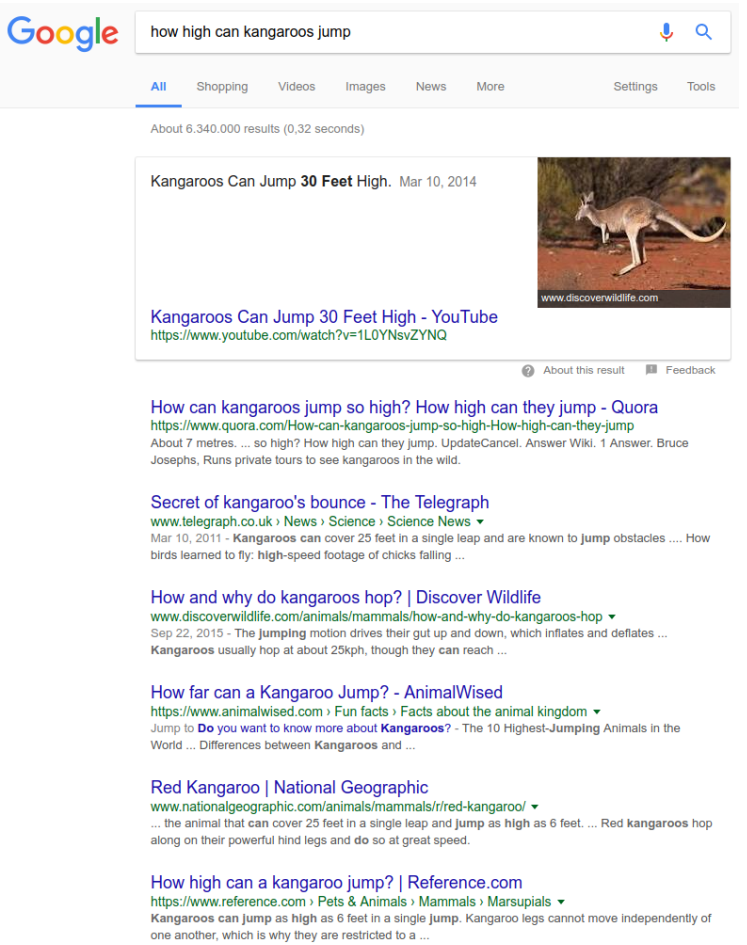


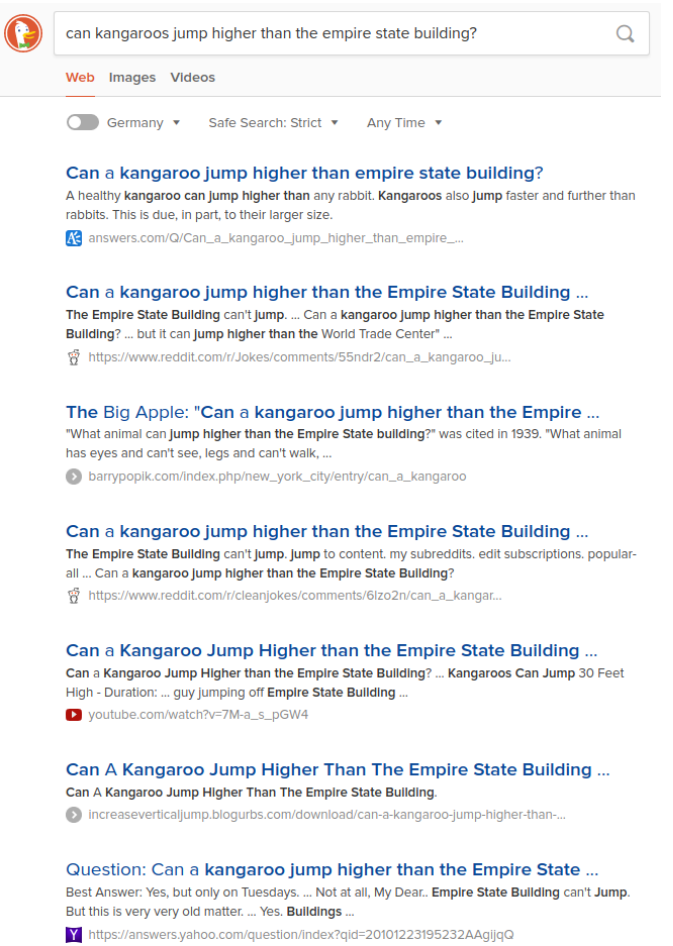


Examples of Retrieval Tasks

Task: *Answer "Can Kangaroos jump higher than the Empire State Building?"*

Search for facts, or ask the question outright.





Remarks:

- Users search for facts; search engines employ various strategies to meet these requests, such as using knowledge bases like Wikidata, or by by extracting factual information from web pages (e.g., Google's Knowledge Graph).
- Google's highlighted top search result appears to answer the question, but the given height is false; it mixes height with distance. Google does not necessarily validate the truth of a statement, but only returns the ones best matching the query. Other snippets mention distances inconsistent with the top one. The two bottom-most snippets claim that kangaroos can jump only about 6 feet high. In any case, a YouTube video may not be a reliable source.
- WolframAlpha allows for asking questions requiring computation, resorting to a knowledge base to fill in required facts.
- Asking the original question directly shows that it is a well-known one; some snippets of DuckDuckGo reveal that it is a riddle by giving away the answer. The odd formulation of the riddle might have revealed this already.

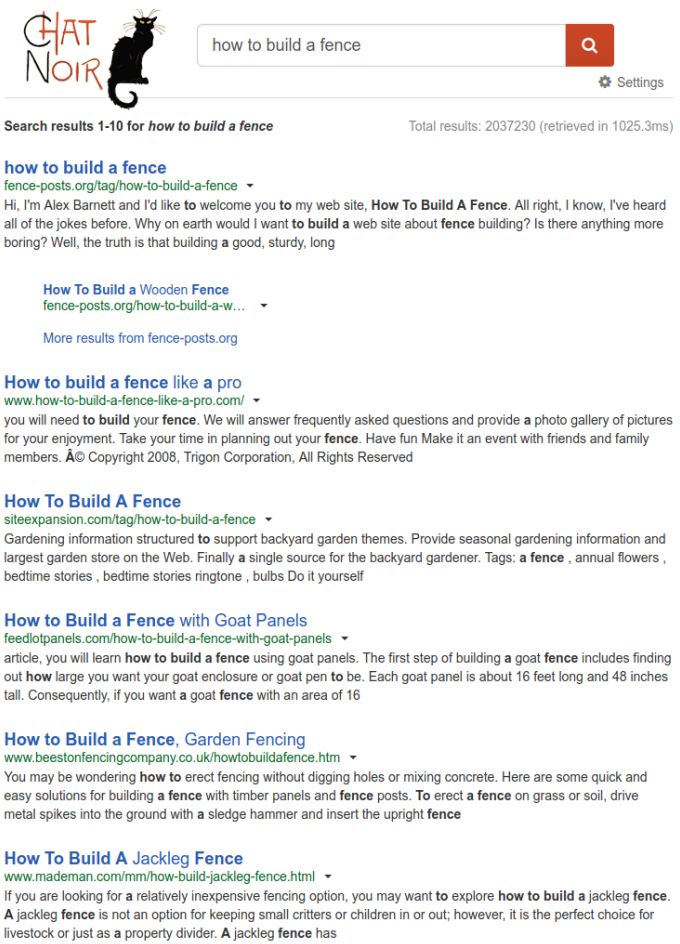
Examples of Retrieval Tasks

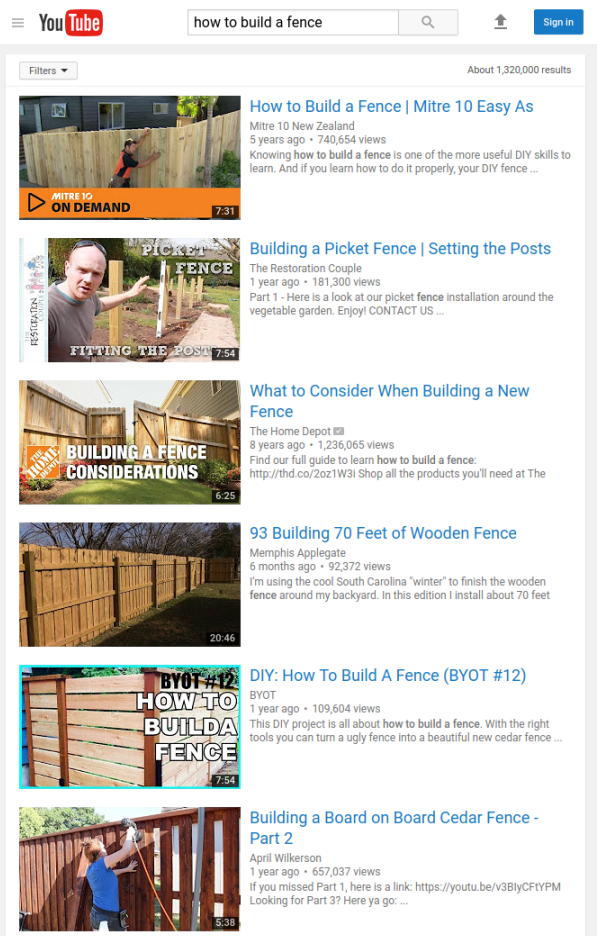
Task: *Build a fence.*

Examples of Retrieval Tasks

Task: *Build a fence.*

Search for tutorials.





Remarks:

- Users search for guidelines on complex tasks. Besides textual information, this includes instructive multimedia contents, e.g., from YouTube.
- ChatNoir is the only publicly available research search engine that operates at scale.

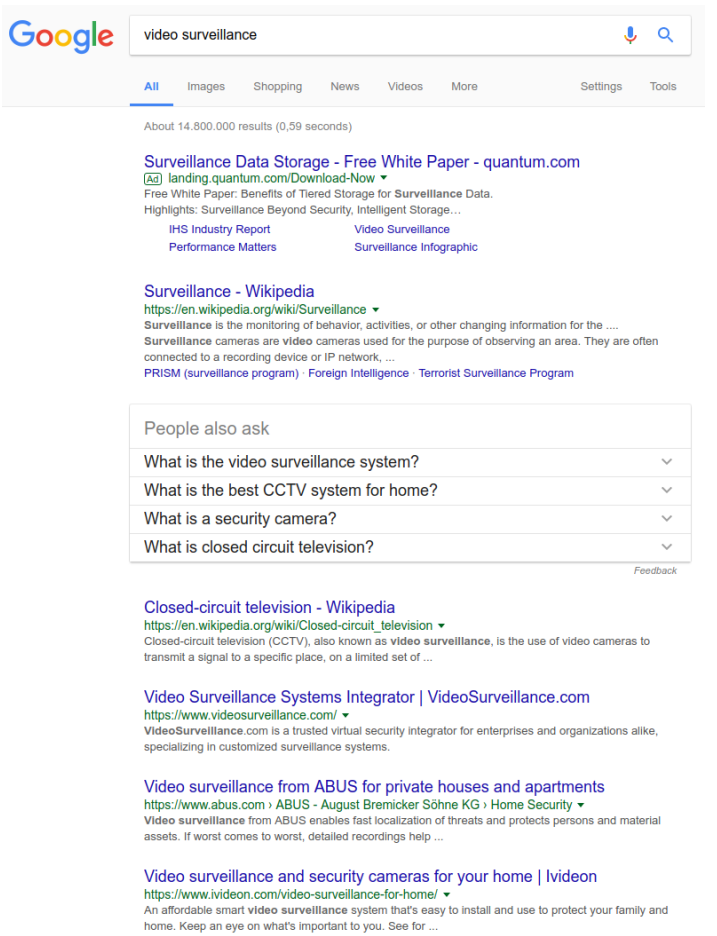
Examples of Retrieval Tasks

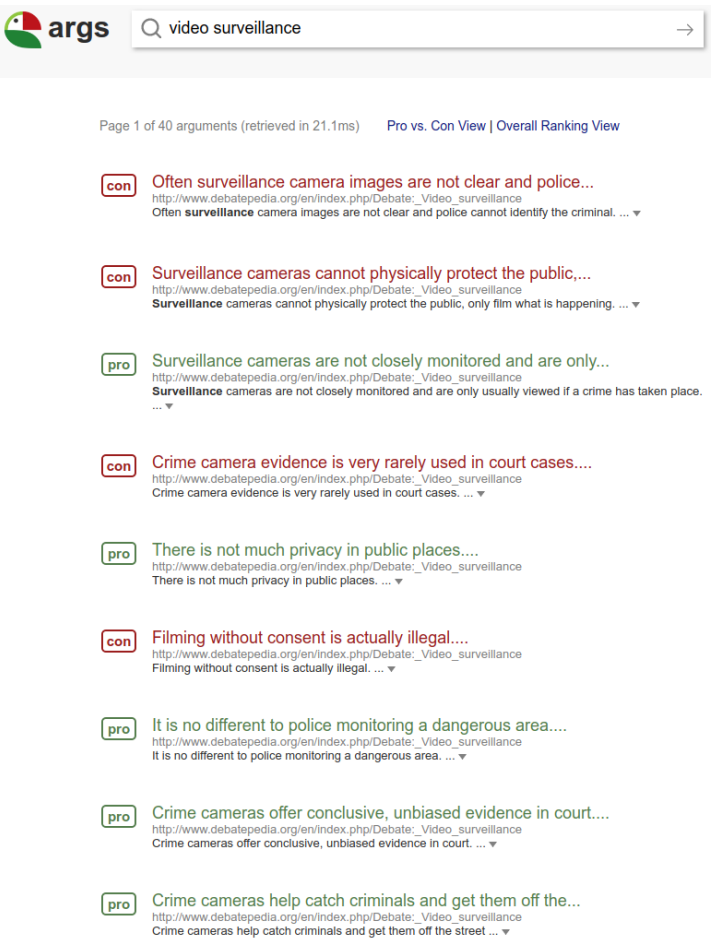
Task: *Write an essay on video surveillance.*

Examples of Retrieval Tasks

Task: *Write an essay on video surveillance.*

Search for other's opinions on video surveillance.





Remarks:

- Users search for other peoples' opinions and reasoning on controversial topics or when deciding what product to buy.
- Google's results include related questions from question answering platforms, concisely presented in a onebox. Yet, there's no hint at the controversiality of the topic. The results referring to Wikipedia are the only way for a user to learn the topic's background, all other results are related to shopping.
- Search engines specialized to argument retrieval, such as Args, retrieve argumentative information alongside stance (pro or con).

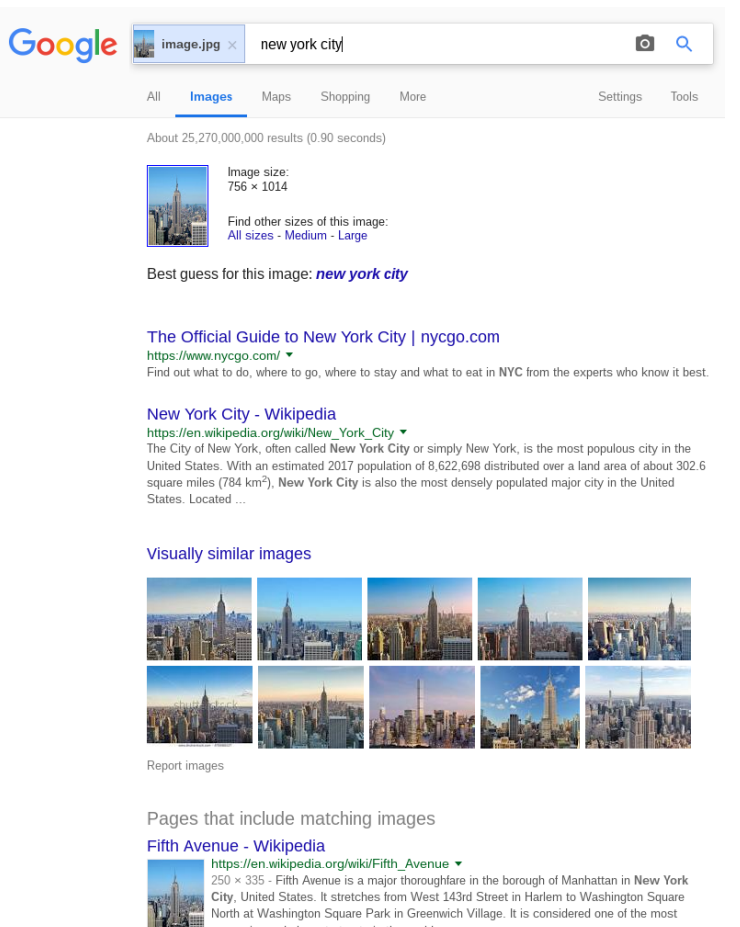
Examples of Retrieval Tasks

Task: *Given an example image, find more like it.*

Examples of Retrieval Tasks

Task: *Given an example image, find more like it.*

The image is an example of the information sought after.



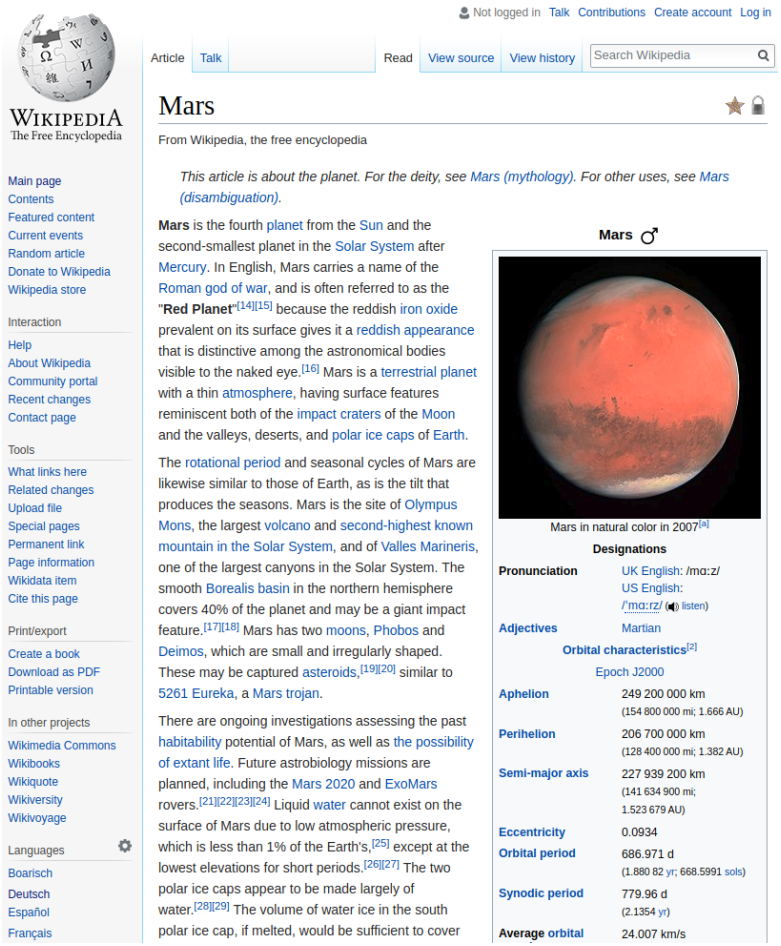
Examples of Retrieval Tasks

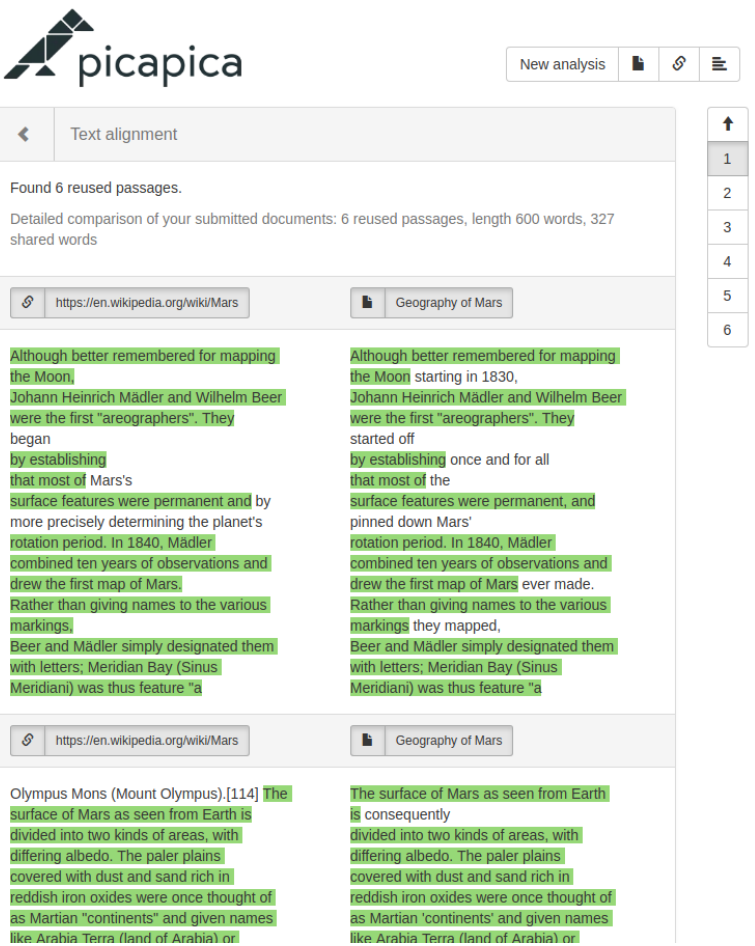
Task: *Given an example text, find more like it.*

Examples of Retrieval Tasks

Task: *Given an example text, find more like it.*

The text is an example of the information sought after.





Remarks:

- Users sometimes cannot express their need as a textual query, but provide an object that best exemplifies the information sought.
- Some search engines are tailored to searching for specific multimedia examples, such as images or audio.
- Using a text as an example, there can be two goals: finding other texts talking about the same subject, or finding other texts which share reused text passages with the text in question. For example, Google Scholar offers the search facet “Related Articles” to search for articles related to one designated from a prior search. Picapica is a search engine for text reuse.

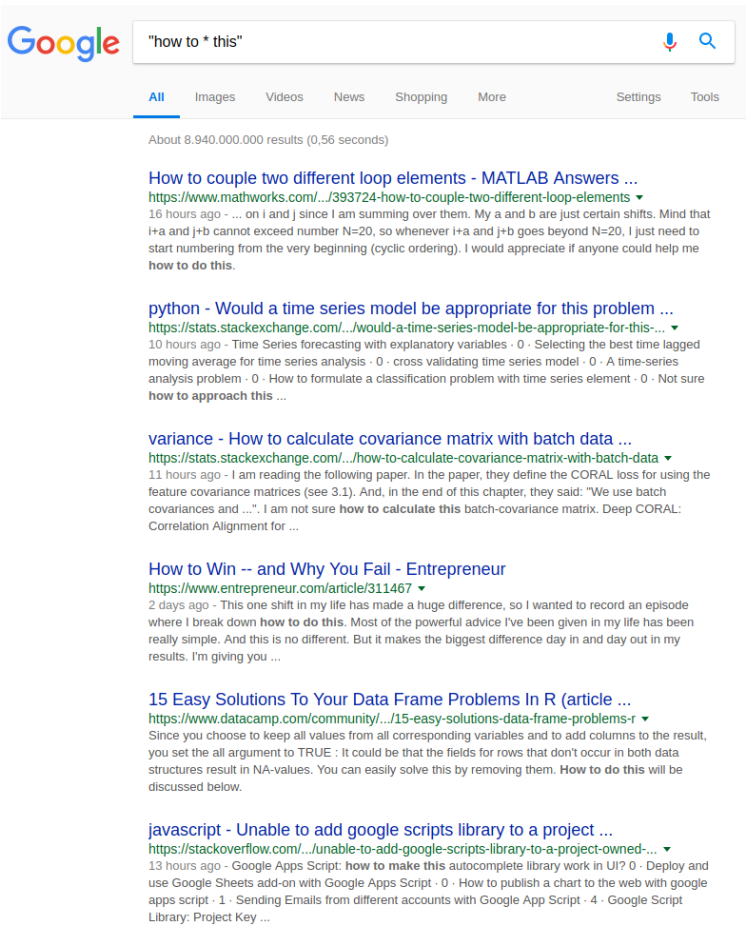
Examples of Retrieval Tasks

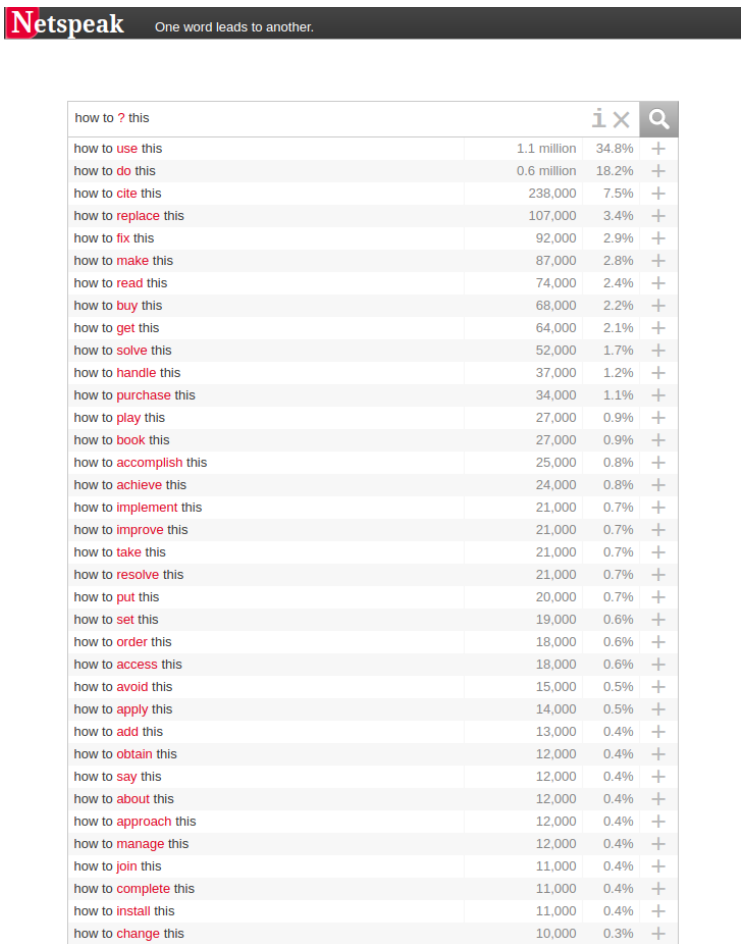
Task: *Find out what people commonly write in the phrase* `how to ? this.`

Examples of Retrieval Tasks

Task: *Find out what people commonly write in the phrase* `how to ? this.`

Use wildcard search operators to find matching phrases.





Remarks:

- Users “misuse” search engines and all other kinds of tools to achieve goals beside their originally intended purpose due to lack of specialized tools, or lack of knowledge of their existence.
- Many web search engines support some wildcard search operators, but they cannot be used to solve this kind of retrieval task. The search engine still interprets the query in terms of its contents, ranking documents according to their relevance to the query. Moreover, only few search results fit on a single page, while many more alternatives may be in practical use.
- Netspeak indexes only short phrases alongside their web frequencies, offering a wildcard search interface tailored to searching for usage commonness.

Terminology

Information science distinguishes the concepts data, information, and knowledge.

Definition 1 (Data)

A sequence of symbols recorded on a storage medium.

Definition 2 (Information)

A useful portion of data.

useful: the data in question serves as a means to an end for a person

Definition 3 (Knowledge)

Product of reflection upon information combined with experience to the extent of recognizing acquaintance, reaching belief, or justified understanding.

Examples: One might claim “I know 2+2=4,” but also “I know God exists.”

Remarks:

- Data is typically organized into documents, each containing purposefully chosen partitions of data. Examples: a book, a video tape. A digital document corresponds to specific sequences of bits on a digital storage medium. Example: files on a hard drive formatted with a file system.
- Definitions of the three concepts, but especially those of information and knowledge, differ wildly among scholars of information science. [Zins 2007] collects 44 different attempts.
- Epistemology studies the nature of knowledge, justification, and the rationality of belief. In mathematics, it is known that 2+2=4, but there is also knowing how to add two numbers, and knowing a person (e.g., oneself), place (e.g., one's hometown), thing (e.g., cars), or activity (e.g., addition). Some philosophers think there is an important distinction between "knowing that" (know a concept), "knowing how" (understand an operation), and "acquaintance-knowledge" (know by relation), with epistemology being primarily concerned with the first of these. Plato's definition of knowledge as *justified true belief* was widely accepted until the 1960s. At this time, a paper written by the American philosopher Edmund Gettier entitled "Is Justified True Belief Knowledge?" called into question the theory of knowledge that had been dominant among philosophers for thousands of years. [Wikipedia]
- The fundamental question of rationality: "Why do you believe what you believe?", or alternatively, "What do you think you know, and how do you think you know it?" [LessWrong]

Terminology

Definition 4 (Information System)

An organized system for collecting, creating, storing, processing, and distributing information, typically including hardware, software, users, and the data itself.

Definition 5 (Information Need)

A user's desire to locate and obtain information to satisfy a conscious or unconscious goal.

Definition 6 (Relevance)

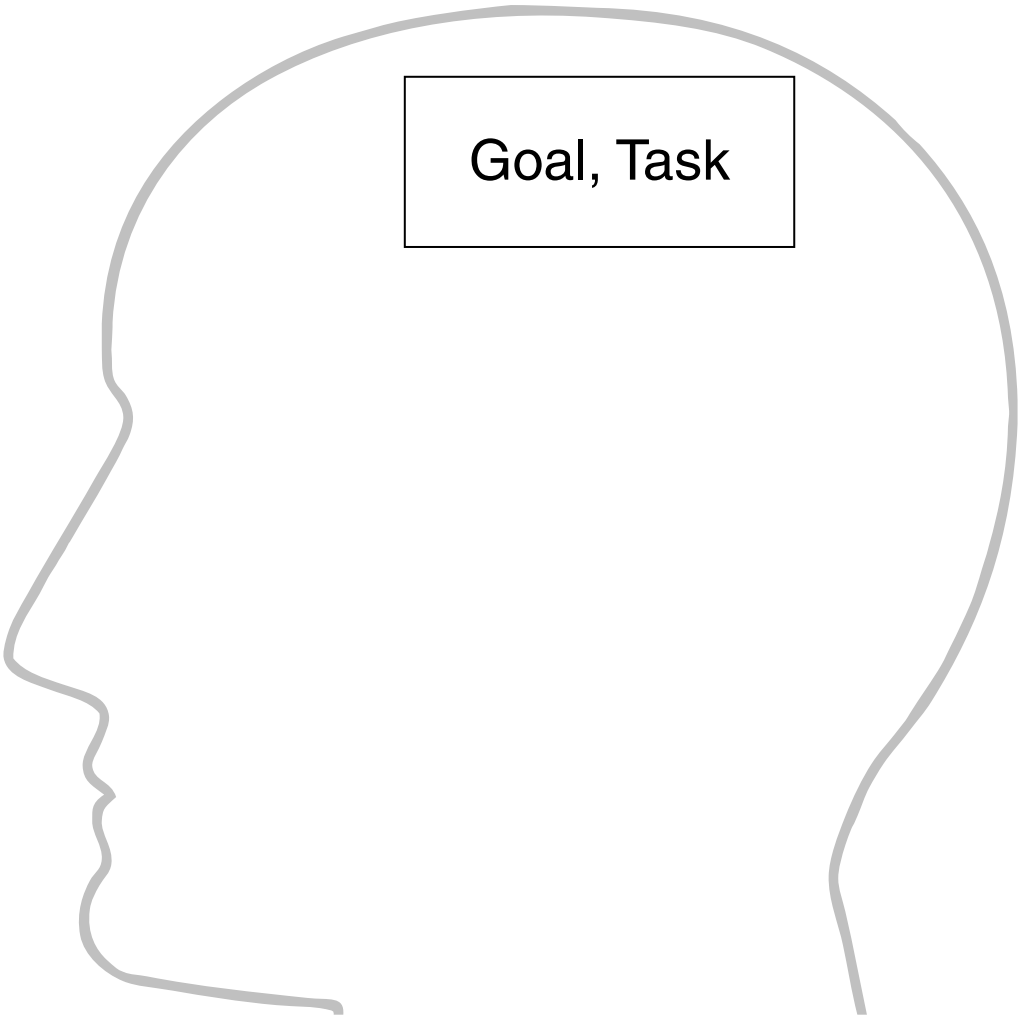
The degree to which a portion of data meets the information need of a user.

A portion of data is said to be relevant to a user's information need, if it informs the user. The closer it brings the user to satisfy their goal, the more relevant it is.

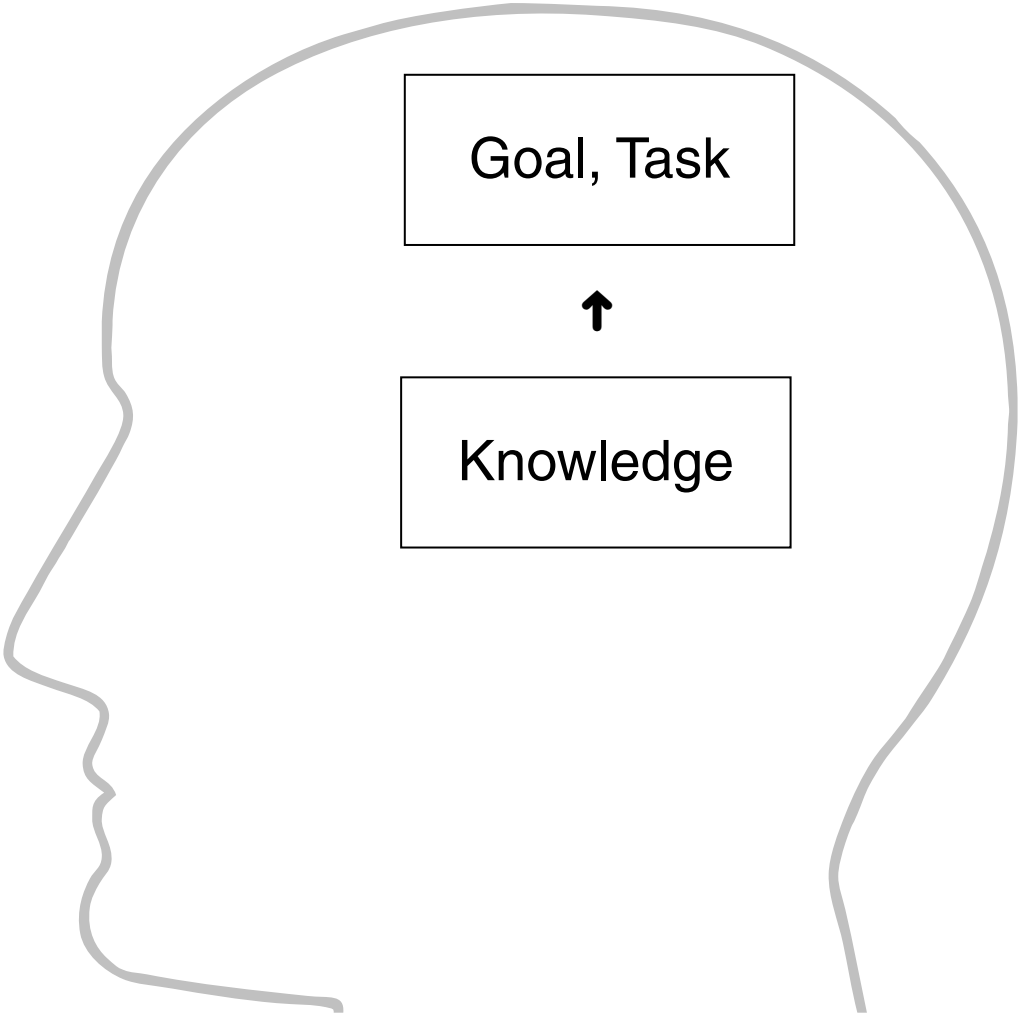
Remarks:

- Information need refers to a cognitive need that is perceived when a gap of knowledge is encountered in the pursuit of a goal.
- The study of information needs has been generalized to the study of information behavior, i.e. “the totality of human behavior in relation to sources and channels of information, including both active and passive information-seeking, and information use.” [Wilson 2000]

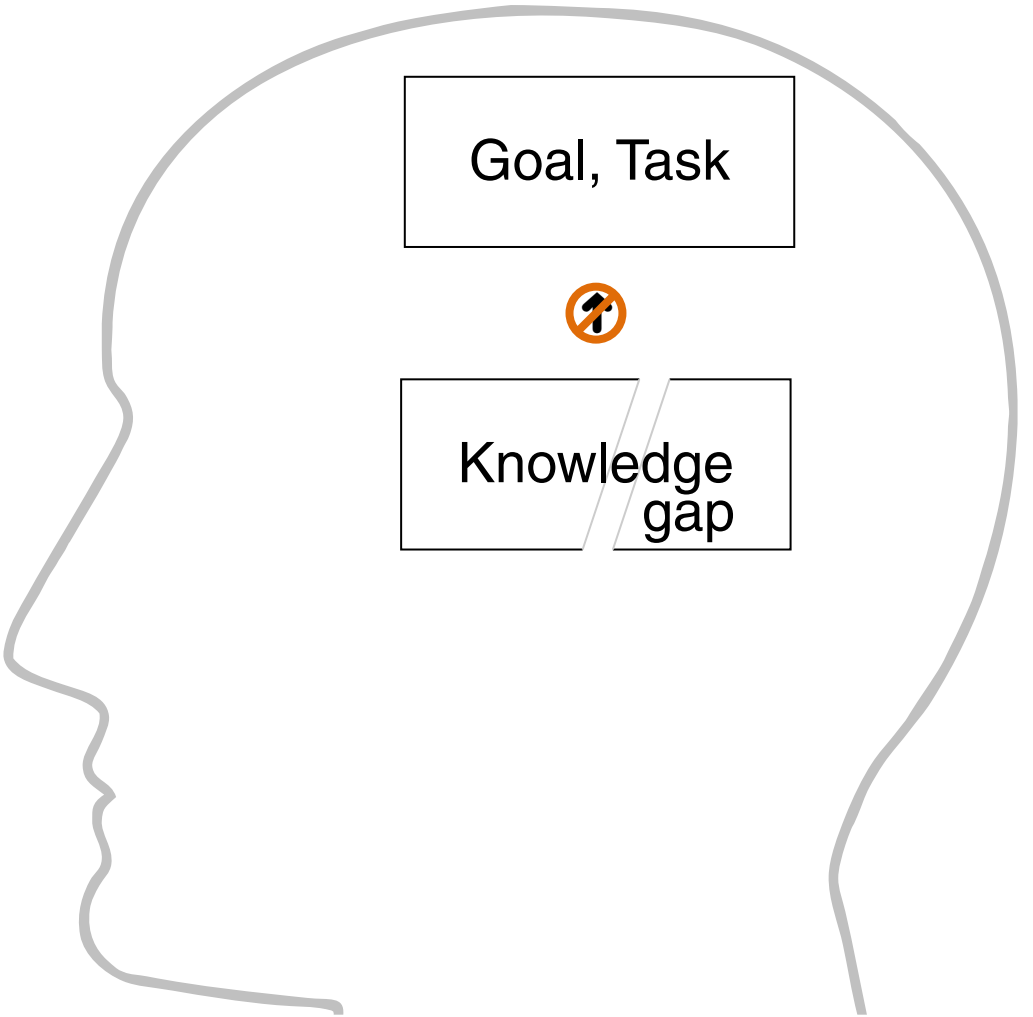
Terminology



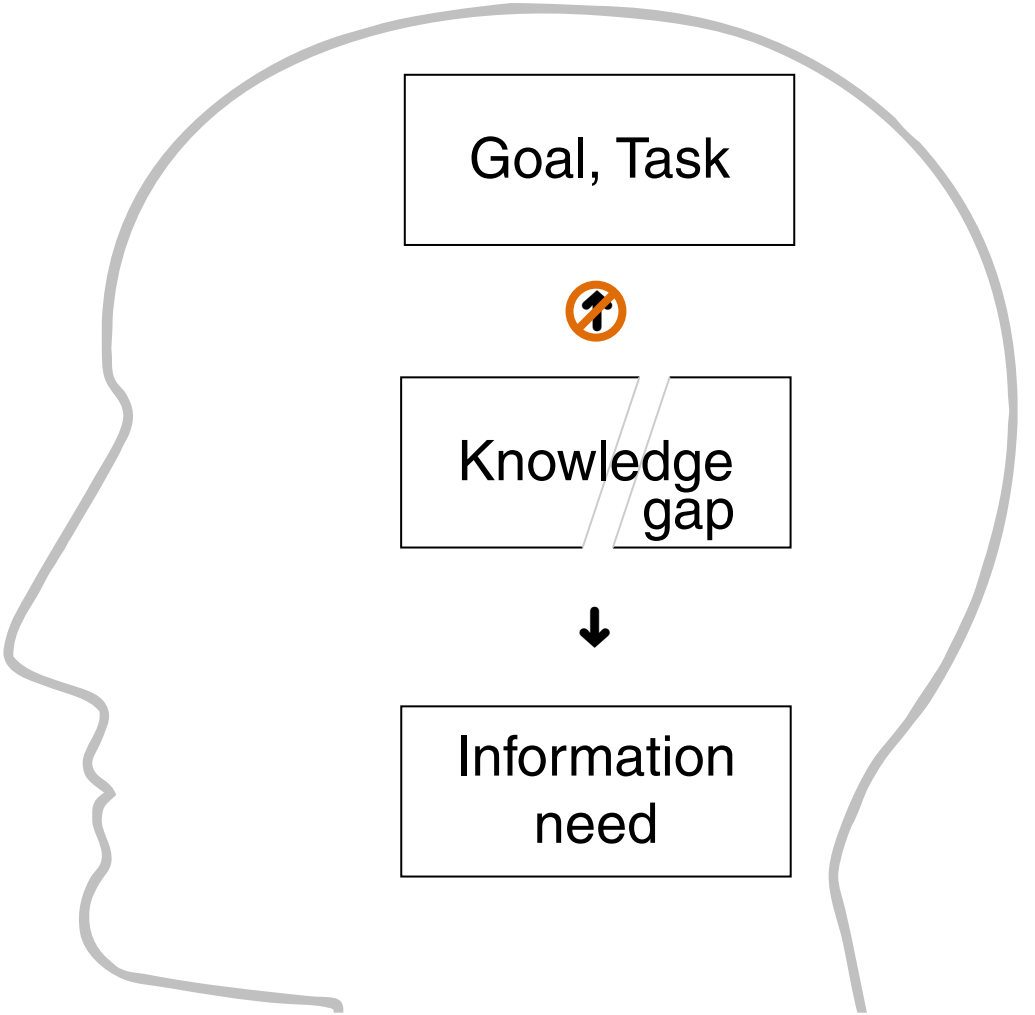
Terminology



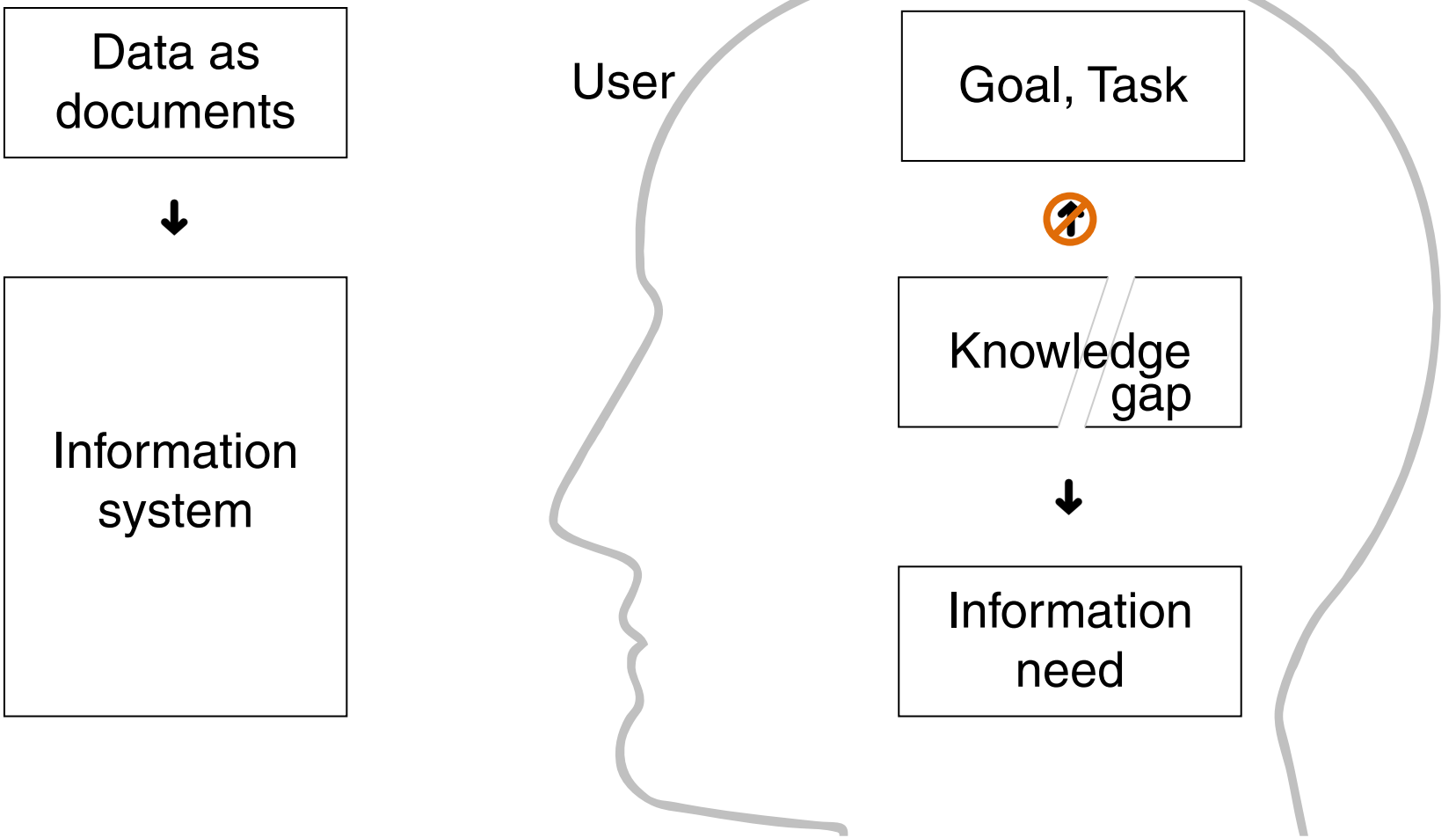
Terminology



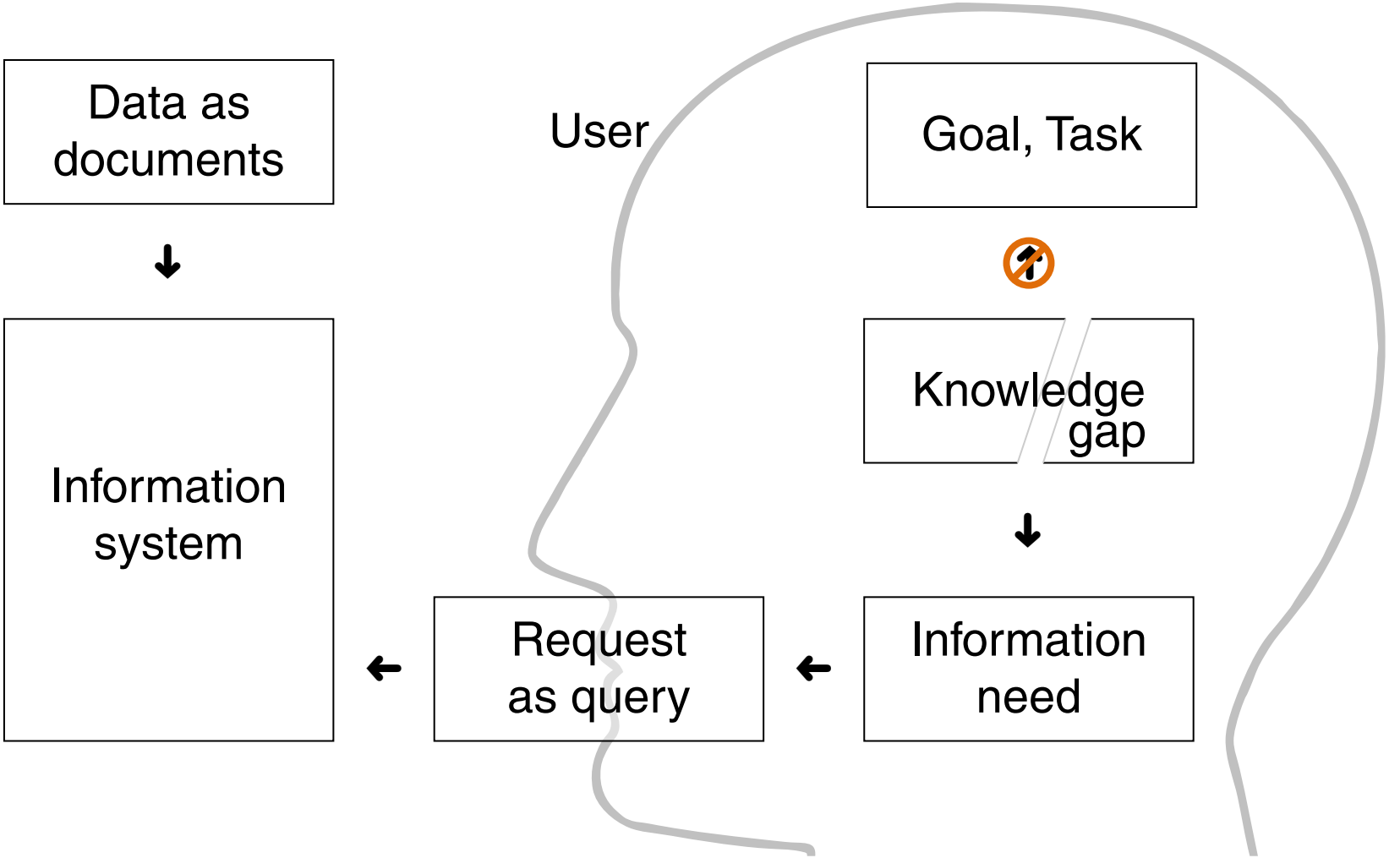
Terminology



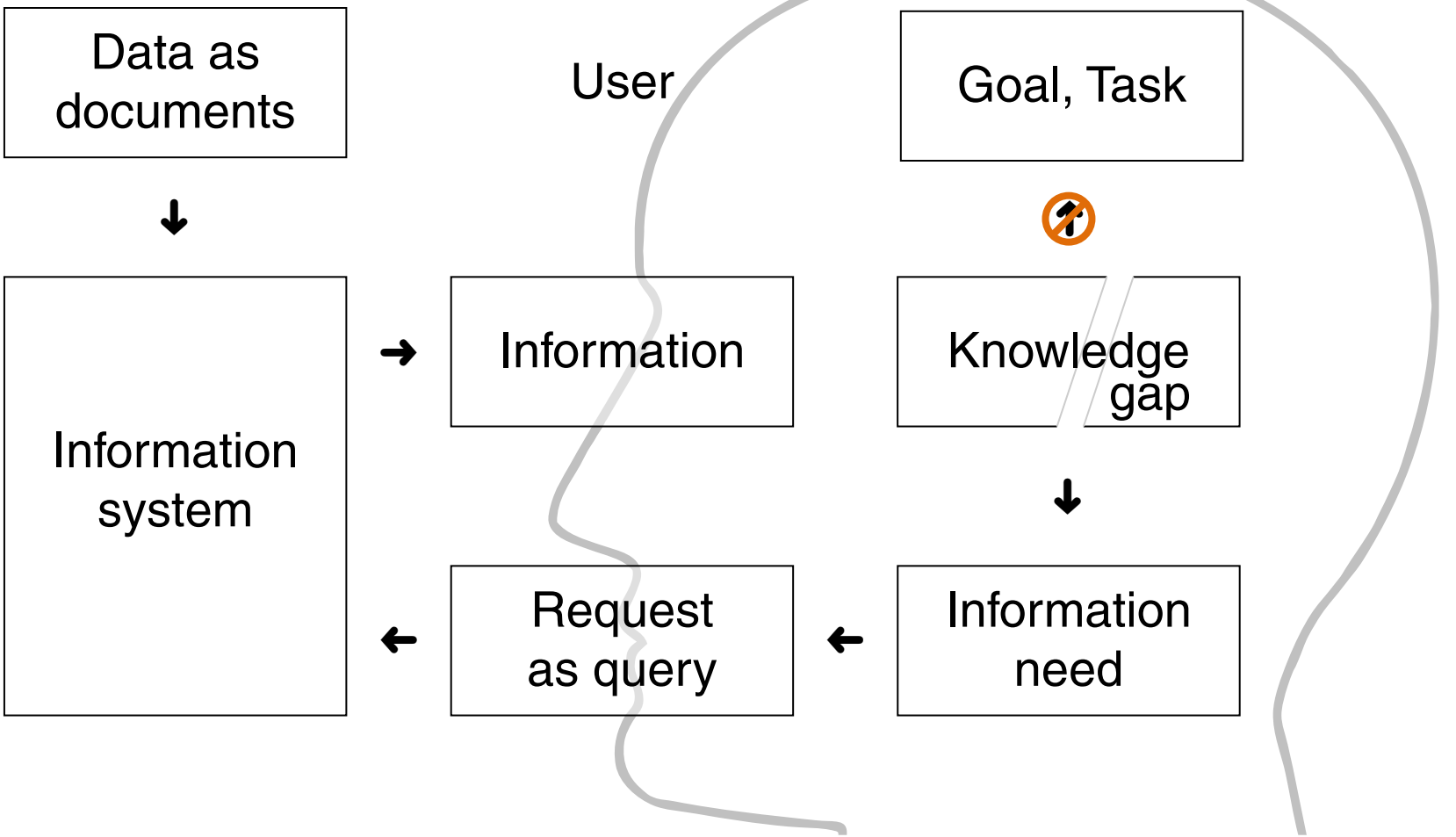
Terminology



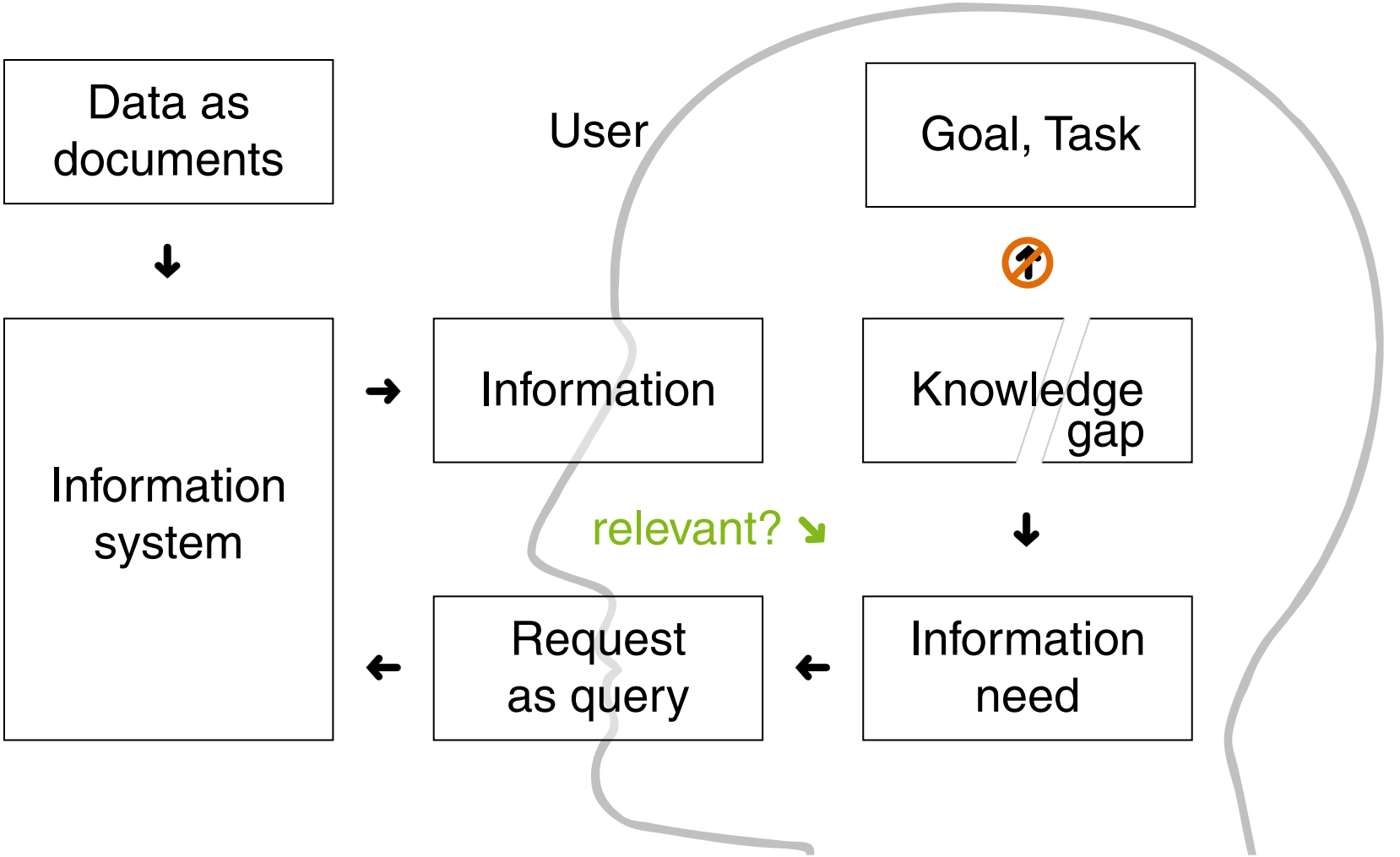
Terminology



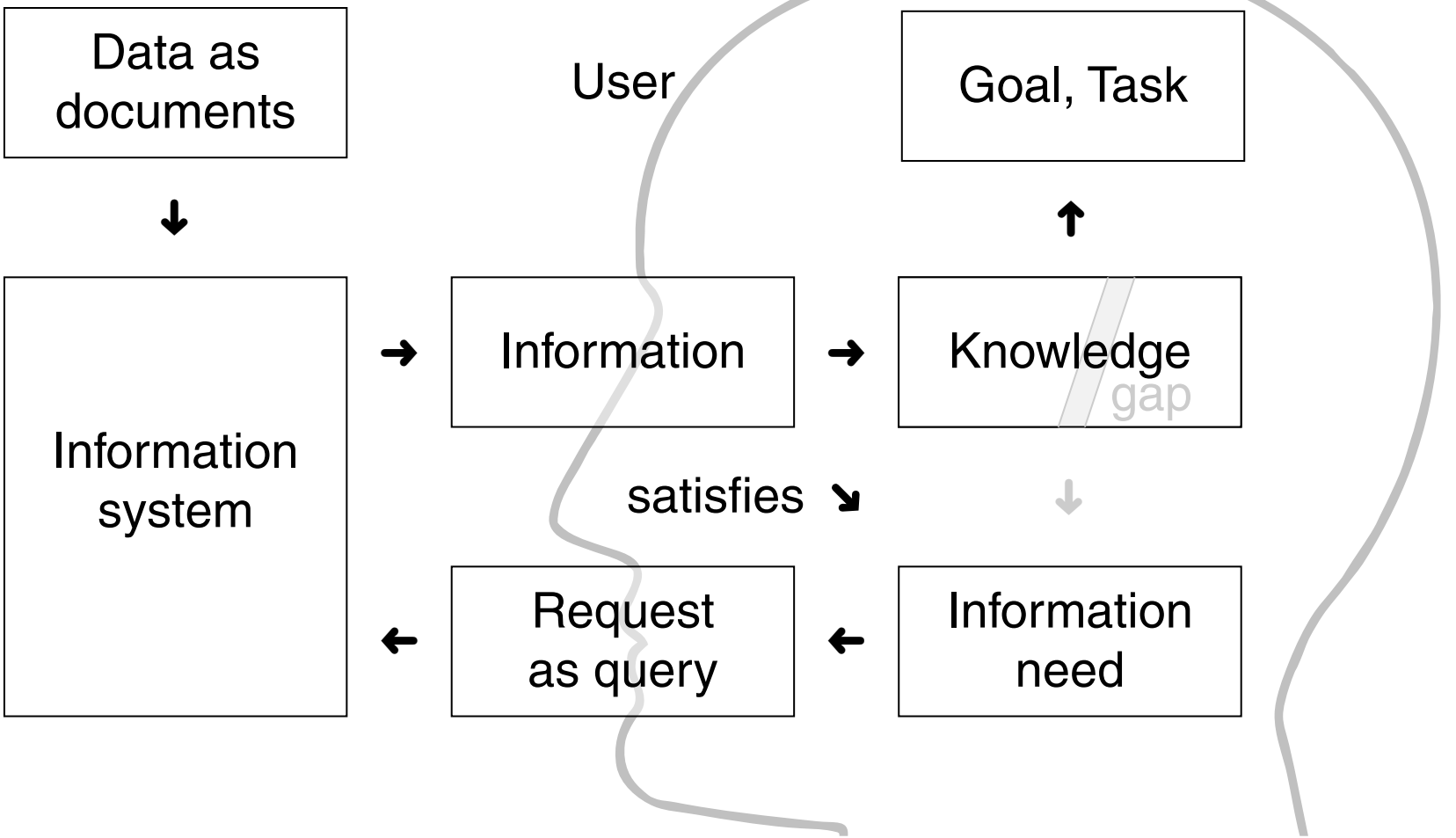
Terminology



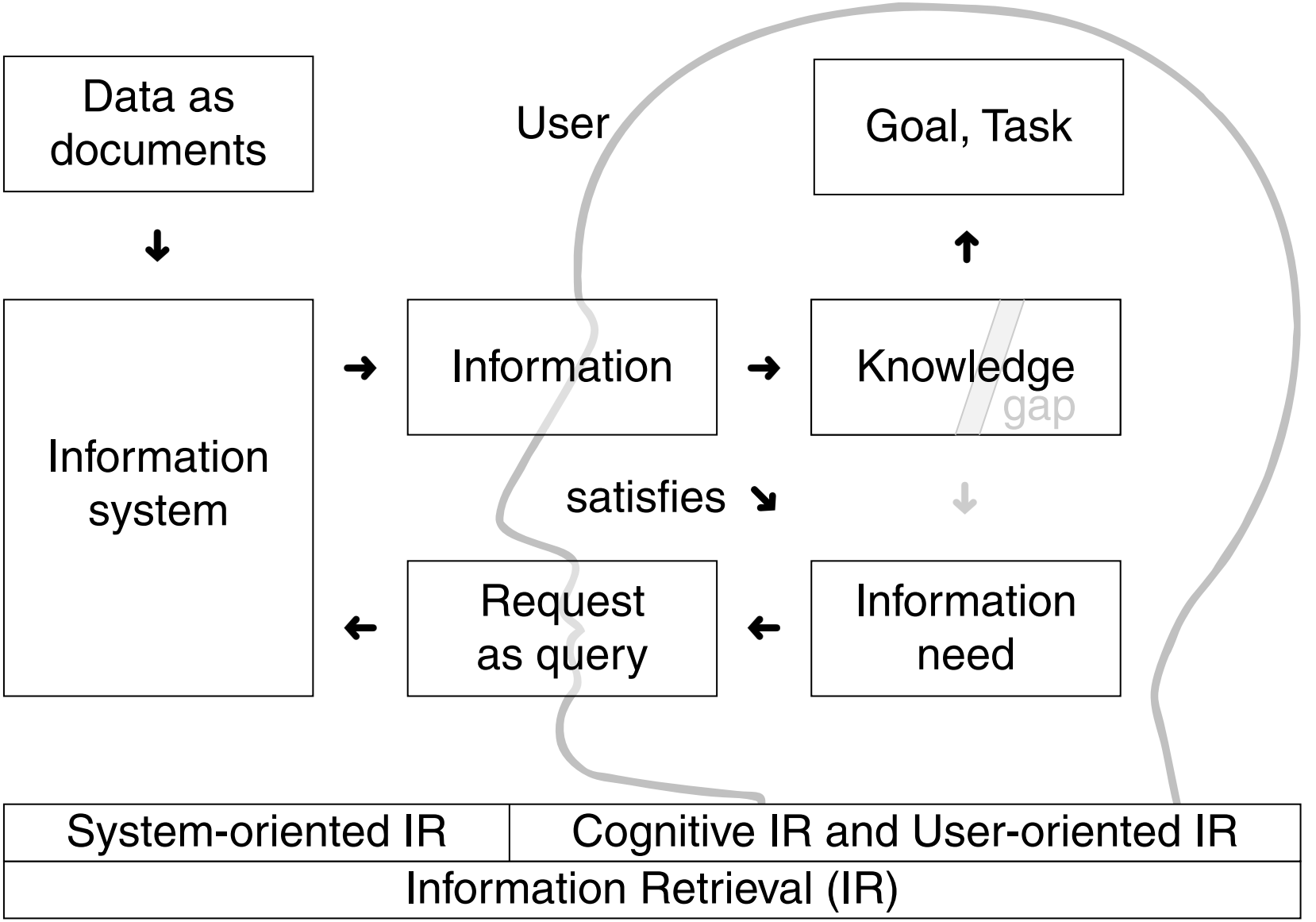
Terminology



Terminology



Terminology



Terminology

Definition 7 (Information Retrieval)

The activity of obtaining information relevant to an information need from data.

As a research field, information retrieval (IR) studies the role of information systems in transferring knowledge via data, as well as the design, implementation, evaluation, and analysis of such systems.

Terminology

Definition 7 (Information Retrieval)

The activity of obtaining information relevant to an information need from data.

As a research field, information retrieval (IR) studies the role of information systems in transferring knowledge via data, as well as the design, implementation, evaluation, and analysis of such systems.

- Role of information systems:

System-oriented IR	information technology
Cognitive IR	human interaction with information technology
User-oriented IR	information systems as sociotechnical systems

- Design — architecture, algorithms, interfaces
- Implementation — hardware, deployment, maintenance
- Evaluation — effectiveness and efficiency
- Analysis — experiments, user studies, log analysis

Terminology

Definition 7 (Information Retrieval)

The activity of obtaining information relevant to an information need from data.

As a research field, information retrieval (IR) studies the role of information systems in transferring knowledge via data, as well as the design, implementation, evaluation, and analysis of such systems.

Major challenges of IR:

1. Vague queries
   Goal not a priori clear; potential vocabulary mismatch; requires interaction / dialog to refine; dependence on previous results; often combining information from multiple data sources.
2. Incomplete and uncertain knowledge
   Results from the limitations of accurately representing semantics; some domains are inherently incomplete / uncertain (e.g., opinion topics like politics, evidence vs. belief topics like religion, interpretation topics like history and news, biased data collections like the web)
3. Accuracy of results
4. Efficiency

Remarks:

- Definitions of system-oriented IR, cognitive IR, and user-oriented IR are vague.
- The goal in real-life IR is to find useful information for an information need situation. [...] In practice, this goal is often reduced to finding documents, document components, or document surrogates, which support the user (the actor) in constructing useful information for her / his information need situation. [...]

  The goal of systems-oriented IR research is to develop algorithms to identify and rank a number of (topically) relevant documents for presentation, given a (topical) request. On the theoretical side, the goals include the analysis of basic problems of IR (e.g., the vocabulary problem between the recipient and the generator, document and query representation and matching) and the development of models and methods for attacking them. [...]

  The user-oriented and cognitive IR research focused [...] on users' problem spaces, information problems, requests, interaction with intermediaries, interface design and query formulation [...]. [Ingwersen 2005]
- User-oriented IR moves the orientation from a "closed system" in which the IR "engine" is tuned to handle a given set of documents and queries, to one that integrates the IR system within a broader information use environment that includes people, and the context in which they are immersed. [Toms 2013]
- "Sociotechnical" refers to the interrelatedness of social and technical aspects of an organization. Sociotechnical systems in organizational development is an approach to complex organizational work design that recognizes the interaction between people and technology in workplaces. The term also refers to the interaction between society's complex infrastructures and human behavior. [Wikipedia]

Delineation

Databases, Data Retrieval [van Rijsbergen 1979]

	Data Retrieval	Information Retrieval
Matching	exact	partial match, best match
Inference	deduction	induction
Model	deterministic	probabilistic
Classification	monothetic	polythetic
Query language	artificial	natural
Query specification	complete	incomplete
Items wanted	matching	relevant
Error response	sensitive	robust

Remarks:

- A major difference between information retrieval (IR) systems and other kinds of information systems is the intrinsic uncertainty of IR. Whereas for database systems, an information need can always (at least for standard applications) be mapped precisely onto a query formulation, and there is a precise definition of which elements of the database constitute the answer, the situation is much more difficult in IR; here neither a query formulation can be assumed to represent uniquely an information need, nor is there a clear procedure that decides whether a database object is an answer or not. Boolean IR systems are not an exception from this statement; they only shift all problems associated with uncertainty to the user. [Fuhr 1992]
- In data retrieval we are most likely to be interested in a monothetic classification, that is, one with classes defined by objects possessing attributes both necessary and sufficient to belong to a class. In IR such a classification is on the whole not very useful, in fact more often a polythetic classification is what is wanted. In such a classification each individual in a class will possess only a proportion of all the attributes possessed by all the members of that class. Hence no attribute is necessary nor sufficient for membership to a class. [van Rijsbergen 1979]

  Example: in a given database, persons are required to possess the attributes name, birth date, gender, etc.; documents about persons may each mention any given subset of these attributes.

Delineation

Semiotics

Data as documents	→	Information	→	Knowledge
syntax layer, sigmatic layer		semantic layer		pragmatic layer
Example: Database as collection of signs		Example: Interpretation of the signs of a database		Transformation of the information to obtain the desired knowledge
10/1/1955		Date		Have I missed my mother's birthday?

Information retrieval is an associative search that particularly addresses the semantics and pragmatics of documents.

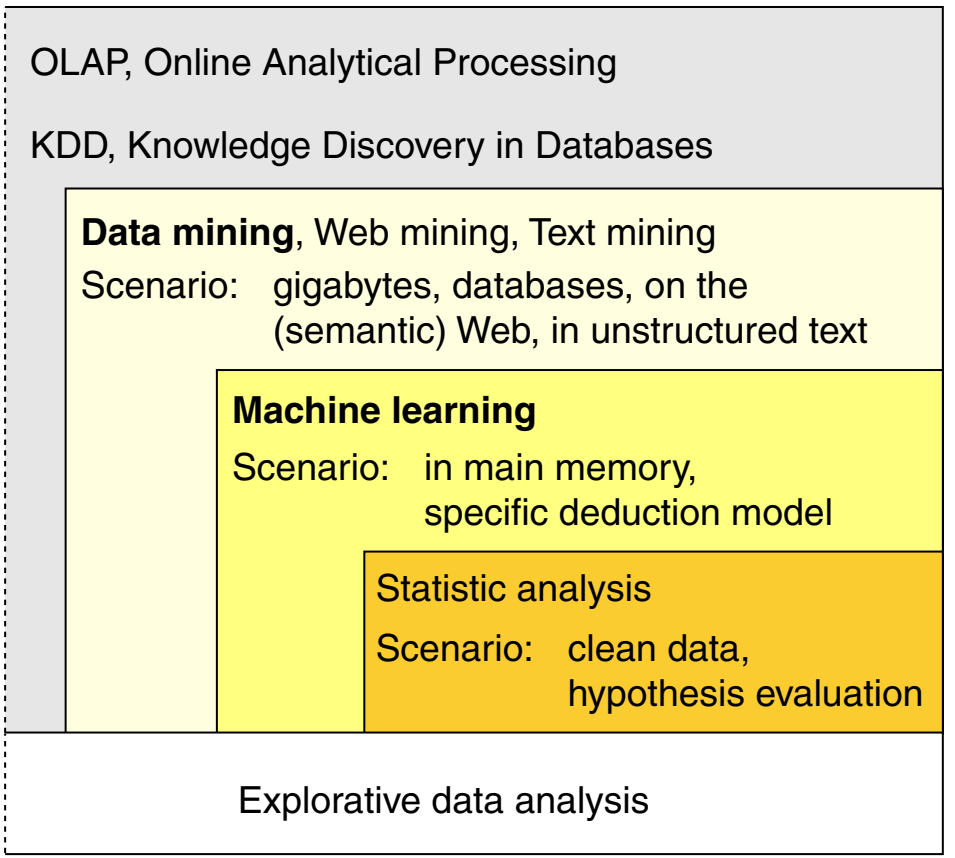
Remarks:

- Semiotics (“sign theory,” derived from greek) is the study of meaning-making, the study of sign process (semiosis) and meaningful communication. Modern semiotics was defined by C.S. Peirce and C.W. Morris, who divided the field into three basic layers: the relations between signs (syntax), those between signs and the things signified (semantics), and those between signs and their users (pragmatics). [Wikipedia]
- K. Georg further distinguishes the relations between signs and the object to which they belong (sigmatics). [Wikipedia]
- The semiotic layers can be aligned with the basic concepts of information science, the latter forming elements of the former:

Layer	Element
Syntax	Sign (e.g., character)
Sigmatics	Data, Document
Semantics	Information
Pragmatics	Message, Knowledge

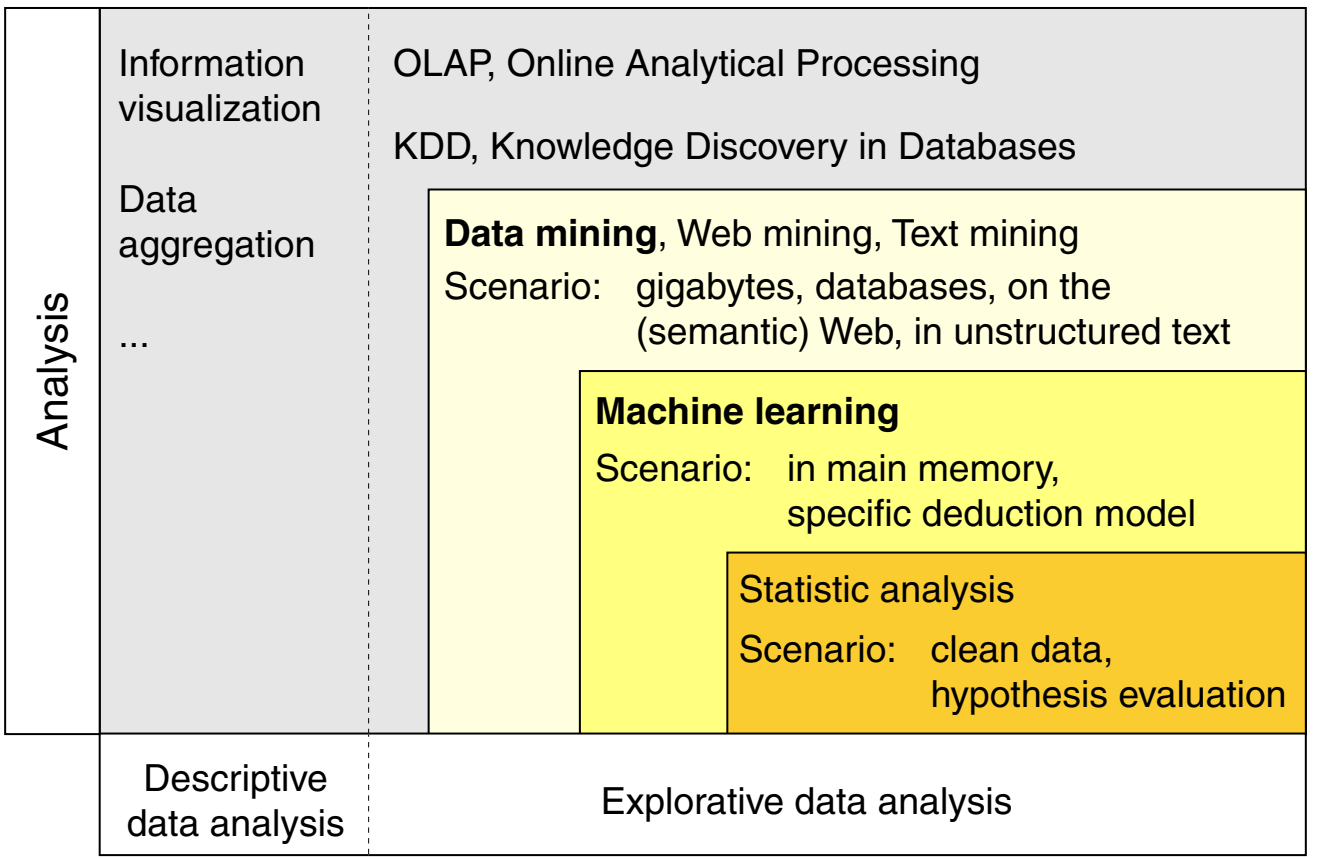
Delineation

Machine Learning, Data Mining



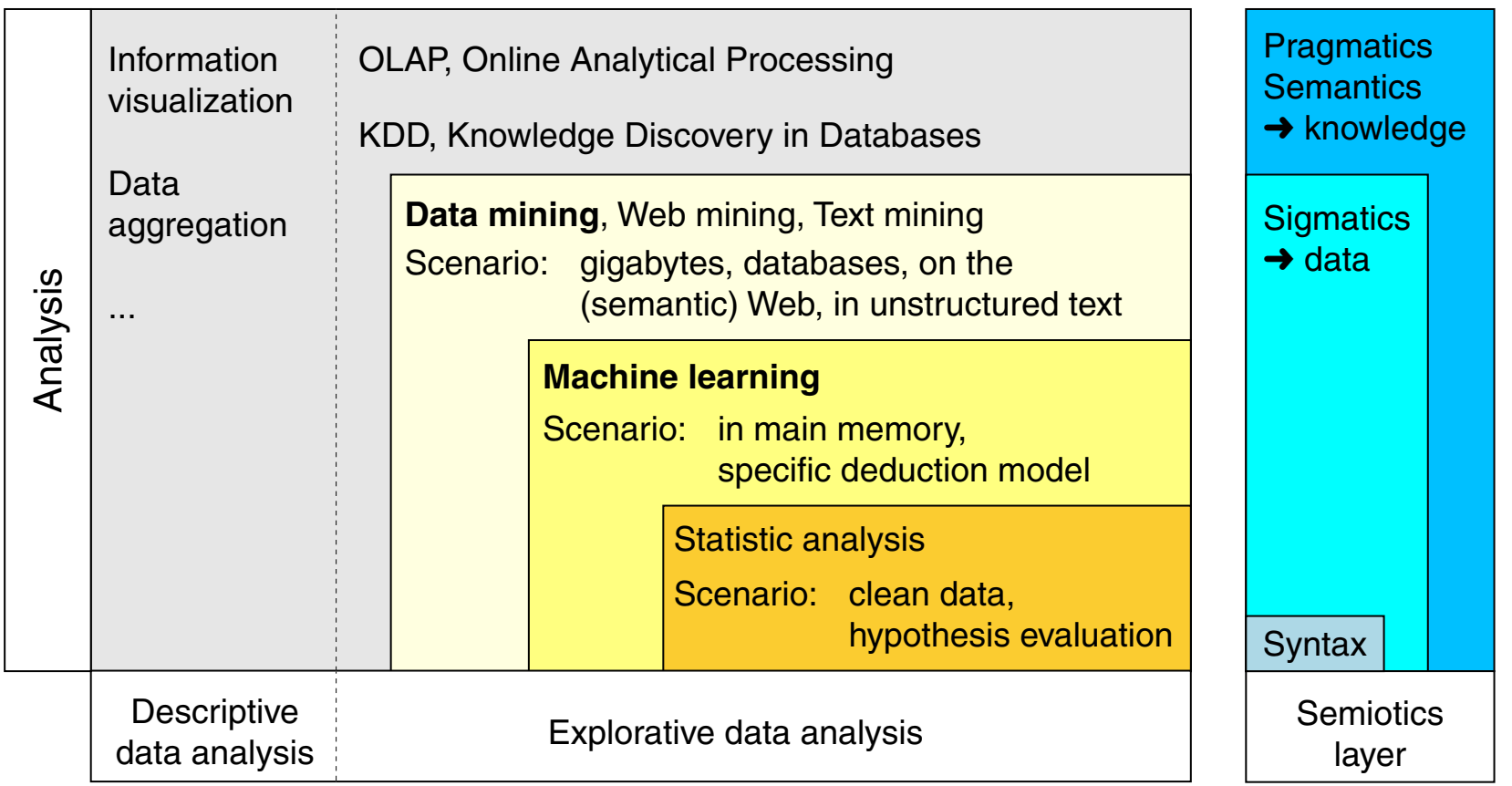
Delineation

Machine Learning, Data Mining



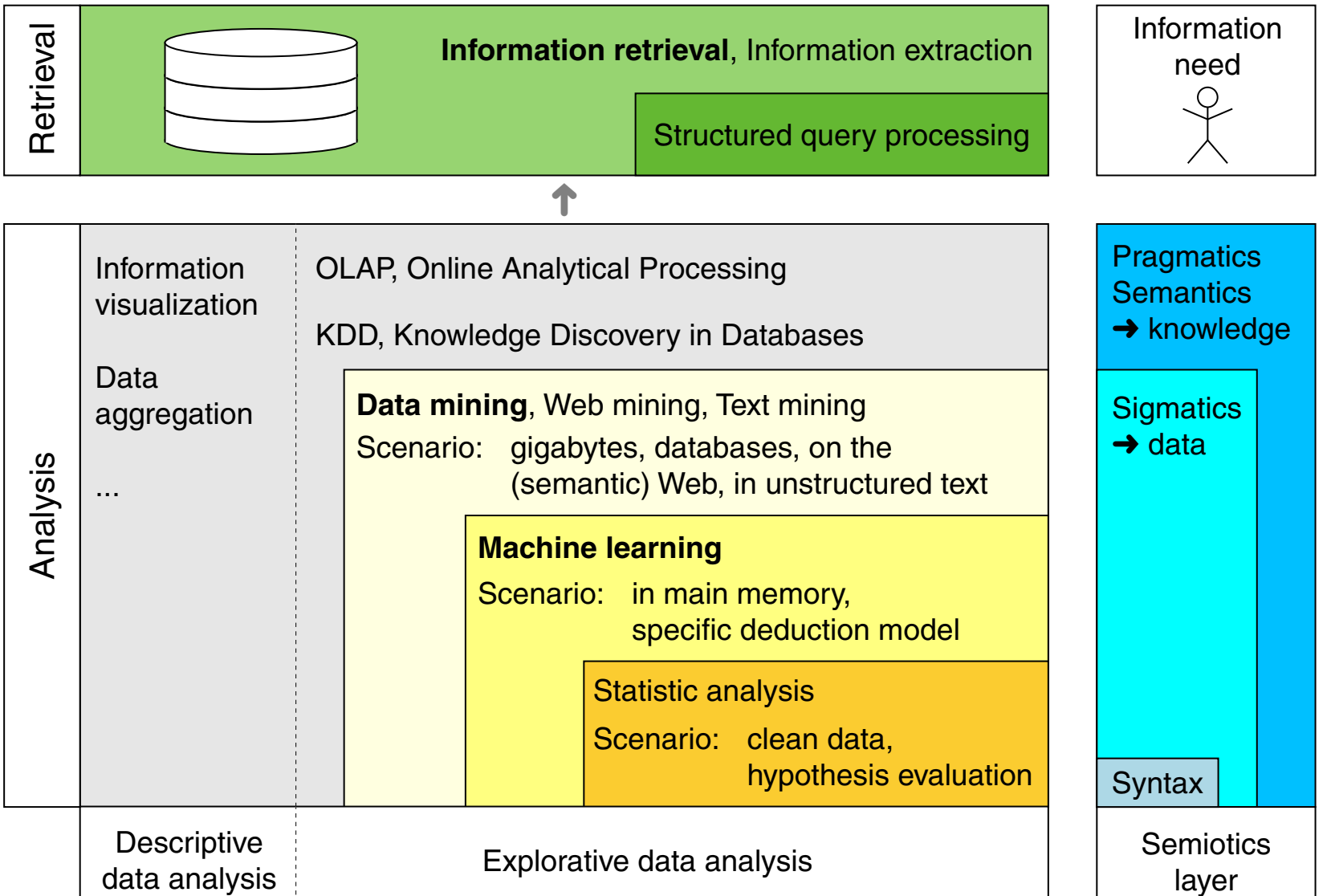
Delineation

Machine Learning, Data Mining



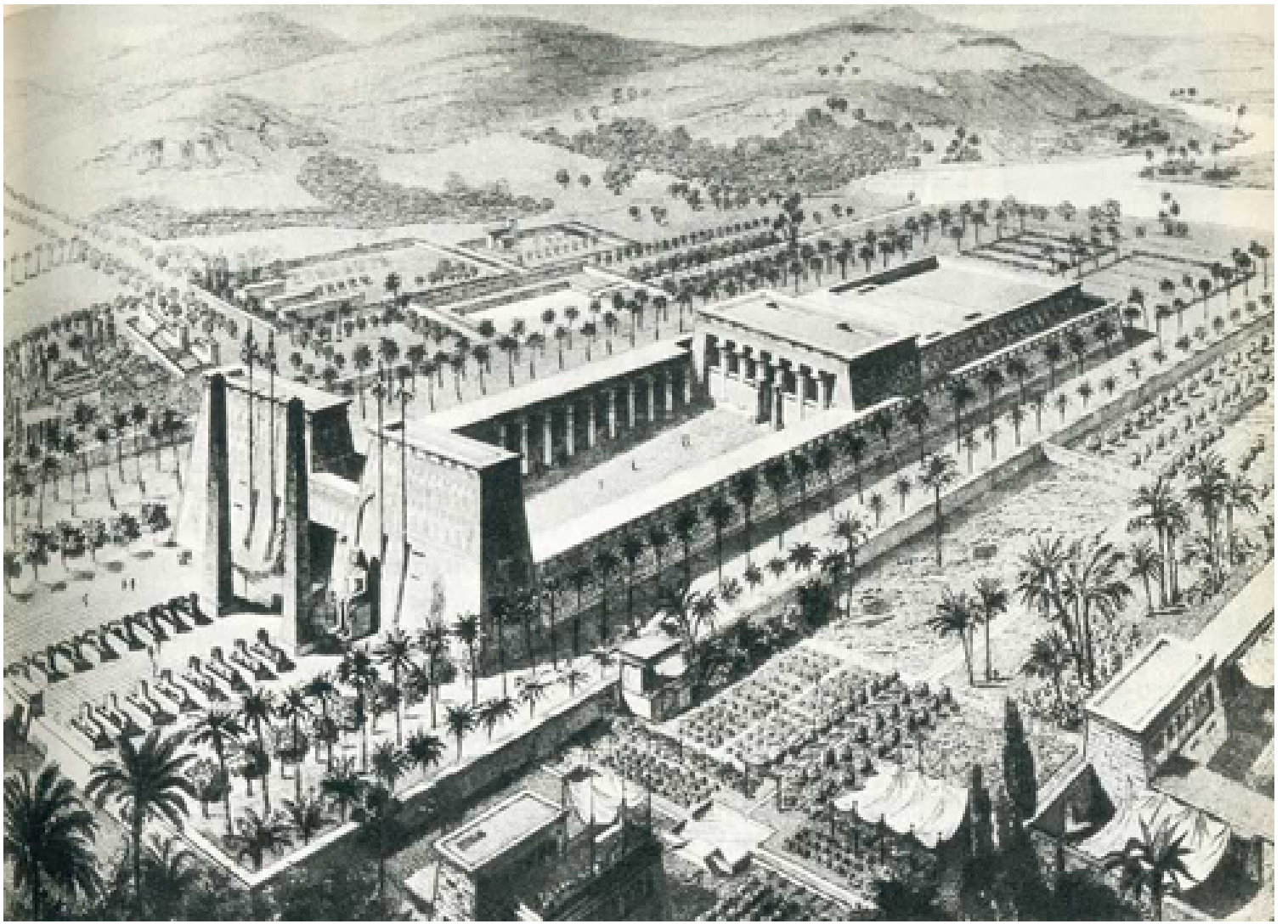
Delineation

Machine Learning, Data Mining



Historical Background

Manual Retrieval

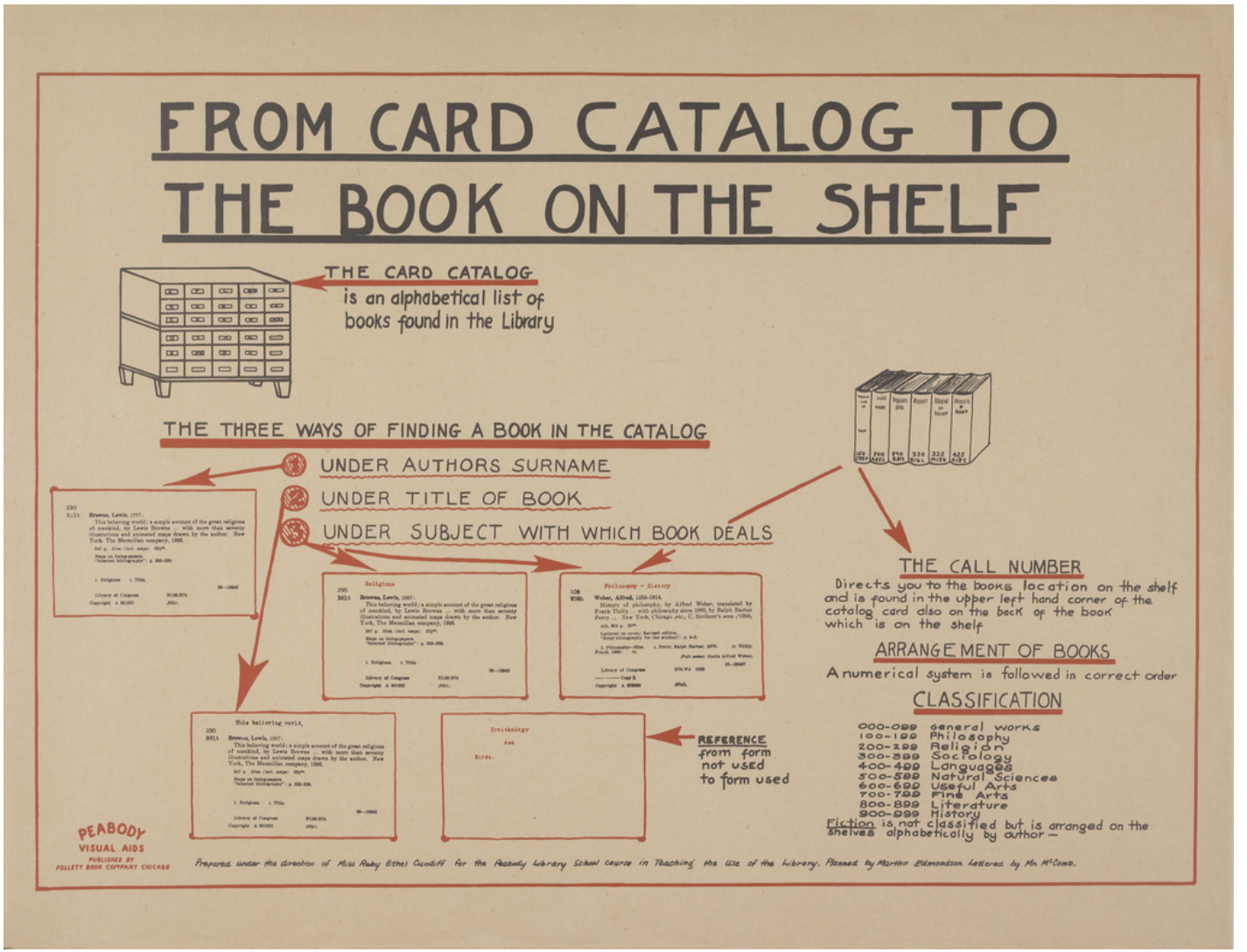


Remarks:

- The Ancient Library of Alexandria, Egypt, was one of the largest and most significant libraries of the ancient world. It flourished under the patronage of the Ptolemaic dynasty and functioned as a major center of scholarship from its construction in the 3rd century BC until the Roman conquest of Egypt in 30 BC. The library was part of a larger research institution called the Musaeum of Alexandria, where many of the most famous thinkers of the ancient world studied. [Wikipedia]
- These include Archimedes, father of engineering; Aristarchus of Samos, who first proposed the heliocentric system of the universe; Callimachus, a noted poet, critic and scholar; Eratosthenes, who argued for a spherical earth and calculated its circumference to near-accuracy; Euclid, father of geometry; Herophilus, founder of the scientific method; Hipparchus, founder of trigonometry; Hero, father of mechanics. [Wikipedia]
- Callimachus' most famous prose work is the Pinakes (*Lists*), a bibliographical survey of authors of the works held in the Library of Alexandria. The Pinakes was one of the first known documents that lists, identifies, and categorizes a library's holdings. By consulting the Pinakes, a library patron could find out if the library contained a work by a particular author, how it was categorized, and where it might be found. Callimachus did not seem to have any models for his pinakes, and invented this system on his own. [Wikipedia]
- The Library held between 400,000 and 700,000 scrolls, grouped together by subject matter. Within the Pinakes, Callimachus listed works alphabetically by author and genre. He did what modern librarians would call adding metadata—writing a short biographical note on each author, which prefaced that author's entry. In addition, Callimachus noted the first words of each work, and its total number of lines. [Phillips 2010]

Historical Background

Manual Retrieval



Historical Background

Manual Retrieval

Remarks:

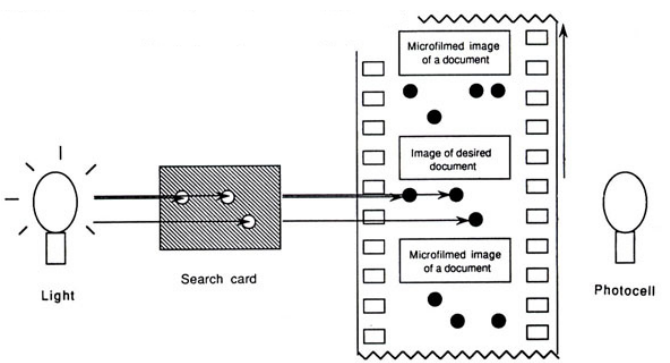
- Today, WorldCat is a union catalog that itemizes the collections of 72,000 libraries in 170 countries and territories. It contains more than 420 million records, representing over 2.6 billion physical and digital assets in 491 languages, as of August 2018. [OCLC]
- What are problems when sorting by author?
- What is necessary to organize library cards by subject?
- Librarians can find books by author, by title, and by subject. What is still missing?

Historical Background

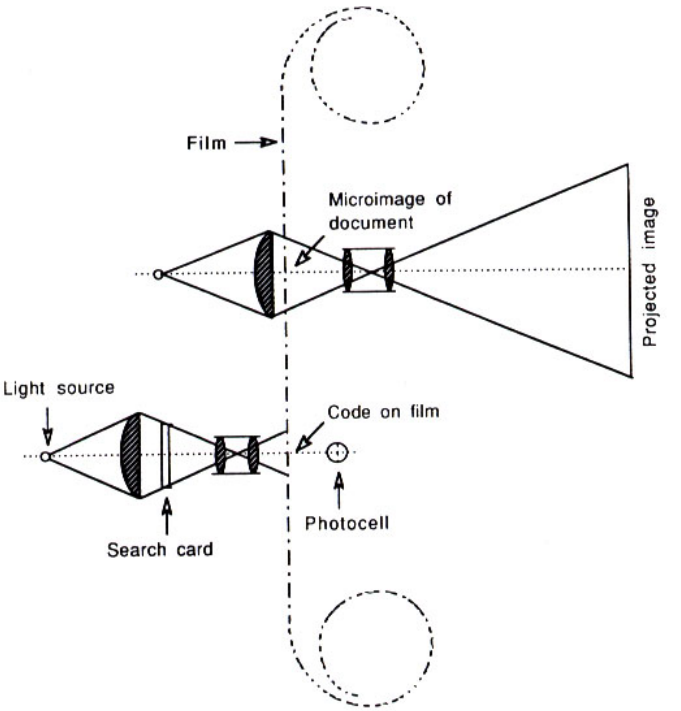
Mechanical Retrieval

Emanuel Goldberg's Statistical Machine [Buckland 1995]:

- Documents on microfilm with associated patterns of holes
- Punch cards as search patterns
- US patent No. 1,838,389, applied 1927, issued 1931



Searching

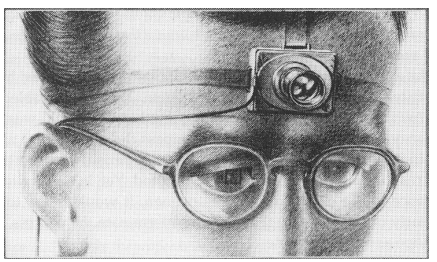


Result presentation

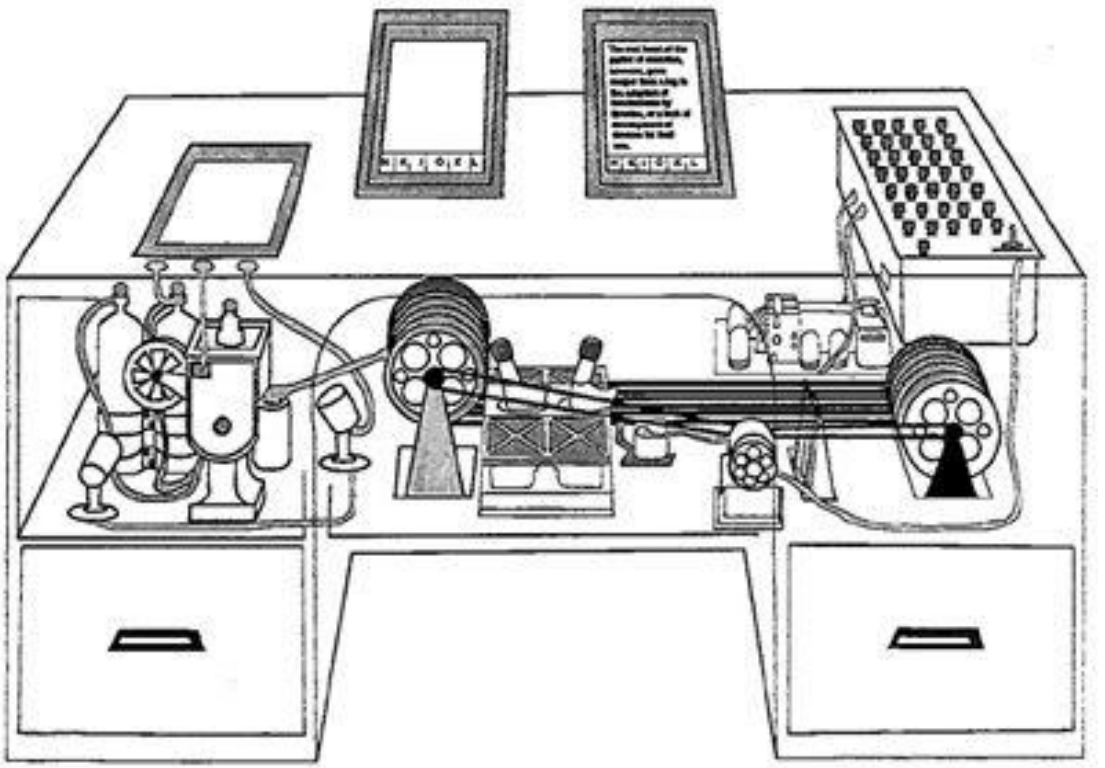
Historical Background

Mechanical Retrieval

Vannevar Bush's Memex [Bush 1945]:



Recording via camera
(early life logging)



Retrieval, Commenting, Browsing, Cross-referencing

Historical Background

Computerized Retrieval

First reference to computer-based search [Holmstrom 1948]:

> Then there is also in America a **machine called the Univac** which has a typewriter keyboard connected to a device whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape.
>
> By this means the **text of a document**, preceded by its subject code symbol, **can be recorded** on the tape by any typist.
>
> For **searching**, the tape is run through the machine which thereupon automatically selects and types out those references which have been coded in any desired way **at a rate of 120 words a minute**--complete with small and capital letters, spacing, paragraphing, indentations and so on.
>
> (If the tape is run through the other way, it obediently types out the text backwards at the same rate!)

Historical Background

Computerized Retrieval

First use of the term “information retrieval” [Mooers 1950]:

> The problem under discussion here is machine searching and retrieval of information from storage according to specification by subject. An example is the library problem of selection of technical abstracts from a listing of such abstracts. 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.

Remarks:

- Serious research into information retrieval started after World War II ended, when scientists of the allied forces turned their attention away from warfare, realizing that the vast quantities of scientific results and other information accumulated throughout the war was too much to make sense of for any individual scientist.
- [Bagley 1951] observed in his Master thesis that “recently published statistics relating to chemical publication show that a search of Chemical Abstracts would have been complete in 1920 after considering twelve volumes containing some **184,000 abstracts**. But in 1935 there would have been fifteen more volumes to search, and these new volumes alone contain about **382,000 abstracts**. By the end of 1950 the forty-four volumes of Chemical Abstracts to be searched contained well **over a million abstracts**. If the present trend in publication continues, the total abstracts published in this one field by 1960 will be almost **1,800,000**.”
- [Sanderson and Croft 2012] compiled a brief history of information retrieval research.

Historical Background

Information Retrieval (1950s)

Indexing and ranked retrieval:

- Coordinate Indexing

  Mortimer Taube proposes “Coordinate Indexing” of documents based on a selection of independent “uniterms,” called (index) terms or keywords today, departing from traditional subject categorization schemes. Assigning uniterms to documents is called indexing. Adding a reference to a document to the specific catalog cards for its uniterms is called posting. Retrieval works by looking up a set of uniterms of interest, collecting documents to which at least a subset of them has been assigned.

- Cranfield paradigm

  Cyril Cleverdon starts the Cranfield projects, introducing lab evaluation of indexing and retrieval based on (1) a document collection, (2) a set of queries, and (3) relevance judgments for pairs of queries and documents, later known as the Cranfield paradigm of IR evaluation.

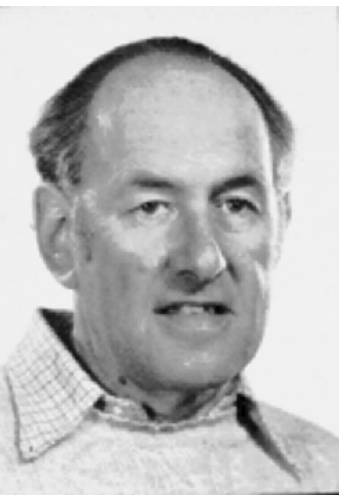
- Term frequency-based ranking

  Hans Peter Luhn proposes to score and rank documents based on their relevance to a query. He suggests term frequency of terms in a document as an approximate measure of term importance during scoring.

Historical Background

Information Retrieval (1960s)

Gerard Salton:



- Eminent IR researcher: “father of Information Retrieval”
- Many seminal works

  Invention of / key contributions to automatic indexing, full-text indexing (i.e., using all words of a document as index terms), term weighting, relevance feedback, document clustering, dictionary construction, term dependency, phrase indexing, semantic indexing via thesauri, passage retrieval, summarization, . . .

- Cosine similarity

  The Vector Space Model, proposed by Paul Switzer, represents documents and queries in high-dimensional space. Salton suggests to measure the similarity between query and document vectors via the cosine of the angle between them, the cosine similarity.

- Integration of the state of the art into the SMART retrieval system.
- First laureate of the Gerard Salton Award in 1983, named in his honor.

Remarks:

- A funny side note; as per [Salton 1968], “information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”
  What is wrong with this definition?
- Interestingly, commercial applications that emerged around this time largely ignored the insights gained from IR research. Not even ranked retrieval was adopted, but basic Boolean retrieval models were employed. This situation did not change until the mid-1990s and even today Boolean search is still very important in some domains like patent retrieval or prior art search, systematic reviews, etc.

Historical Background

Information Retrieval (1970s)

tf · *idf*-weighted Vector Space Model:

- Inverse document frequency

  Karen Spärck Jones proposes the Inverse Document Frequency to measure term importance within document collections, complementing Luhn's term frequency to form the well-known *tf* · *idf* term weighting scheme.

- Vector space model

  Supposed formalization of "A Vector Space Model for Information Retrieval" by Salton, Wong, and Yang; this attribution has been debunked [Dubin 2004].

Probabilistic retrieval:

- Probability ranking principle

  Stephen Robertson formalizes the probability ranking principle, stating that "documents should be ranked in such a way that the probability of the user being satisfied by any given rank position is a maximum."

- C.J. "Keith" van Rijsbergen proposes to incorporate term dependency into probabilistic retrieval models.

Historical Background

Information Retrieval (1980s - mid-1990s)

- BM25

  Stephen Robertson et al. introduce BM25 (Best Match 25) as an alternative to $tf \cdot idf$.

- Latent semantic indexing

  Scott Deerwester et al. propose to embed document and query representations in low-dimensional space using singular value decomposition of the term-document matrix.

- Stemming

  Introduction of Porter's stemming algorithm into the indexing pipeline to conflate words sharing the same stem.

- TREC-style evaluation: shared tasks

  Ellen Vorhees and Donna Harman organize the first Text REtrieval Conference (TREC), focusing on large-scale IR systems evaluation under the Cranfield paradigm, repeating it annually to this day.

- Learning to rank

  Norbert Fuhr describes the foundations of learning to rank, the application of machine learning to ranked retrieval, where relevance is learned from training samples of pairs of queries and (ir)relevant documents.

Historical Background

Information Retrieval (mid-1990s - 2000s)

Web search:

- Web crawlers are developed for the rapidly growing web.
- PageRank and HITS

  Spam pages increasingly pollute search results. Sergey Brin and Larry Page propose PageRank to identify authoritative web pages based on link structure, laying the foundation of Google. In parallel, John M. Kleinberg proposes HITS.
- Query log analysis

  Thorsten Joachims renders learning to rank feasible, exploiting clickthrough data for training. Others develop query suggestion, spell correction, query expansion, etc. based on logs.
- Anchor text indexing

  Oliver A. McBryan proposes the use of anchor text indexing to gain additional information about a web page, and to undo spam.
- Maximum marginal relevance for diversity

  Jaime Carbonell and Jade Goldstein propose maximum marginal relevance (MMR) to allow for search result diversity.
- Language modeling for IR

  Jay M. Ponte and W. Bruce Croft first apply language modeling to IR.

It’s been a long way

Web Search

DuckDuckGo

bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

Web Search

DuckDuckGo

bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

Web Search