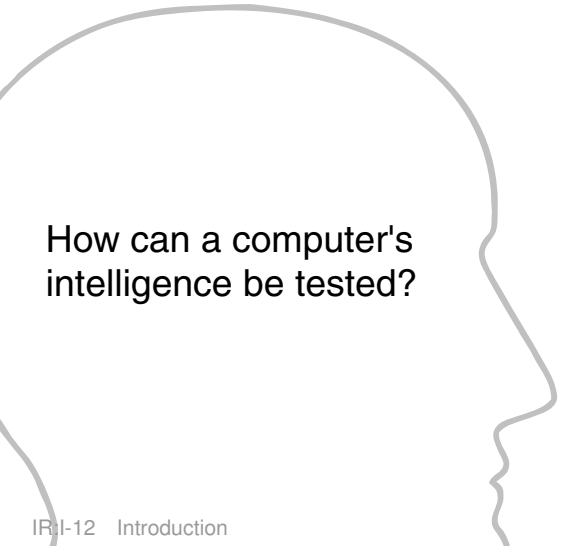
Chapter IR:I

I. Introduction

- Information Retrieval in a Nutshell
- Examples of Retrieval Problems
- Terminology
- Delineation
- Historical Background
- Architecture of a Search Engine

Information Retrieval in a Nutshell

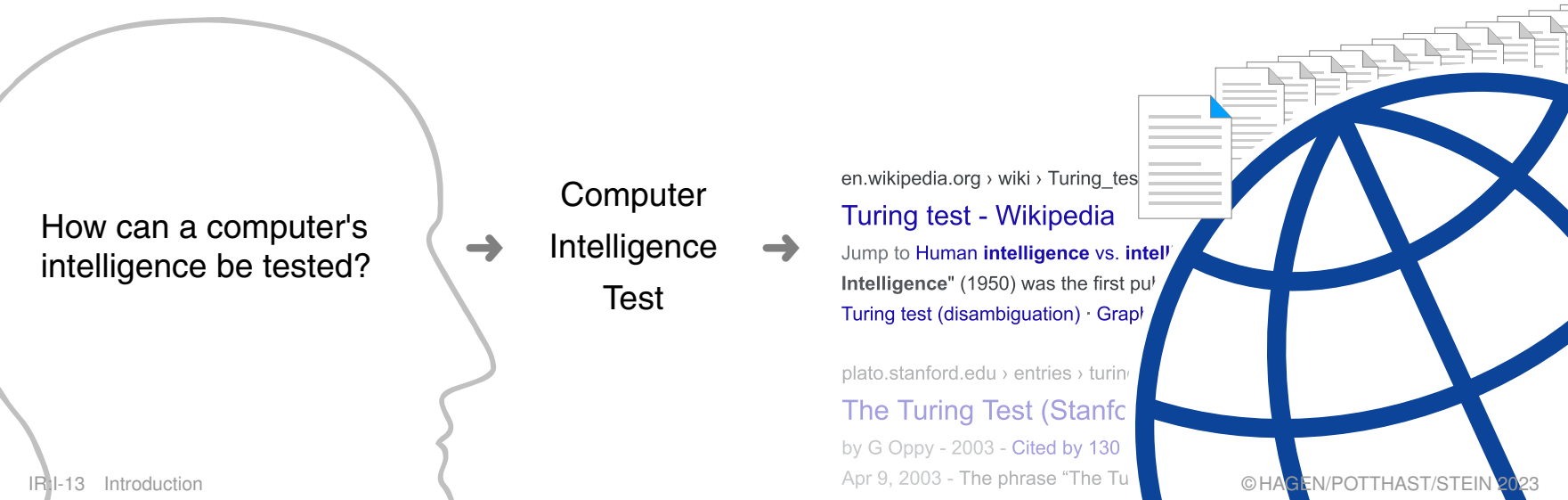
- A vague request.
  Expression of a complex information need: a question
- Billions of documents.
  Text, images, audio files, videos, . . .



Information Retrieval in a Nutshell

- A vague request.
  Expression of a complex information need: a question, or just a few keywords.
- Billions of documents.
  Text, images, audio files, videos, . . .
- High class imbalance.
  Only a tiny fraction of all documents are relevant to the request.

➜ Retrieve relevant documents in milliseconds.



Chapter IR:I

I. Introduction

- Information Retrieval in a Nutshell
- Examples of Retrieval Problems
- Terminology
- Delineation
- Historical Background
- Architecture of a Search Engine

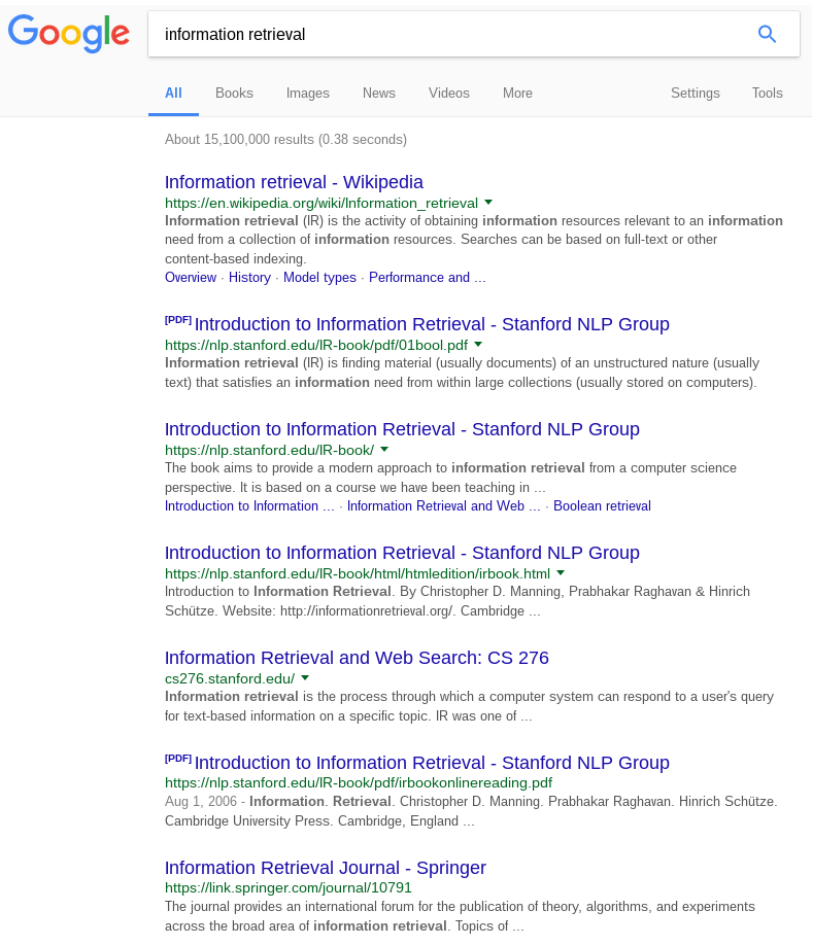
Examples of Retrieval Problems

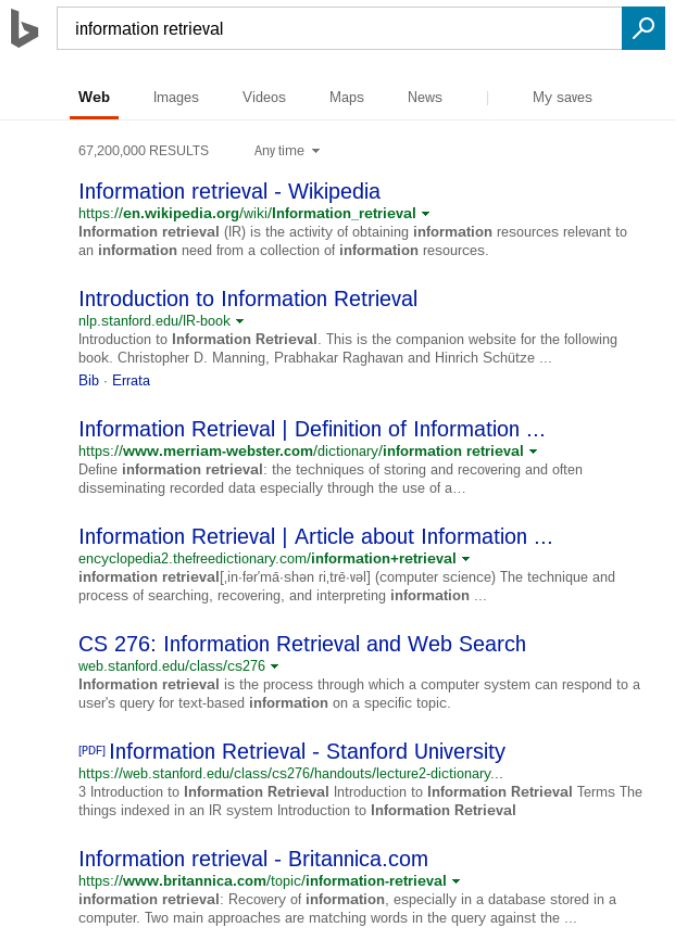
Learn everything there is to learn about information retrieval.

Examples of Retrieval Problems

Learn everything there is to learn about information retrieval.

Search for texts containing ‘information’ and ‘retrieval’.

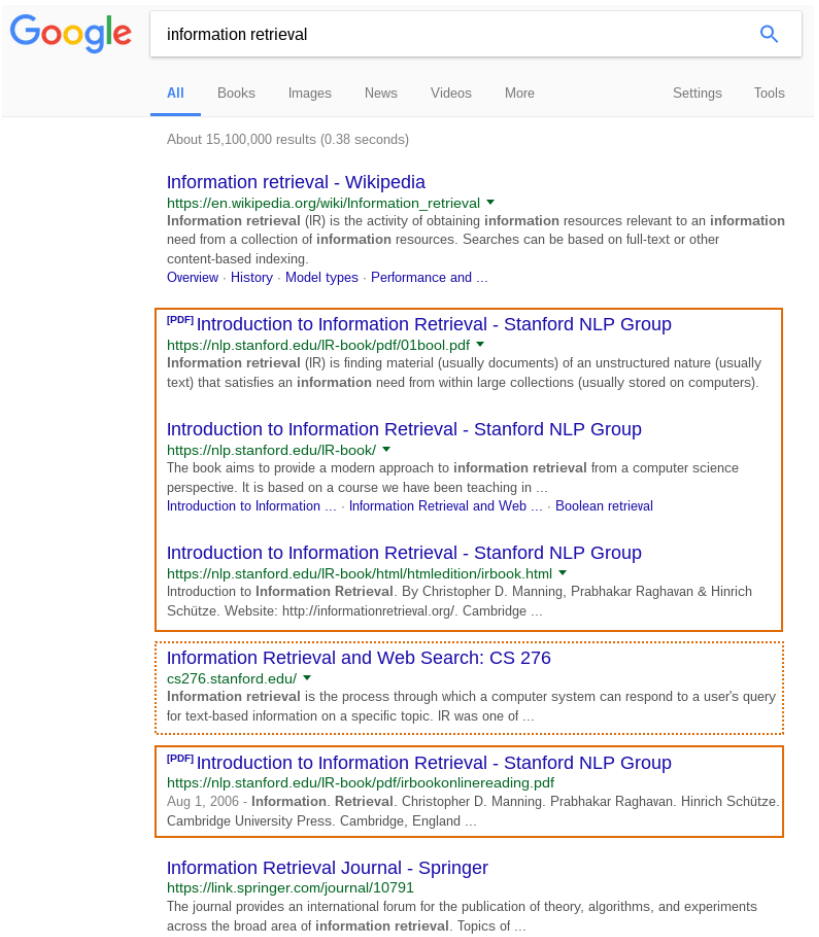


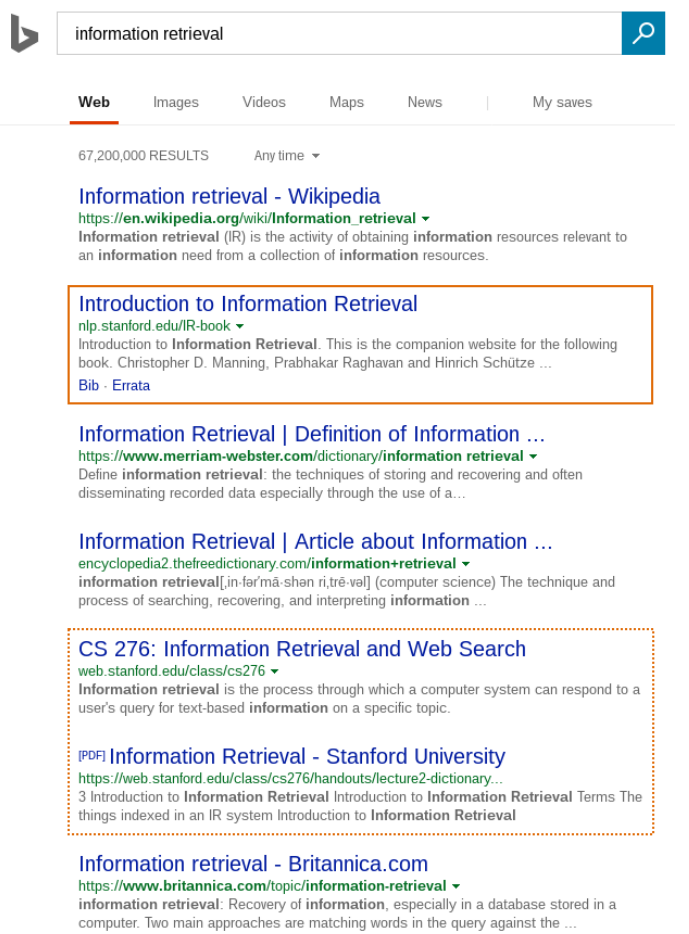


Examples of Retrieval Problems

Learn everything there is to learn about information retrieval.

Search for texts containing ‘information’ and ‘retrieval’.





Remarks:

- Here the search engine is treated like a database of documents. It searches for those documents that contain certain words the user expects to find in a relevant document. Unlike a database, the search engine ranks the retrieved documents according to its estimation of how useful they are to the user.
- Compare the different total numbers of search results. The discrepancy may be due to the difference in the number of documents indexed, suggesting that Bing indexes many more documents than Google. However, for competitive reasons, search engines have long ago stopped disclosing the number of documents they index, and these numbers are only estimates based on a partial search. As a rule, these estimates overestimate the actual number of documents that can be retrieved.
- 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 included in all retrieved documents in that order, greatly reducing the number of results. It is reasonable to assume that all documents dealing with information retrieval contain this phrase at least once, whereas documents dealing with something else do not. Interestingly, only about 1.12% of all search queries contain this or other search operators [White and Morris, 2007].
- Google’s results 2, 3, 4, and 6 point to the same book’s website; the 5th result points to a lecture based on the book at the same organization. Bing’s results 3, 4, and 7 are dictionary pages. What is wrong with that?
- The snippet of Bing’s 6th result is flawed.

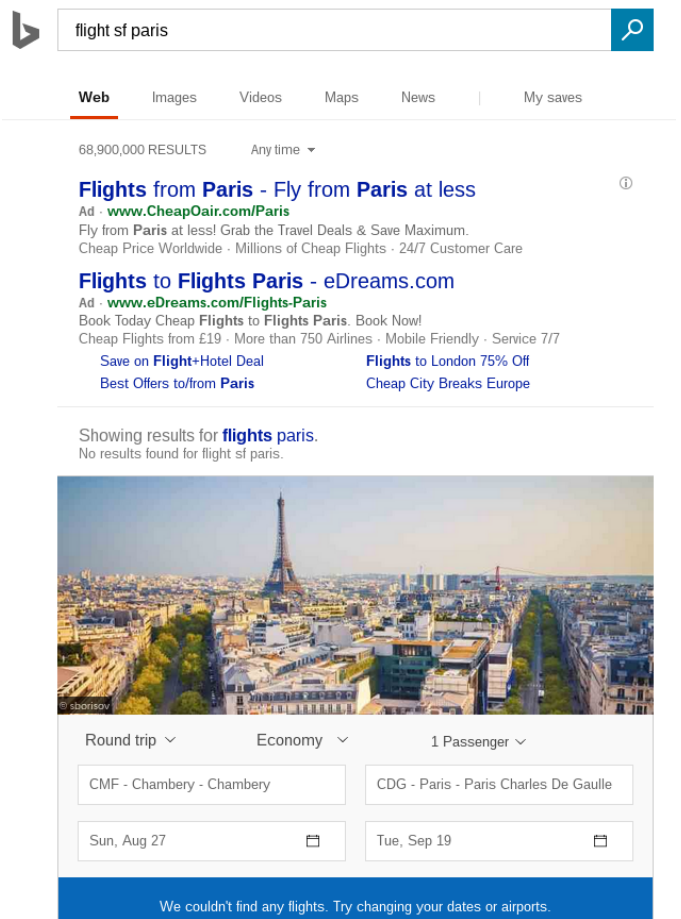
Examples of Retrieval Problems

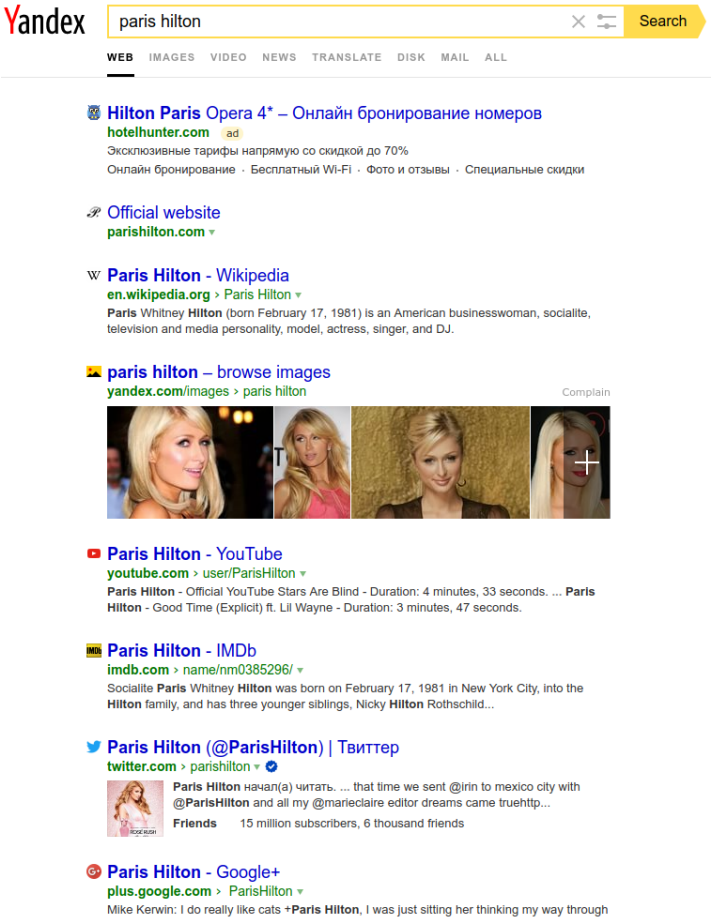
Plan a trip from San Francisco to Paris, France.

Examples of Retrieval Problems

Plan a trip from San Francisco to Paris, France.

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





Remarks:

- Users use casual language to describe their information needs.
- Bing does not understand ‘sf’ as an abbreviation for San Francisco. ‘SFO’, the location identifier for San Francisco airport, would have worked. Google at least offers a spelling correction for ‘sf’ to ‘sfo’.
- Users use ambiguous search queries whose semantics depend on context. For example, when searching for the Hilton hotel in Paris, Yandex returns results about the celebrity Paris Hilton. Only the ad on top of the search results gives an indication of the intended semantics. Searching for “hilton paris” returns better results.
- Search engines allow you to solve problems directly on the search results page (so-called oneboxes). The flight search box in Bing’s results is one example.
- Search engines adopt the paradigm of “universal search” and offer different types of results. The images in Yandex’ results are an example of this.

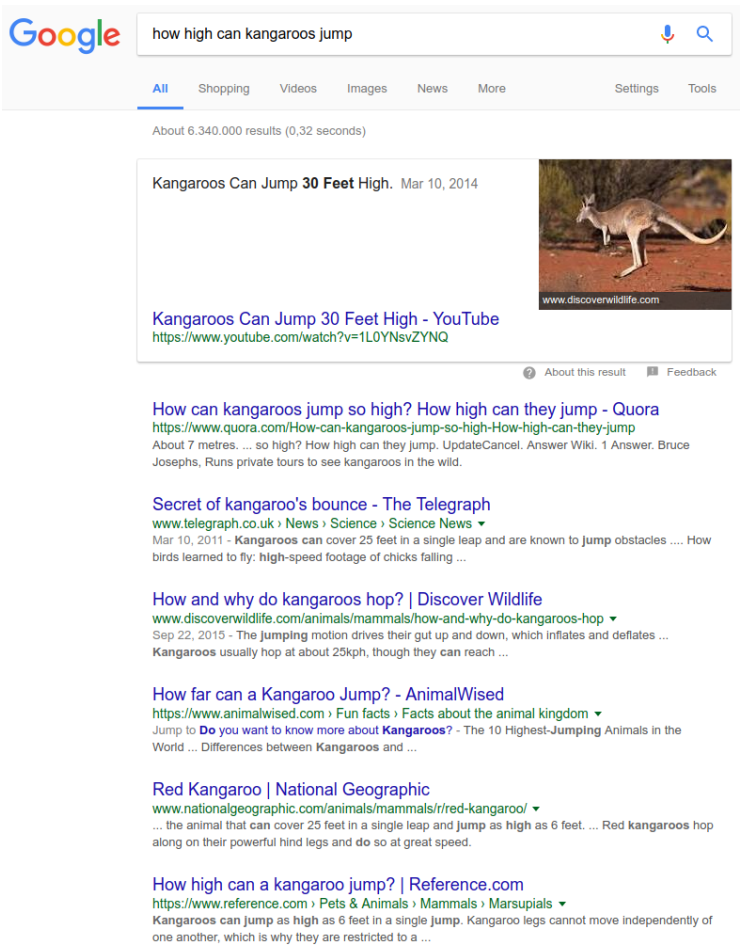
Examples of Retrieval Problems

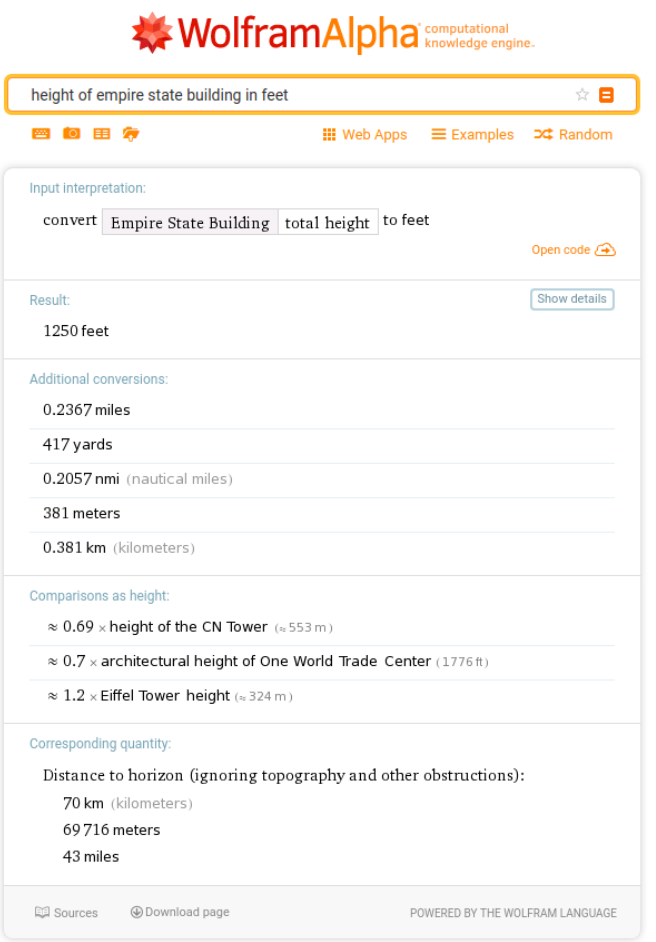
Answer “Can Kangaroos jump higher than the Empire State Building?”

Examples of Retrieval Problems

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

Search for facts

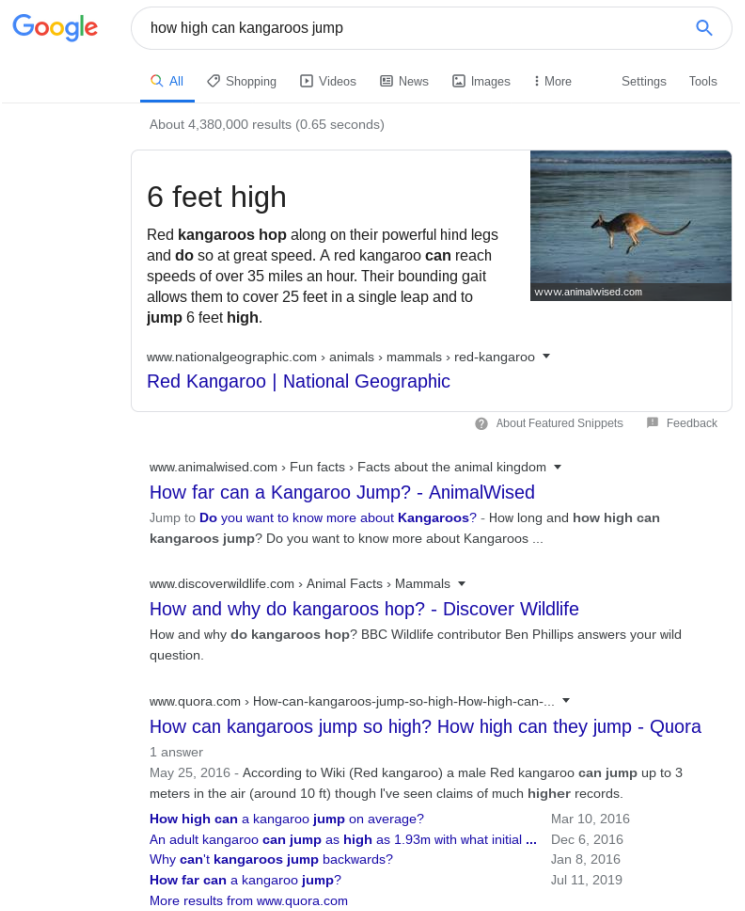


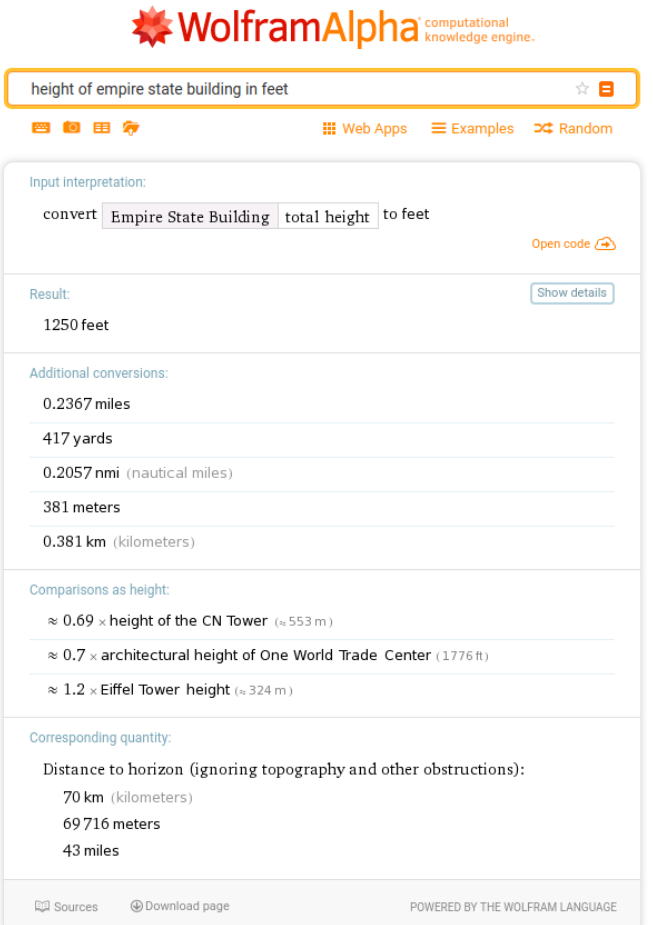


Examples of Retrieval Problems

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

Search for facts

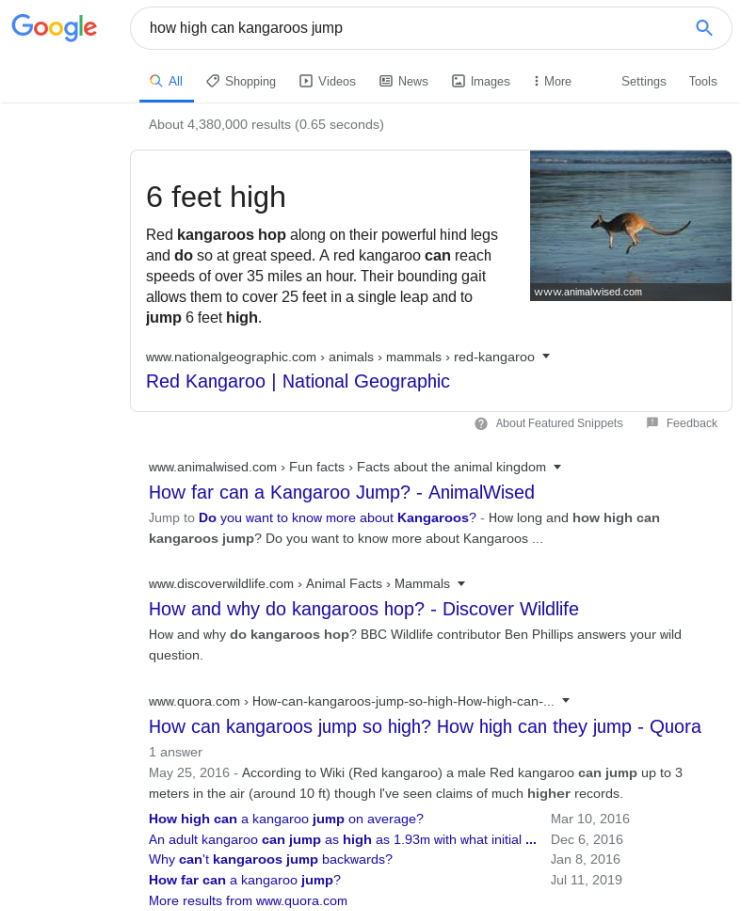


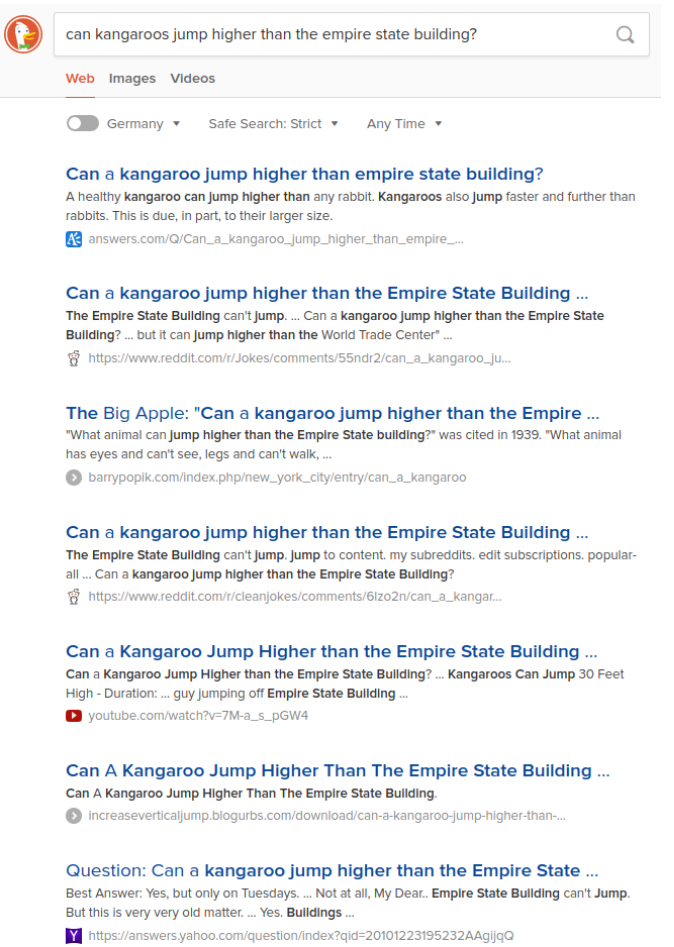


Examples of Retrieval Problems

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

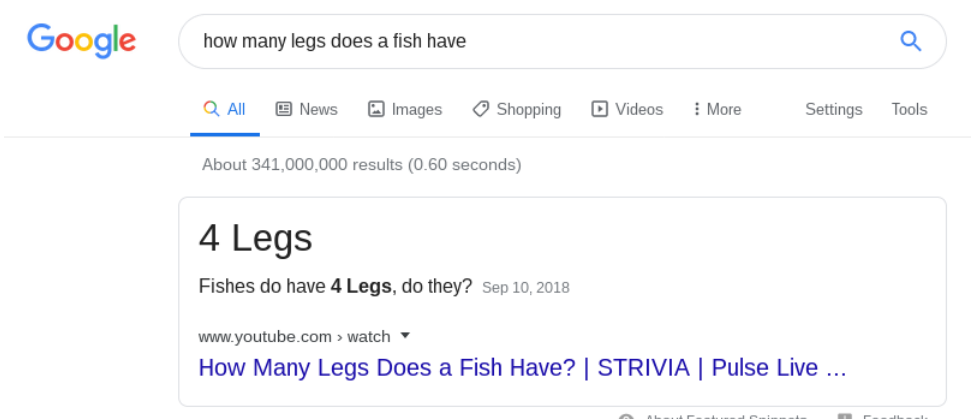
Search for facts, or ask the question outright.





Remarks:

- Users search for facts. Search engines use different strategies to fulfill such queries, using knowledge bases like Wikidata or extracting facts from web pages.
- The highlighted top search result from Google seems to answer the question. However, the height given is wrong. Height is confused with distance. Google does not check the truth of a statement, but outputs the results that best match the query. In other snippets, distances are given that do not match the top one. In the bottom two snippets, it is claimed that kangaroos can only jump about 6 feet high. YouTube videos often are an unreliable source.
- WolframAlpha allows asking questions that require computations. It draws on knowledge bases to supplement required facts.
- Asking the original question directly reveals that it is a well-known joke question. Some snippets from DuckDuckGo reveal the answer.
- Search engines lack common sense: [@webis_de]



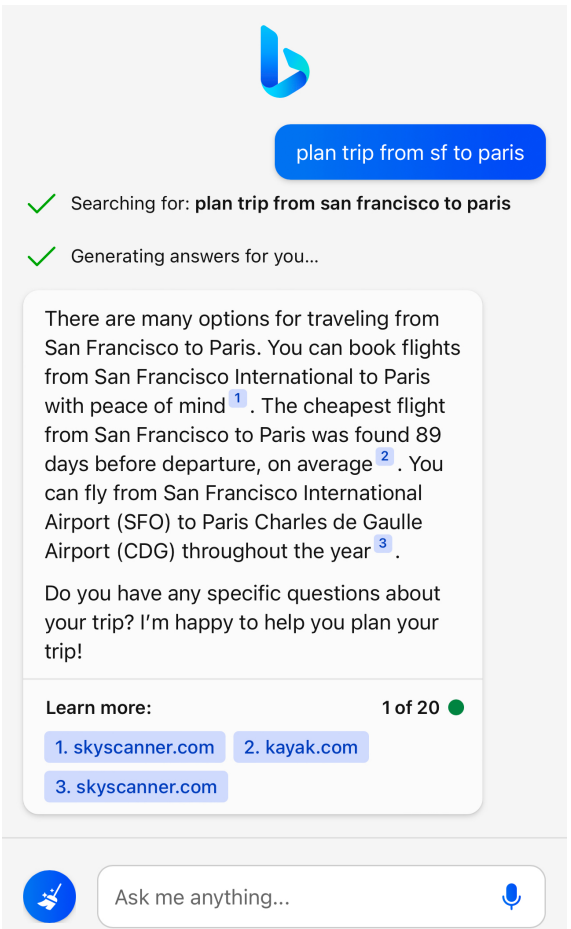
Examples of Retrieval Problems

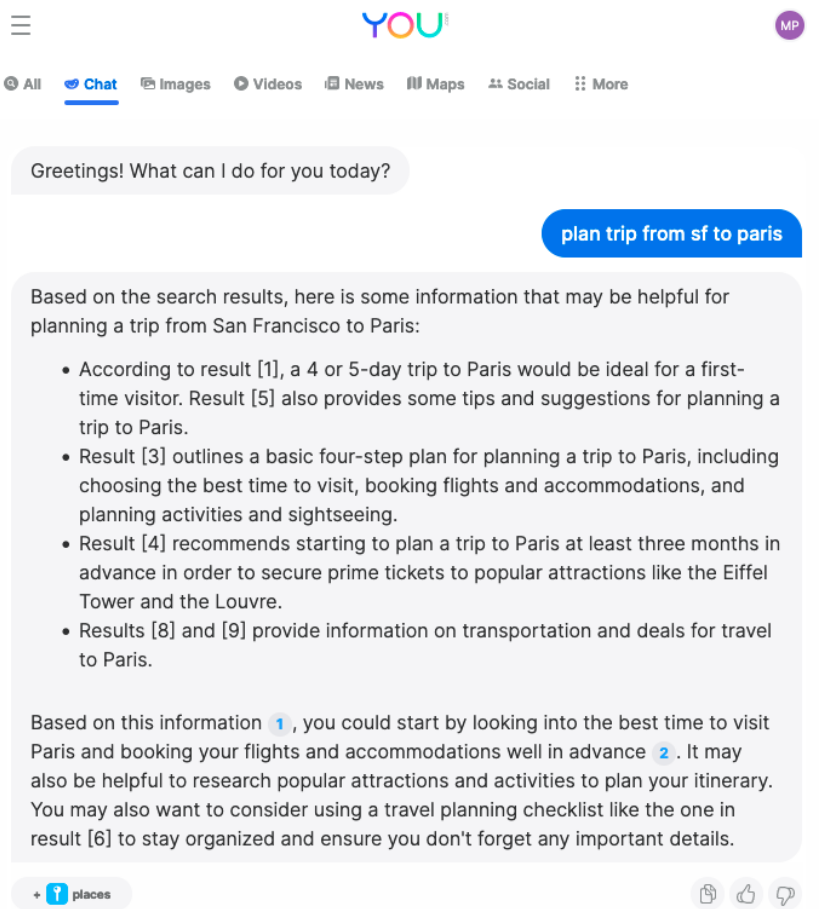
Do I really need to read so many search results to solve these problems?

Examples of Retrieval Problems

Do I really need to read so many search results to solve these problems?

Do a conversational search with an AI assistant.

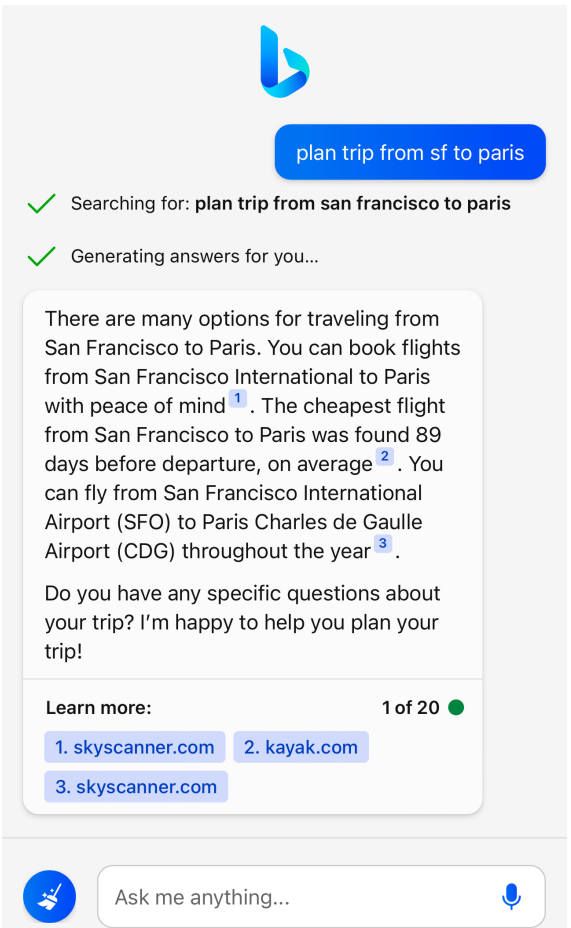


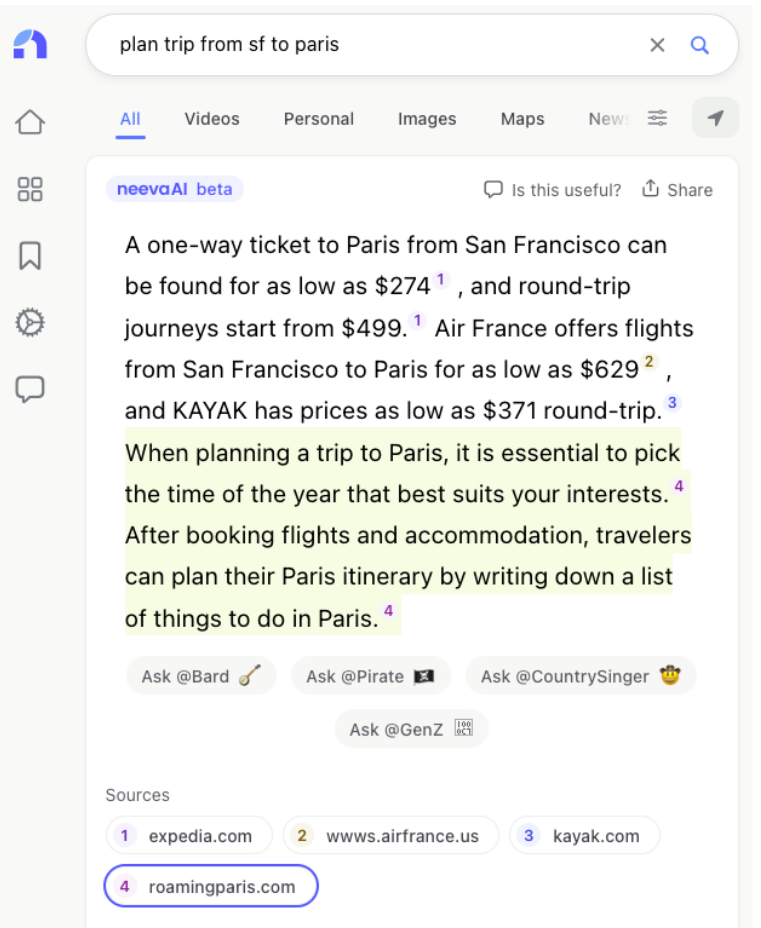


Examples of Retrieval Problems

Do I really need to read so many search results to solve these problems?

Do a conversational search with an AI assistant.





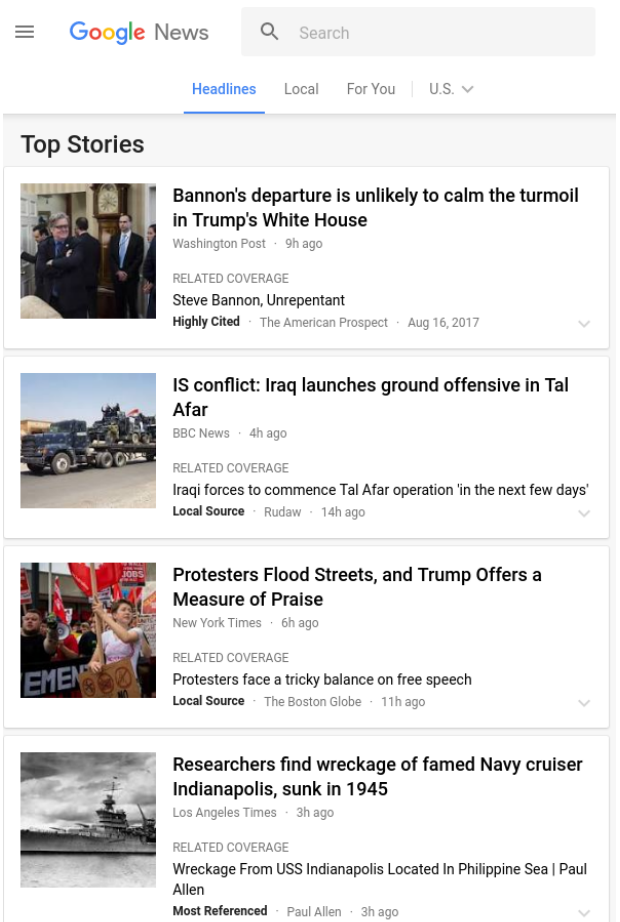
Examples of Retrieval Problems

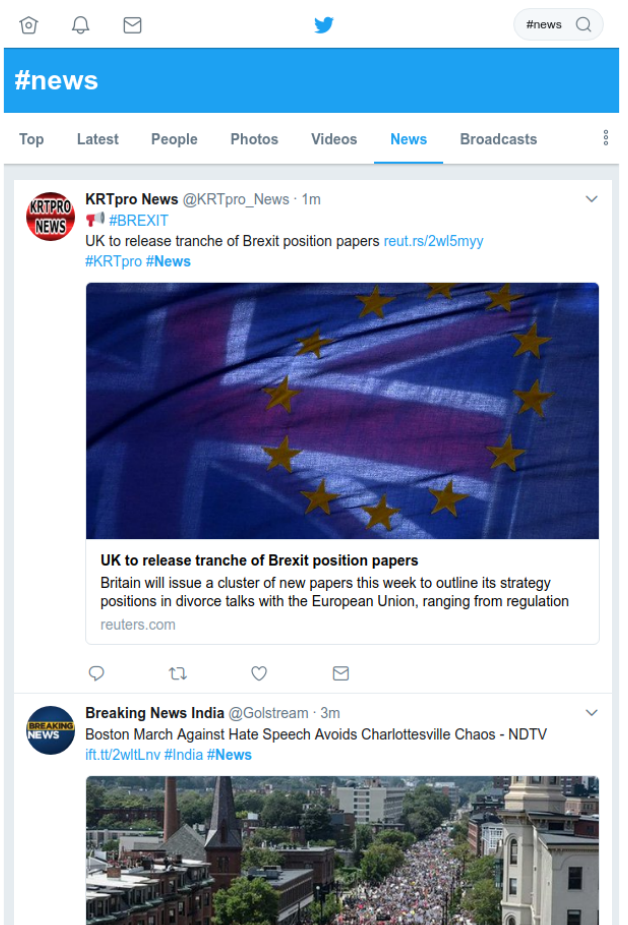
What were the news today?

Examples of Retrieval Problems

What were the news today?

Check out the news feeds.

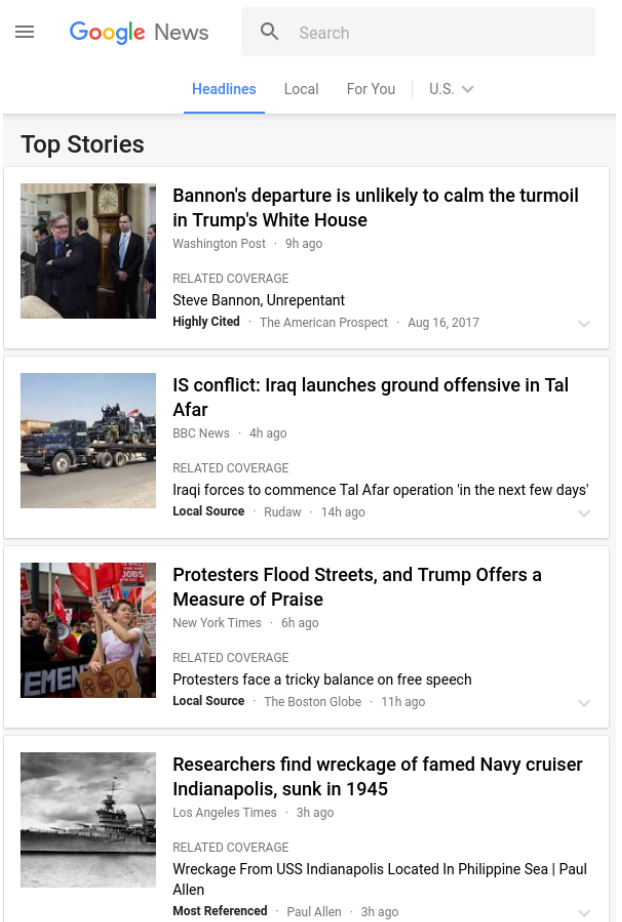




Examples of Retrieval Problems

What were the news today?

Check out the news feeds.



Remarks:

- One cannot search for things one does not know. Instead of searching for news, they are explored. The information systems for this purpose are news aggregators and social networks, but also the front pages of newspapers. The former recommend news based on user preferences.
- Since 2017, Google News no longer displays preview snippets of news articles, but only the headlines. While this change is justified by better usability, it coincides with increasing pressure from news publishers and the legislation passed on ancillary copyright in various countries.
- Google News shows brief explanation labels indicating the relevance of a new item: “Highly Cited”, “Local Source”, “Most Referenced”, etc.
- Facebook’s role in providing Americans with political news has never been stronger—or more controversial. Scientists worry that the social network can create “echo chambers” where users see posts only from like-minded friends and media sources.

  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]

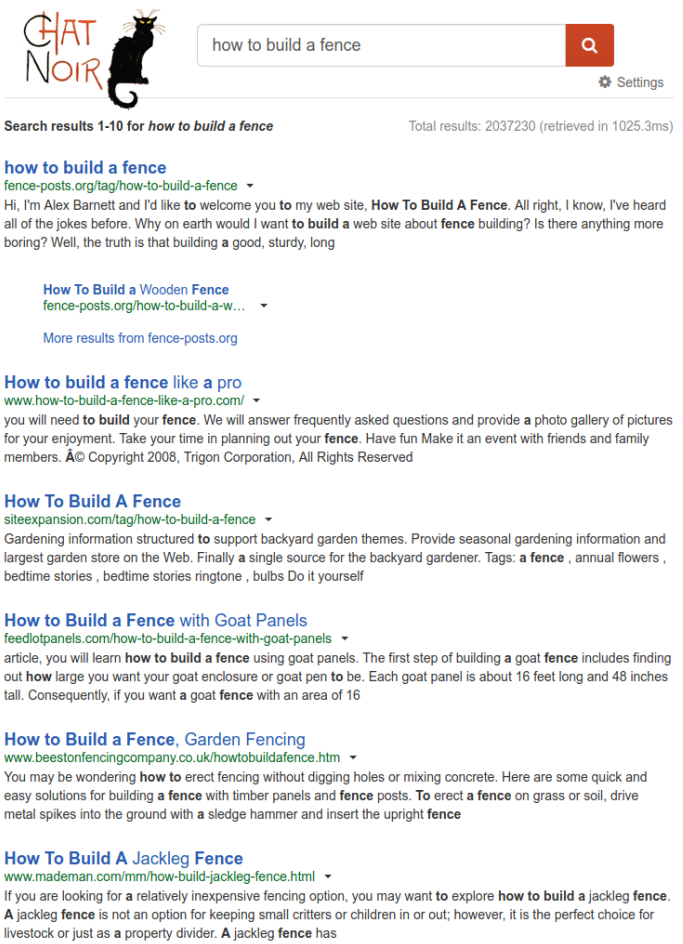
Examples of Retrieval Problems

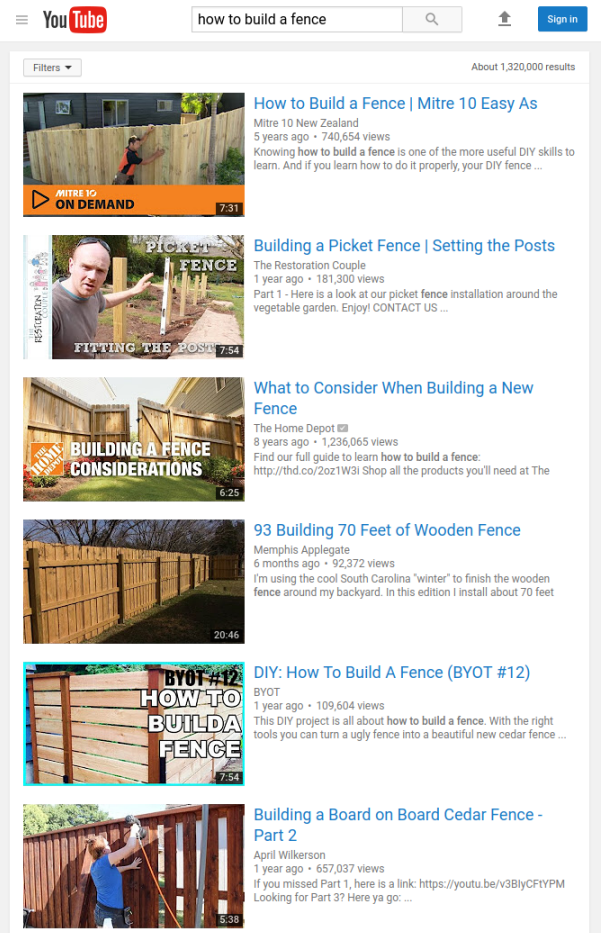
Build a fence.

Examples of Retrieval Problems

Build a fence.

Search for tutorials.





Remarks:

- Users search for instructions for complex tasks. In addition to textual information, this also includes instructive multimedia content, for example from YouTube.
- ChatNoir is the only publicly available research search engine that operates at scale.

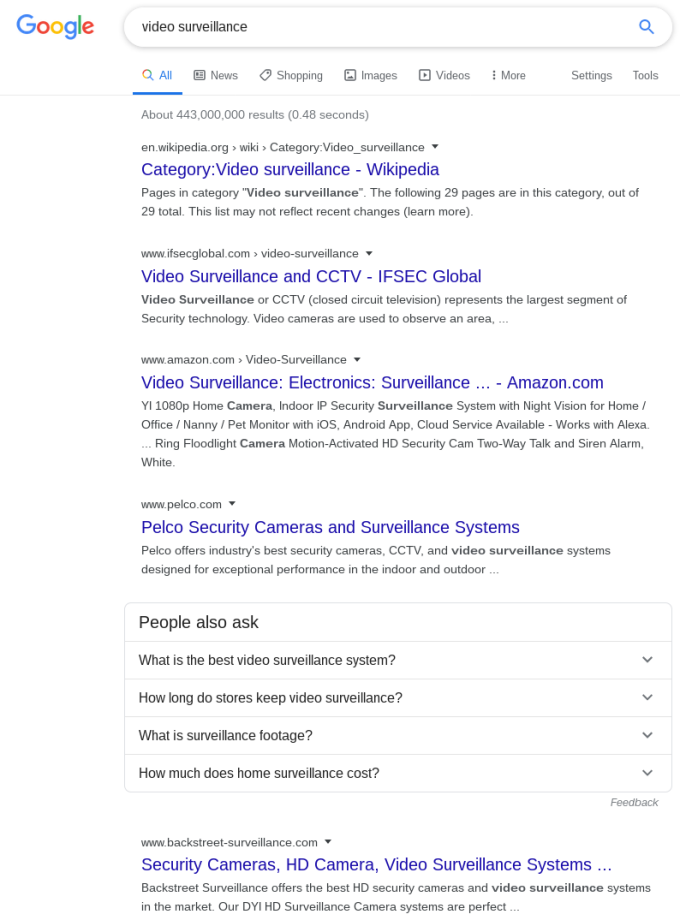
Examples of Retrieval Problems

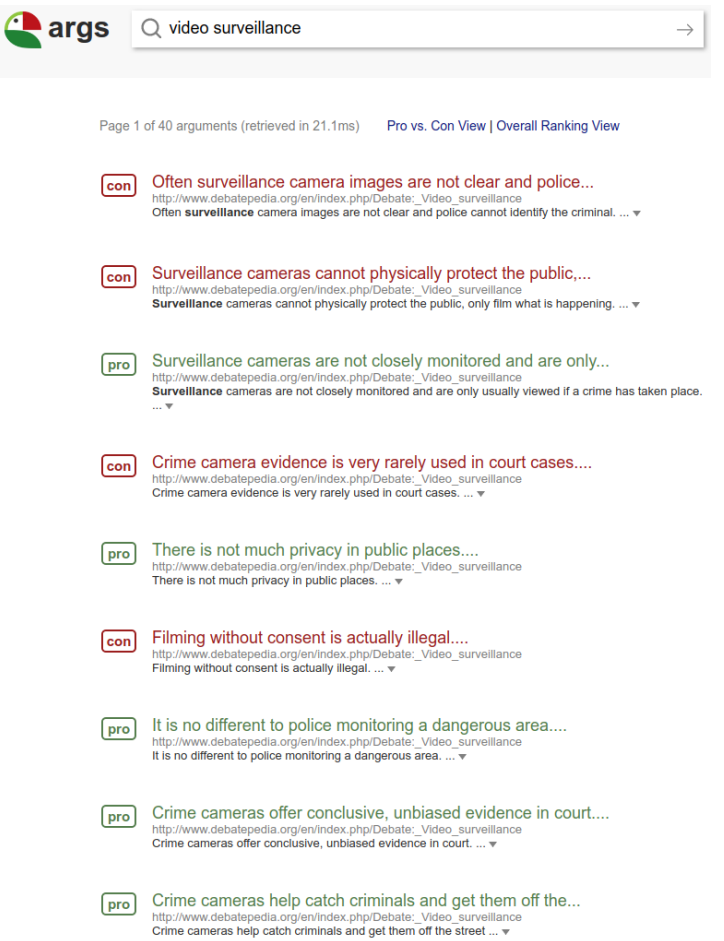
Write an essay about video surveillance.

Examples of Retrieval Problems

Write an essay about video surveillance.

Search for opinions on video surveillance.





Remarks:

- Users search for opinions and arguments on controversial topics or when making a purchase decision.
- Google's results include related questions from question answering platforms in a onebox. Nevertheless, there is no indication of the controversial nature of the topic. The results that link to Wikipedia are the only way to get background information. All others are commercial results.
- Search engines for argument retrieval, such as Args, retrieve arguments along with their stance (pro or con).

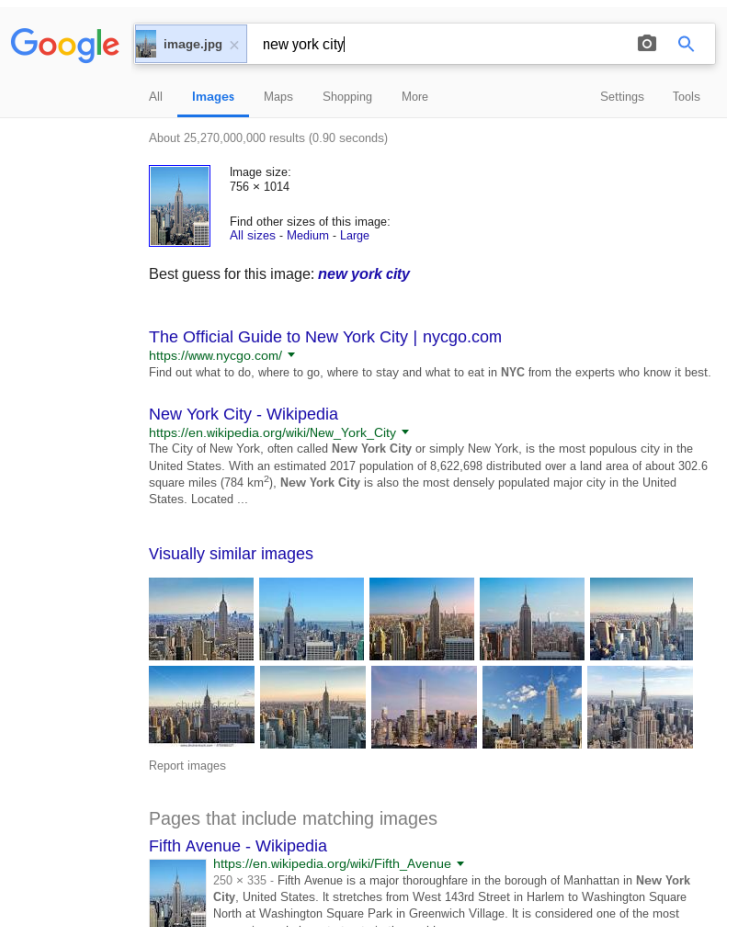
Examples of Retrieval Problems

Given an example image, find more like it.

Examples of Retrieval Problems

Given an example image, find more like it.

The image is an example of the information sought.



Examples of Retrieval Problems

Given an example text, find more like it.

Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikibooks
Wikiquote
Wikiversity
Wikivoyage

Languages
Boarisch
Deutsch
Español
Français

Mars

From Wikipedia, the free encyclopedia

This article is about the planet. For the deity, see Mars (mythology). For other uses, see Mars (disambiguation).

Mars is the fourth planet from the Sun and the second-smallest planet in the Solar System after Mercury. In English, Mars carries a name of the Roman god of war, and is often referred to as the "**Red Planet**"[14][15] because the reddish iron oxide prevalent on its surface gives it a reddish appearance that is distinctive among the astronomical bodies visible to the naked eye.[16] Mars is a terrestrial planet with a thin atmosphere, having surface features reminiscent both of the impact craters of the Moon and the valleys, deserts, and polar ice caps of Earth.

The rotational period and seasonal cycles of Mars are likewise similar to those of Earth, as is the tilt that produces the seasons. Mars is the site of Olympus Mons, the largest volcano and second-highest known mountain in the Solar System, and of Valles Marineris, one of the largest canyons in the Solar System. The smooth Borealis basin in the northern hemisphere covers 40% of the planet and may be a giant impact feature.[17][18] Mars has two moons, Phobos and Deimos, which are small and irregularly shaped. These may be captured asteroids,[19][20] similar to 5261 Eureka, a Mars trojan.

There are ongoing investigations assessing the past habitability potential of Mars, as well as the possibility of extant life. Future astrobiology missions are planned, including the Mars 2020 and ExoMars rovers.[21][22][23][24] Liquid water cannot exist on the surface of Mars due to low atmospheric pressure, which is less than 1% of the Earth's,[25] except at the lowest elevations for short periods.[26][27] The two polar ice caps appear to be made largely of water.[28][29] The volume of water ice in the south polar ice cap, if melted, would be sufficient to cover

Mars ♂

Mars in natural color in 2007[a]

Designations

Pronunciation	UK English: /mɑːz/ US English: /ˈmɑːrz/ (listen)
Adjectives	Martian

Orbital characteristics[2]

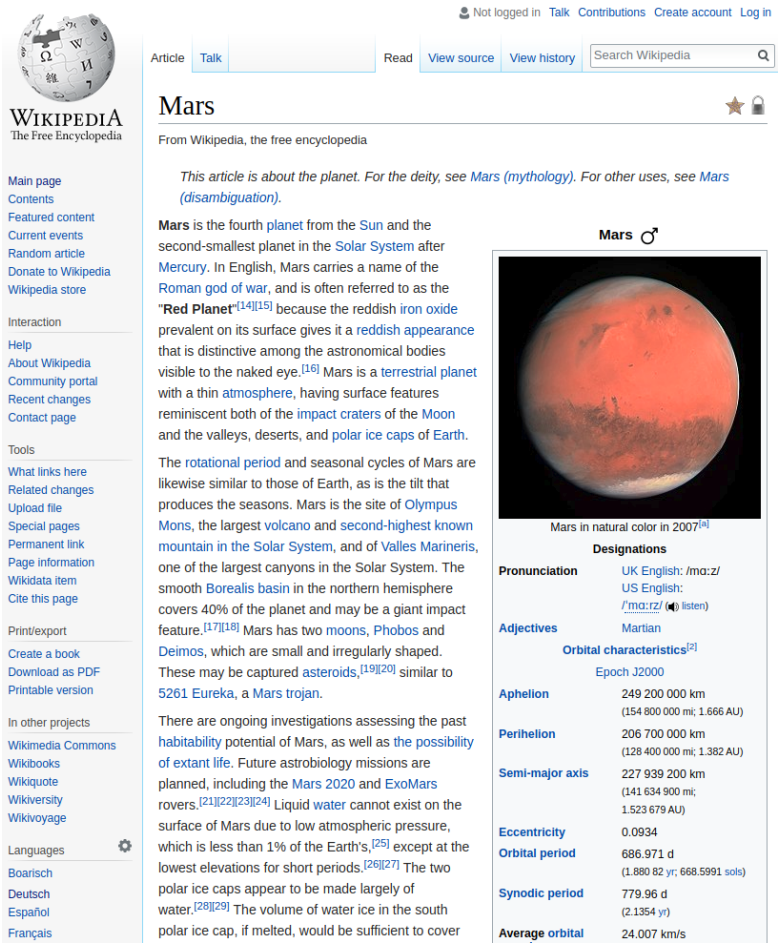
Epoch J2000

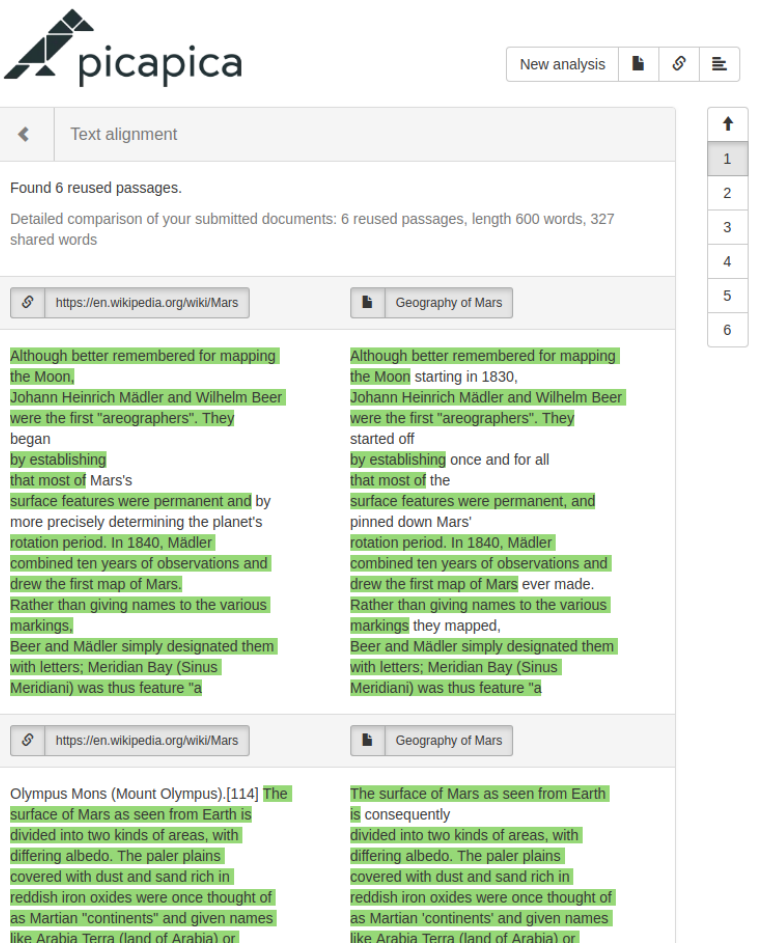
Aphelion	249 200 000 km (154 800 000 mi; 1.666 AU)
Perihelion	206 700 000 km (128 400 000 mi; 1.382 AU)
Semi-major axis	227 939 200 km (141 634 900 mi; 1.523 679 AU)
Eccentricity	0.0934
Orbital period	686.971 d (1.880 82 yr; 668.5991 sols)
Synodic period	779.96 d (2.1354 yr)
Average orbital	24.007 km/s

Examples of Retrieval Problems

Given an example text, find more like it.

The text is an example of the information sought.





Examples of Retrieval Problems

Given an example text, find more like it.

The text is an example of the information sought.

Query By Humming
Musical Information Retrieval in
An Audio Database

Asif Ghias Jonathan Logan David Chamberlin
Brian C. Smith

Cornell University
{ghias,bsmith}@cs.cornell.edu, logan@ghs.com, chamber@engr.sgi.com

ABSTRACT
The emergence of audio and video data types in databases will require new information retrieval methods adapted to the specific characteristics and needs of these data types. An effective and natural way of querying a musical audio database is by humming the tune of a song. In this paper, a system for querying an audio database by humming is described along with a scheme for representing the melodic information in a song as relative pitch changes. Relevant difficulties involved with tracking pitch are enumerated, along with the approach we followed, and the performance results of system indicating its effectiveness are presented.

KEYWORDS: Musical information retrieval, multimedia databases, pitch tracking

Introduction
Next generation databases will include image, audio and video data in addition to traditional text and numerical data. These data types will require query methods that are more appropriate and natural to the type of respective data. For instance, a natural way to query an image database is to retrieve images based on operations on images or sketches supplied as input. Similarly a natural way of querying an audio database (of songs) is to hum the tune of a song.

Such a system would be useful in any multimedia database containing musical data by providing an alternative and natural way of querying. One can also imagine a widespread use of such a system in commercial music industry, music radio and TV stations, music stores and even for one's personal use.

In this paper, we address the issue of how to specify a hummed query and report on an efficient query execution implementation using approximate pattern matching. Our approach hinges upon the observation that melodic contour, defined as the sequence of relative differences in pitch between successive notes, can be used to discriminate between melodies. Handel[3] indicates that melodic contour is one of the most important methods that listeners use to determine similarities between melodies. We currently use an alphabet of three possible relationships between pitches ('U', 'D', and 'S'), representing the situations where a note is above, below or the same as the previous note, which can be pitch-tracked quite robustly. With the current implementation of our system we are successfully able to retrieve most songs within 12 notes. Our database currently comprises a collection of all parts (melody and otherwise) from 183 songs, suggesting that three-way discrimination would be useful for finding a particular song among a private music collection, but that higher resolutions will probably be necessary for larger databases.

This paper is organized as follows. The first section describes the architecture of the current system. The second section describes what pitch is, why it is important in representing the melodic contents of songs, several techniques for tracking

Examples of Retrieval Problems

Given an example text, find more like it.

The text is an example of the information sought.

Query By Humming
Musical Information Retrieval in
An Audio Database

Asif Ghias Jonathan Logan David Chamberlin
Brian C. Smith

Cornell University
{ghias,bsmith}@cs.cornell.edu, logan@ghs.com, chamber@engr.sgi.com

ABSTRACT
The emergence of audio and video data types in databases will require new information retrieval methods adapted to the specific characteristics and needs of these data types. An effective and natural way of querying a musical audio database is by humming the tune of a song. In this paper, a system for querying an audio database by humming is described along with a scheme for representing the melodic information in a song as relative pitch changes. Relevant difficulties involved with tracking pitch are enumerated, along with the approach we followed, and the performance results of system indicating its effectiveness are presented.

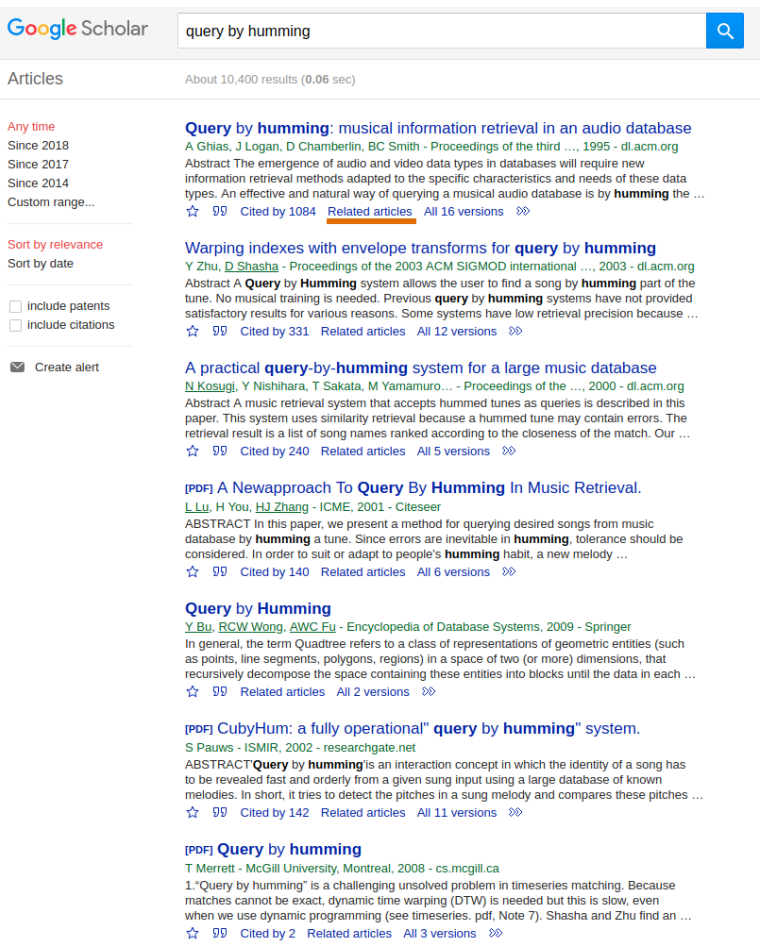
KEYWORDS: Musical information retrieval, multimedia databases, pitch tracking

Introduction
Next generation databases will include image, audio and video data in addition to traditional text and numerical data. These data types will require query methods that are more appropriate and natural to the type of respective data. For instance, a natural way to query an image database is to retrieve images based on operations on images or sketches supplied as input. Similarly a natural way of querying an audio database (of songs) is to hum the tune of a song.

Such a system would be useful in any multimedia database containing musical data by providing an alternative and natural way of querying. One can also imagine a widespread use of such a system in commercial music industry, music radio and TV stations, music stores and even for one's personal use.

In this paper, we address the issue of how to specify a hummed query and report on an efficient query execution implementation using approximate pattern matching. Our approach hinges upon the observation that melodic contour, defined as the sequence of relative differences in pitch between successive notes, can be used to discriminate between melodies. Handel[3] indicates that melodic contour is one of the most important methods that listeners use to determine similarities between melodies. We currently use an alphabet of three possible relationships between pitches ('U', 'D', and 'S'), representing the situations where a note is above, below or the same as the previous note, which can be pitch-tracked quite robustly. With the current implementation of our system we are successfully able to retrieve most songs within 12 notes. Our database currently comprises a collection of all parts (melody and otherwise) from 183 songs, suggesting that three-way discrimination would be useful for finding a particular song among a private music collection, but that higher resolutions will probably be necessary for larger databases.

This paper is organized as follows. The first section describes the architecture of the current system. The second section describes what pitch is, why it is important in representing the melodic contents of songs, several techniques for tracking



Google Scholar query by humming

Articles About 10,400 results (0.06 sec)

Any time
Since 2018
Since 2017
Since 2014
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

Query by **humming**: musical information retrieval in an audio database
A Ghias, J Logan, D Chamberlin, BC Smith - Proceedings of the third ..., 1995 - dl.acm.org
Abstract The emergence of audio and video data types in databases will require new information retrieval methods adapted to the specific characteristics and needs of these data types. An effective and natural way of querying a musical audio database is by **humming** the ...
☆ Cited by 1084 Related articles All 16 versions

Warping indexes with envelope transforms for **query** by **humming**
Y Zhu, D Shasha - Proceedings of the 2003 ACM SIGMOD international ..., 2003 - dl.acm.org
Abstract A **Query** by **Humming** system allows the user to find a song by **humming** part of the tune. No musical training is needed. Previous **query** by **humming** systems have not provided satisfactory results for various reasons. Some systems have low retrieval precision because ...
☆ Cited by 331 Related articles All 12 versions

A practical **query**-by-**humming** system for a large music database
N Kosugi, Y Nishihara, T Sakata, M Yamamuro... - Proceedings of the ..., 2000 - dl.acm.org
Abstract A music retrieval system that accepts hummed tunes as queries is described in this paper. This system uses similarity retrieval because a hummed tune may contain errors. The retrieval result is a list of song names ranked according to the closeness of the match. Our ...
☆ Cited by 240 Related articles All 5 versions

[PDF] A Newapproach To **Query** By **Humming** In Music Retrieval.
L Lu, H You, HJ Zhang - ICME, 2001 - Citeseer
ABSTRACT In this paper, we present a method for querying desired songs from music database by **humming** a tune. Since errors are inevitable in **humming**, tolerance should be considered. In order to suit or adapt to people's **humming** habit, a new melody ...
☆ Cited by 140 Related articles All 6 versions

Query by **Humming**
Y Bu, RCW Wong, AWC Fu - Encyclopedia of Database Systems, 2009 - Springer
In general, the term Quadtree refers to a class of representations of geometric entities (such as points, line segments, polygons, regions) in a space of two (or more) dimensions, that recursively decompose the space containing these entities into blocks until the data in each ...
☆ Related articles All 2 versions

[PDF] CubyHum: a fully operational" **query** by **humming**" system.
S Pauws - ISMIR, 2002 - researchgate.net
ABSTRACT'**Query** by **humming**'is an interaction concept in which the identity of a song has to be revealed fast and orderly from a given sung input using a large database of known melodies. In short, it tries to detect the pitches in a sung melody and compares these pitches ...
☆ Cited by 142 Related articles All 11 versions

[PDF] **Query** by **humming**
T Merrett - McGill University, Montreal, 2008 - cs.mcgill.ca
1."Query by humming" is a challenging unsolved problem in timeseries matching. Because matches cannot be exact, dynamic time warping (DTW) is needed but this is slow, even when we use dynamic programming (see timeseries. pdf, Note 7). Shasha and Zhu find an ...
☆ Cited by 2 Related articles All 3 versions

Remarks:

- Sometimes users cannot express their information need as a textual query, but instead specify an object that best illustrates the information they are looking for.
- Some search engines are tailored to search for specific multimedia examples, such as images, audio, or video.
- Using a text as an example, two goals can be pursued: finding other texts that deal with the same topic, or finding other texts that share reused text passages with the text in question. Google Scholar, for example, offers the search facet “Related Articles” to search for articles that correspond to a text found earlier. Picapica is a search engine for reused text.

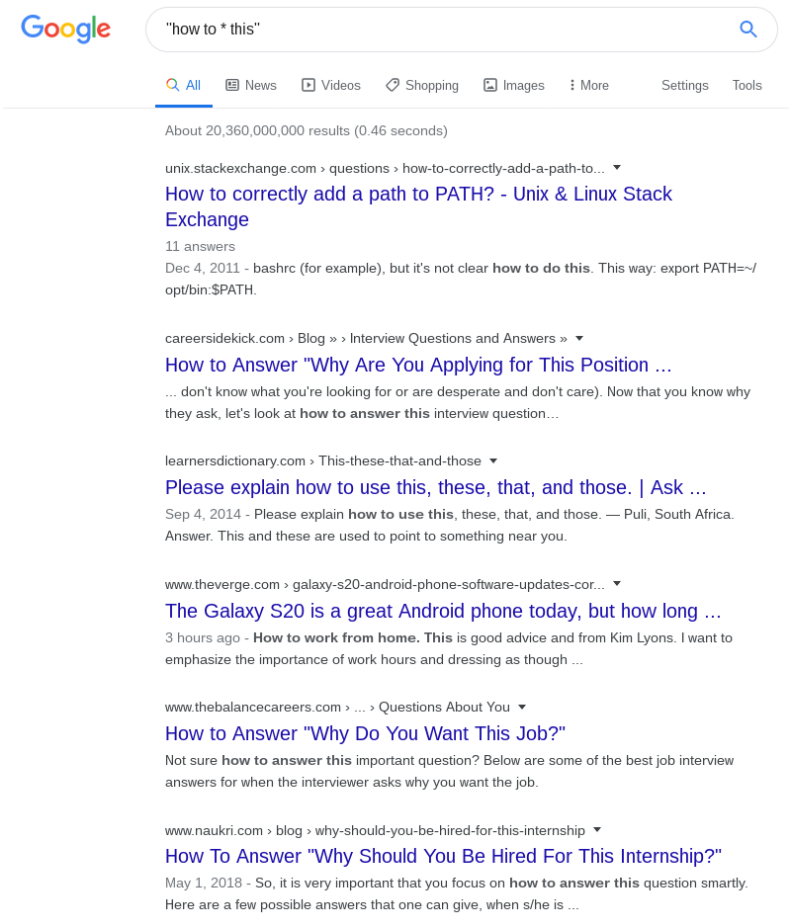
Examples of Retrieval Problems

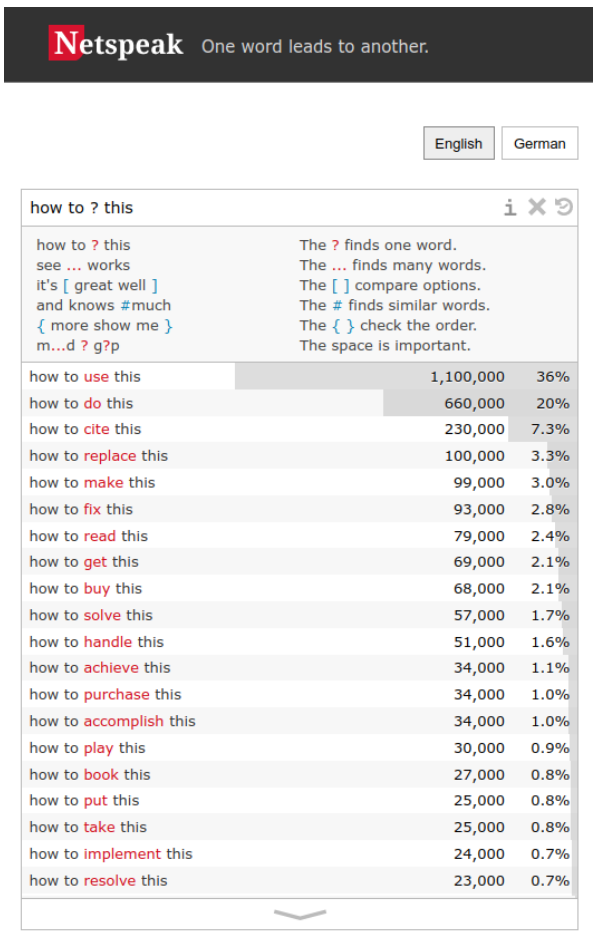
Figure out what people commonly write in the phrase `how to ? this.`

Examples of Retrieval Problems

Figure out what people commonly write in the phrase `how to ? this.`

Use wildcard search operators to find matching phrases.





Remarks:

- Users “misuse” search engines, as well as all other types of tools, to achieve goals that do not meet the tools’ originally intended purpose, either because of a lack of specialized tools or because the users do not know that specialized tools exist.
- While many web search engines support basic wildcard search operators, they cannot be used to solve this type of common formulation search task. The search engine interprets such a query in terms of its content and ranks the documents according to their relevance to the query. Moreover, only a few search results fit on a single page, while in practice many more alternatives may be available.
- Netspeak indexes short phrases along with their usage frequency on the web, and provides a wildcard search interface tailored to search by frequency of use.

Chapter IR:I

I. Introduction

- Information Retrieval in a Nutshell
- Examples of Retrieval Problems
- Terminology
- Delineation
- Historical Background
- Architecture of a Search Engine

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)

Data that is useful.

"useful": meaningful, interpretable, factual, instructional, informative, important

Definition 3 (Knowledge)

Knowledge is a thought characterized by one's justifiable belief that it is true.

Knowledge derives from factual or instructional information and enables the knower to act (e.g., to solve a problem).

Facts (+ references to justifying information) stored in a (structured) database can be considered a form of "externalized knowledge", often called a "knowledge base".

Remarks:

- Definitions of the three terms, but especially of information and knowledge, vary widely among scholars of epistemology and information science. Zins 2007 collects 44 different attempts. Our definitions are based on those of the DIKW pyramid.
- Data is usually organized in documents. Examples: a book, a videotape. A digital document corresponds to a specific bit sequence on a digital storage medium. Example: files on a hard disk that is formatted with a file system.
- Information can also be described as *data + queries*. Imagine a piece of information such as "The Earth has only one moon." This can be transformed into a query + binary answer, such as "Does the Earth have only one moon?" + "Yes". Subtract the "Yes" (at most 1 bit of information) and virtually all the semantic content remains, yet the query is neither true nor false. It is said to be de-alethicised (from aletheia, the Greek word for truth). [Floridi, 2015]
- The usefulness of a piece of information depends on context and on who is asking. Consider the sequence of symbols "$a^2 + b^2 = c^2$". Without any mathematical knowledge, it is useless. With some knowledge, it might prove useful to a pupil being asked about the Pythagorean theorem. With more knowledge, it becomes less useful again in most situations, since many people have memorized it a school and can still reproduce it.

Remarks: (continued)

- The analysis of knowledge forms the basis of epistemology. Most epistemologists have found it overwhelmingly plausible that what is false cannot be known. The idea behind the belief condition is that one can only know what one believes. To identify knowledge only with true belief would be implausible because a belief might be true even though it is formed improperly. This tripartite analysis of knowledge is abbreviated as the "JTB" analysis, for "justified true belief". It became something of a convenient fiction to suppose that this analysis was widely accepted throughout much of the history of philosophy. In fact, the JTB analysis was first articulated in the twentieth century by its attackers. Many counterexamples form the basis for new approaches to define knowledge.
  [Ichikawa and Steup, 2017]

- The JTB analysis is applicable in practice to one's own knowledge, as well as that claimed by others. Given proposition a, does one believe a, is the belief in a justified, and is a true?

  Example propositions:

  - a_1 = "I know Berlin is the capital of Germany."
  - a_2 = "I know Bonn is the capital of Germany."
  - a_3 = "I know the soccer match Bayern München vs. Dortmund ends with a tie."

Terminology

Definition 4 (Information System)

An organized system for collecting, creating, storing, processing, and distributing information, including hardware, software, operators, 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)

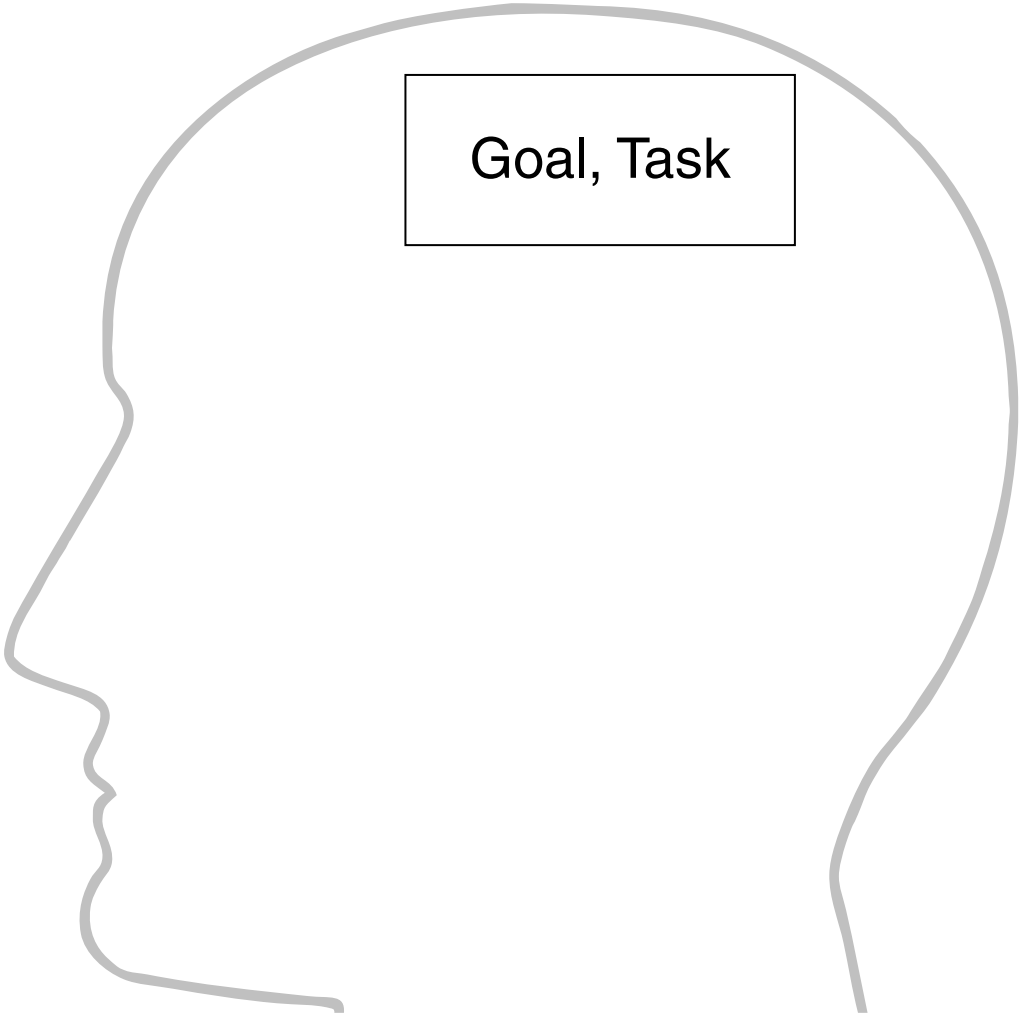
The degree to which a portion of data satisfies the information need of a user.

A portion of data is said to be relevant, if it is (partially) useful to satisfying a given information need. The closer it brings the user to satisfaction, the more relevant it is.

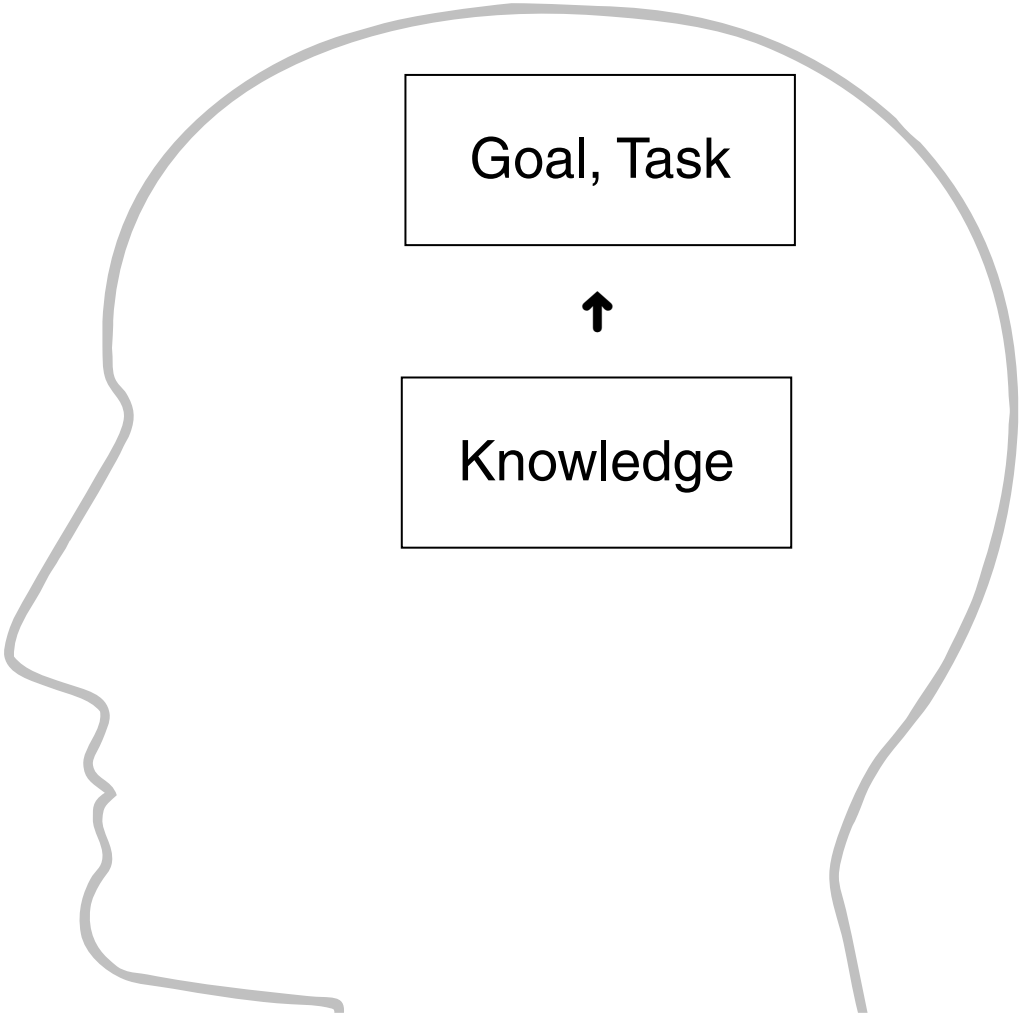
Remarks:

- 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]

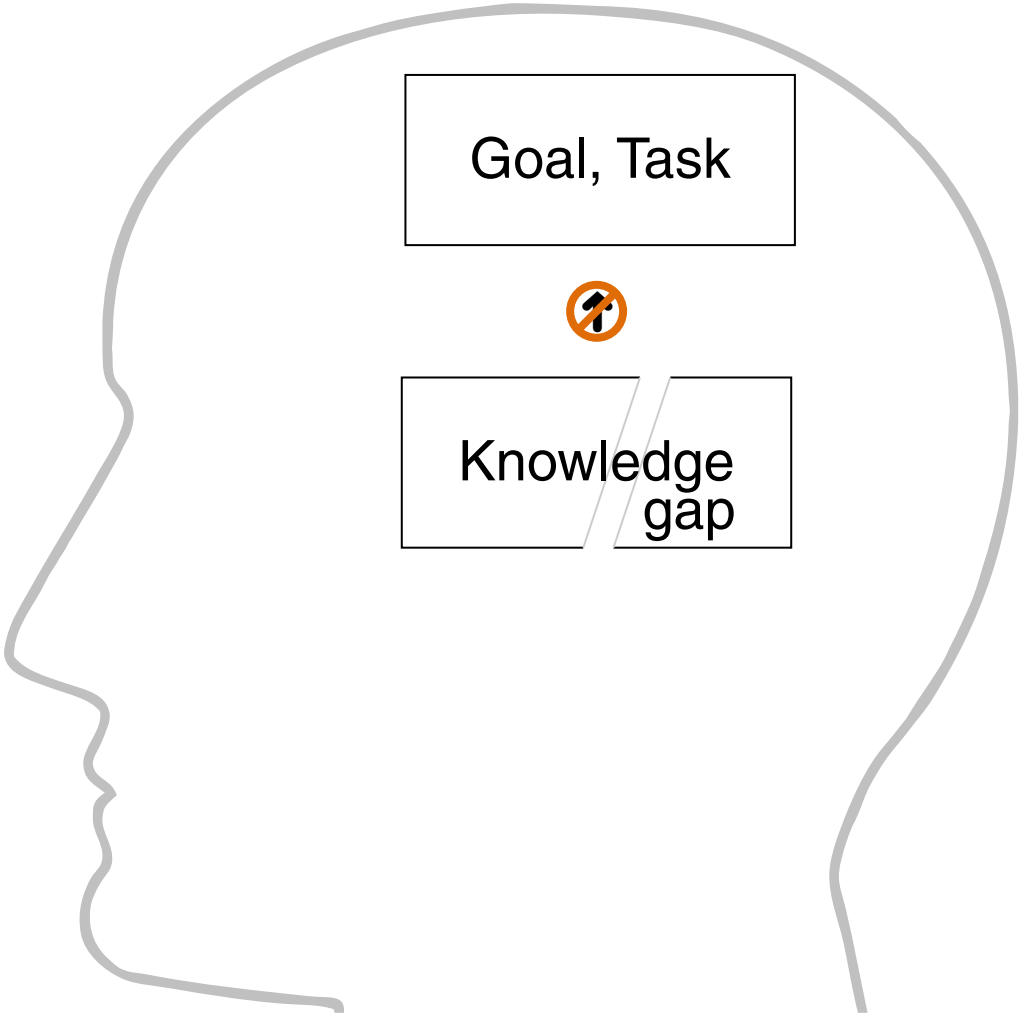
Terminology



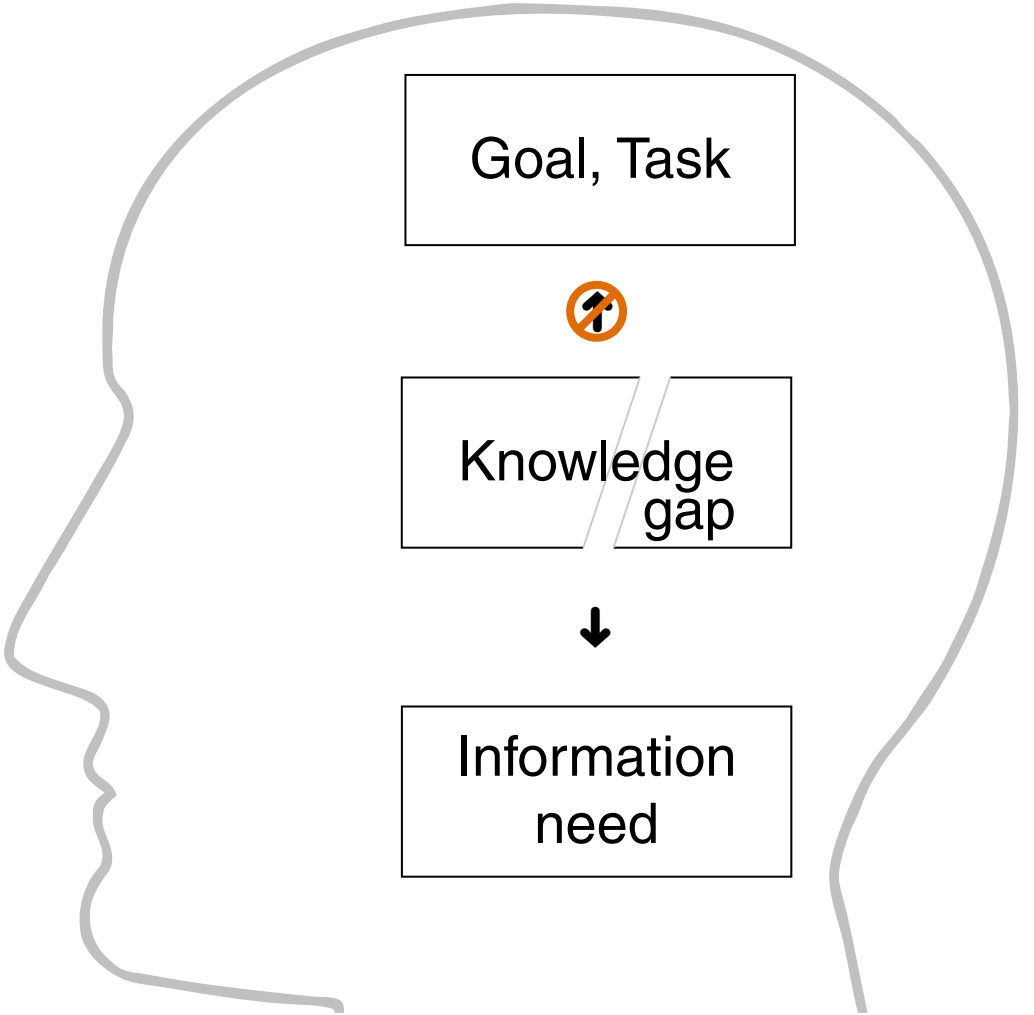
Terminology



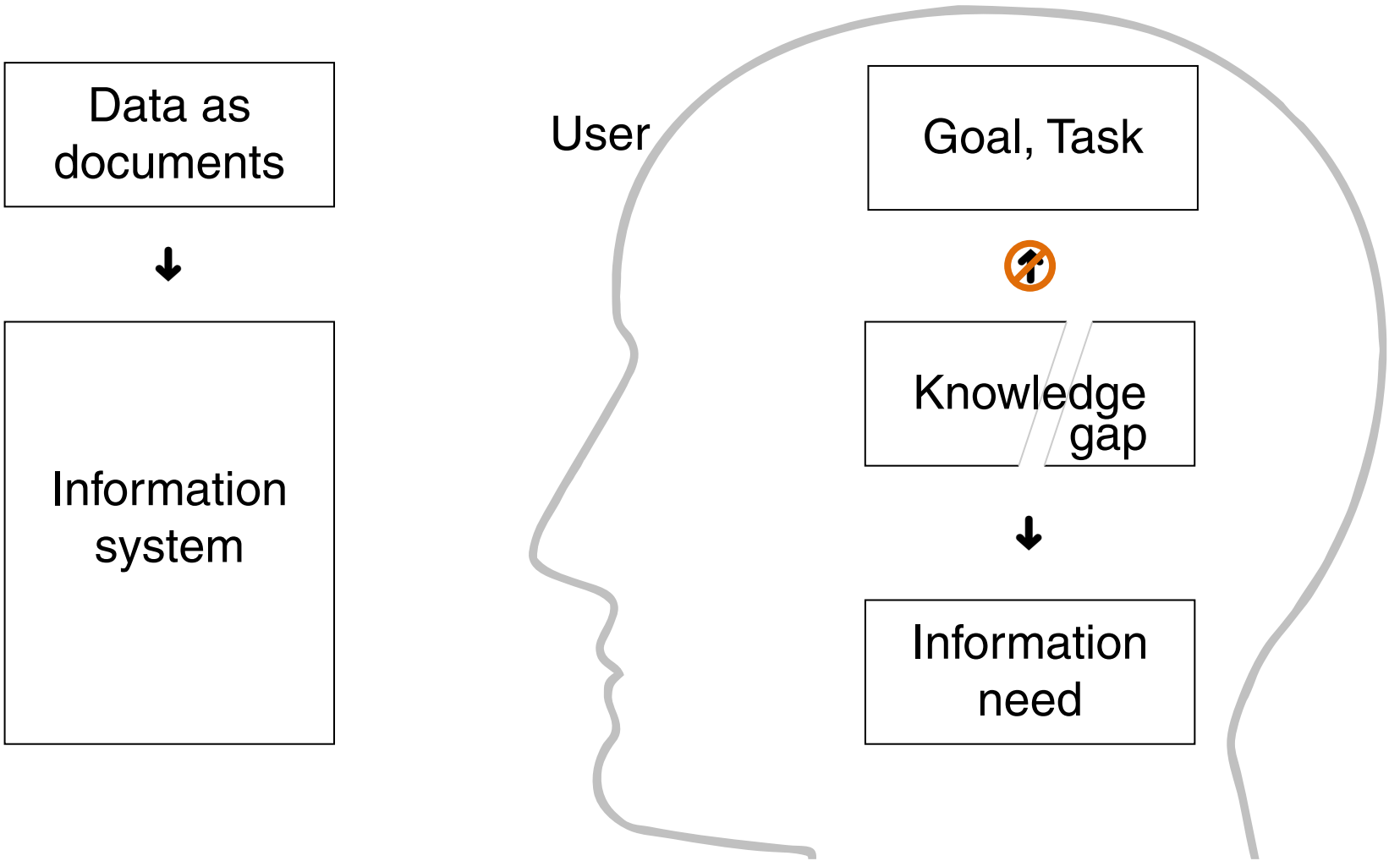
Terminology



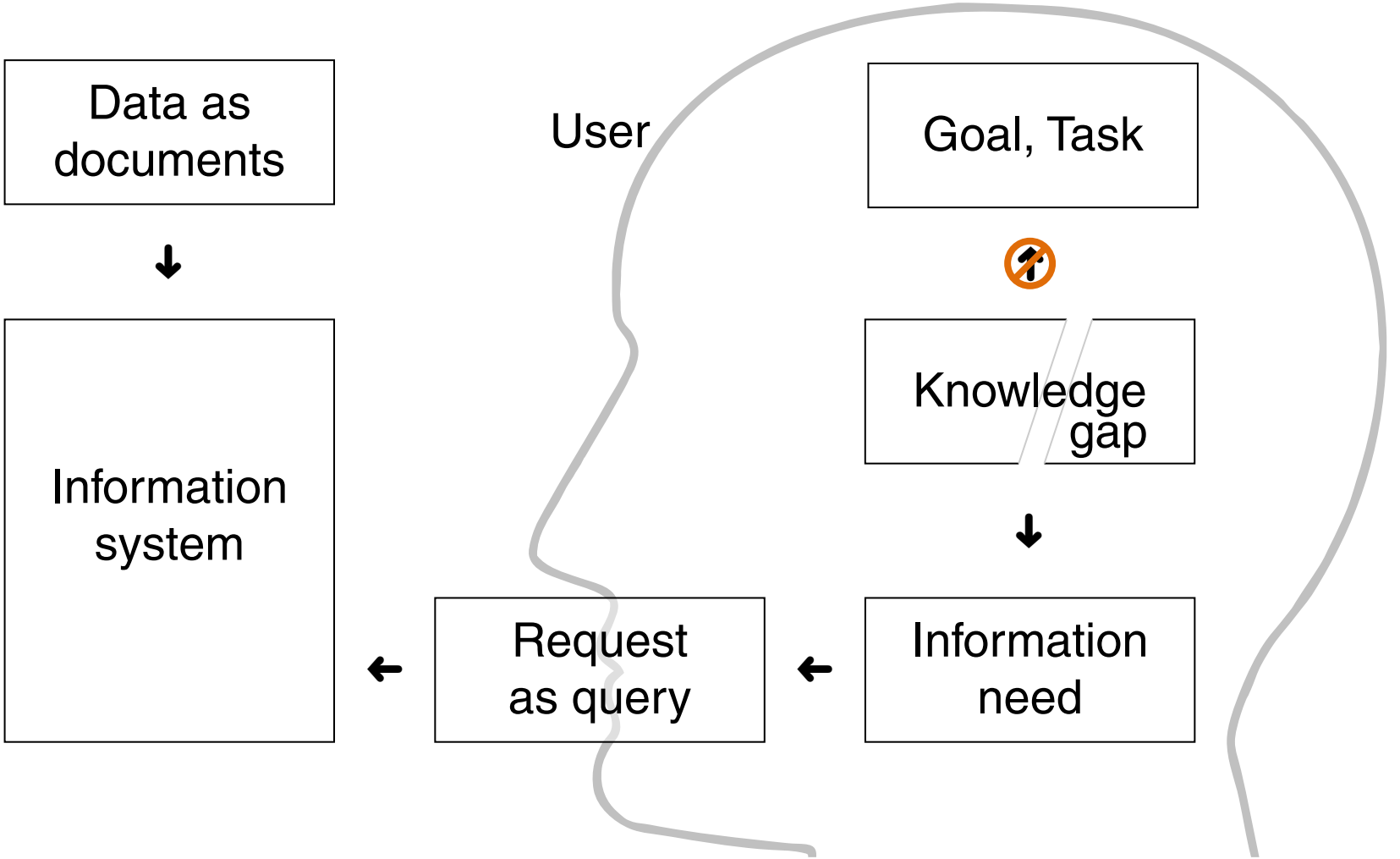
Terminology



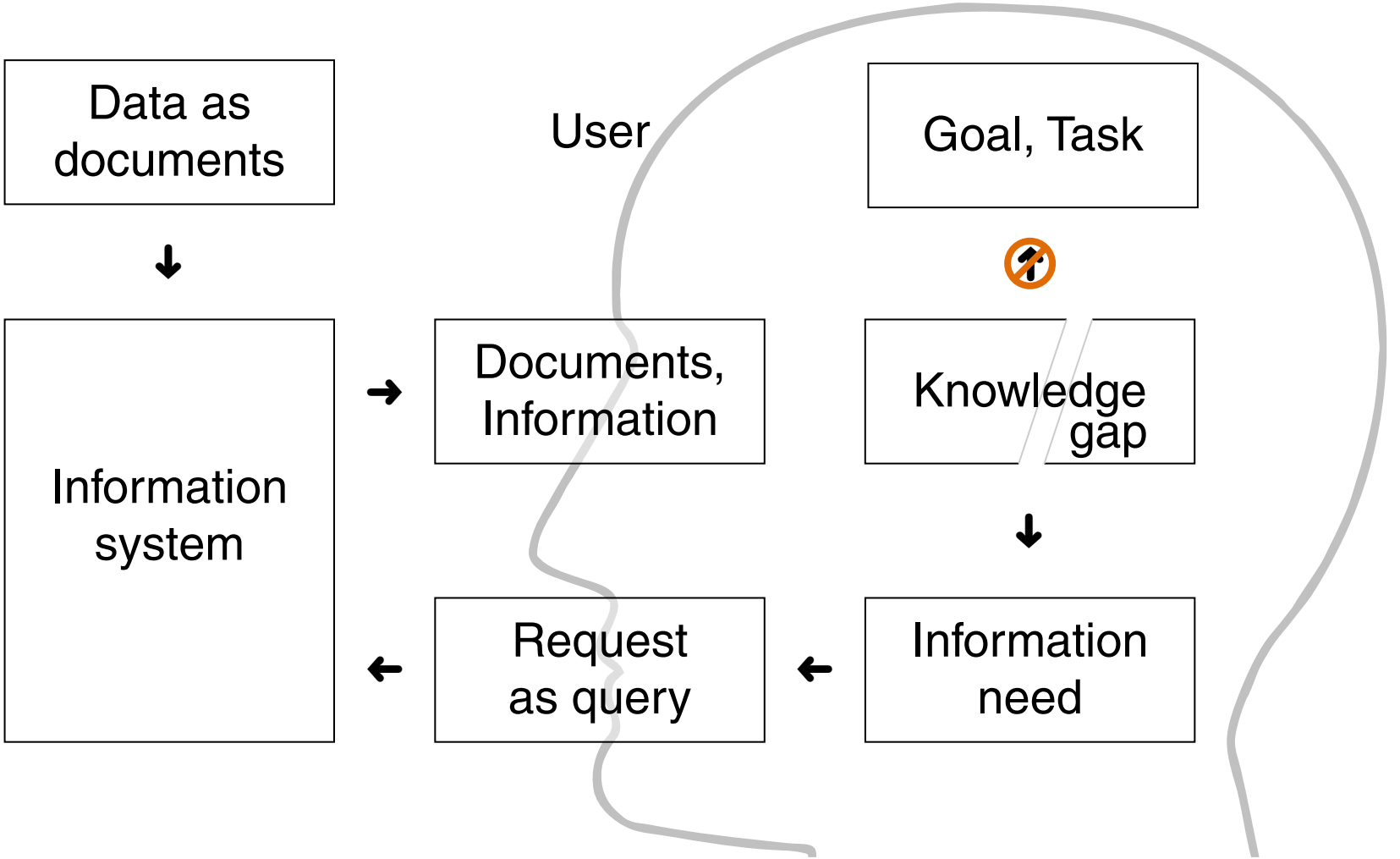
Terminology



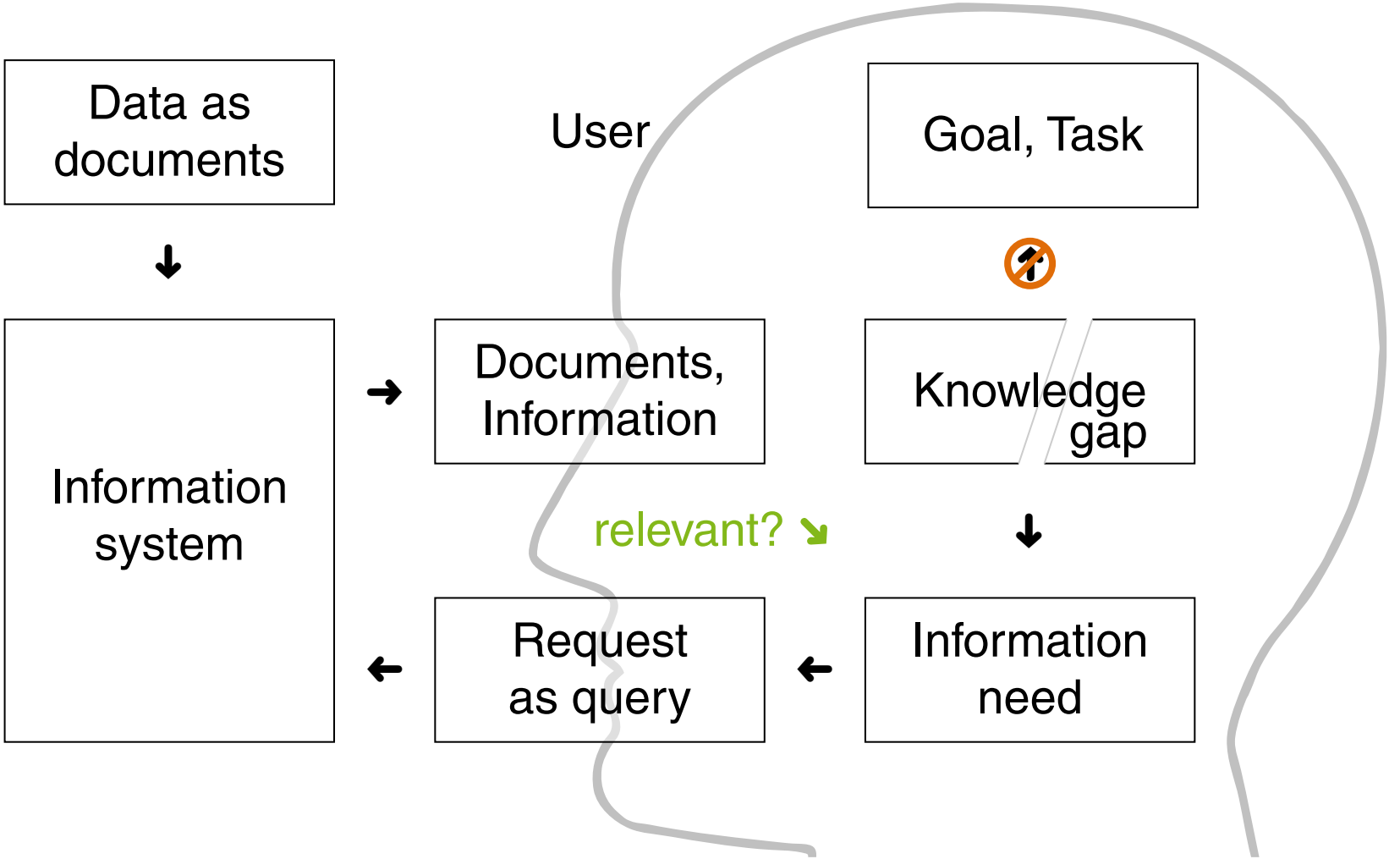
Terminology



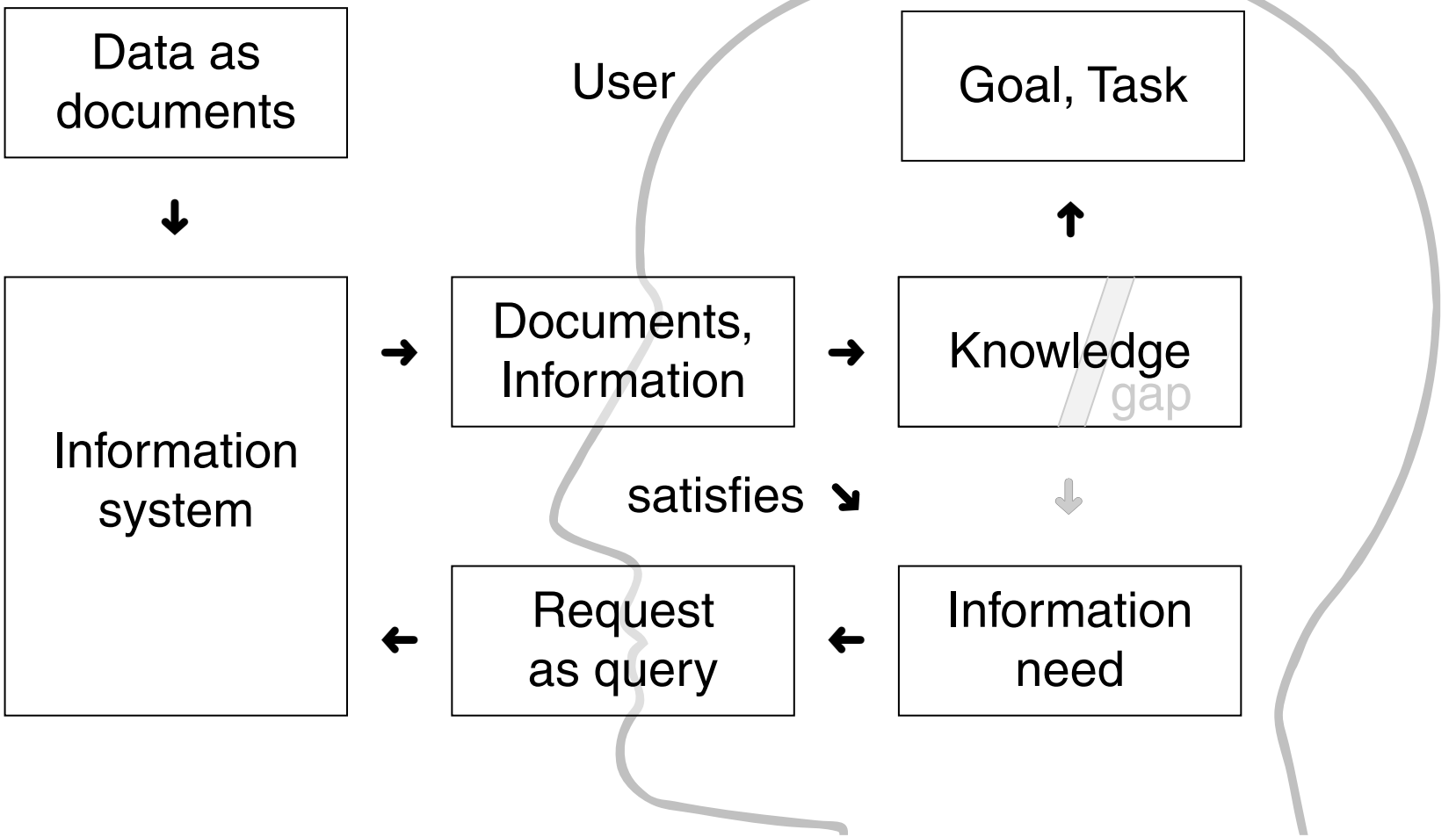
Terminology



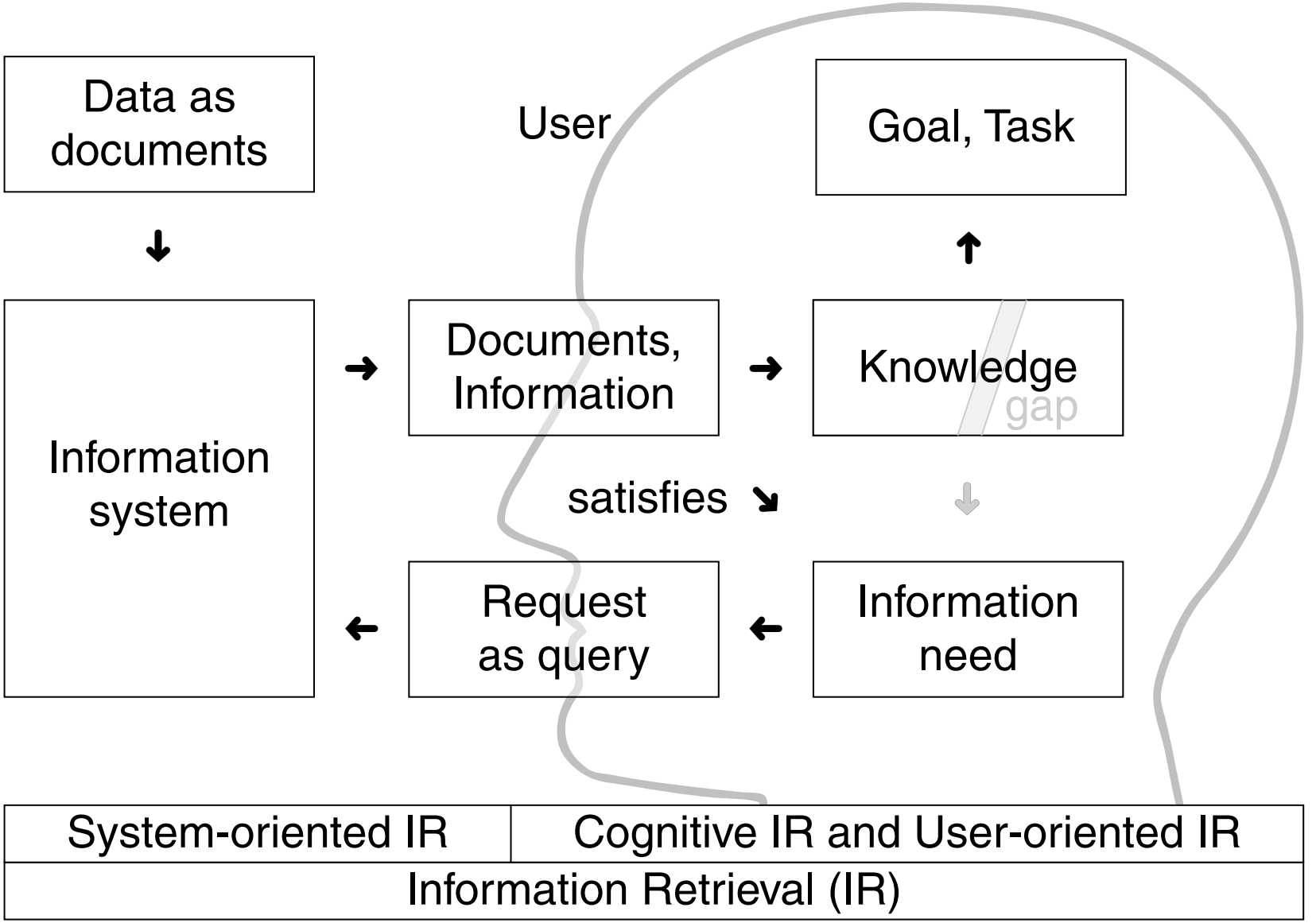
Terminology



Terminology



Terminology



Terminology

Definition 7 (Information Retrieval, IR)

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

As a research field, information retrieval 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, IR)

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

As a research field, information retrieval 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	retrieval technology
Cognitive IR	human interaction with retrieval 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, IR)

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

As a research field, information retrieval 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
   Unclear goals due to, e.g., vocabulary mismatch; refinement through interaction / dialog. Answers may depend on previous results or combine information from multiple 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]

Chapter IR:I

I. Introduction

- Information Retrieval in a Nutshell
- Examples of Retrieval Problems
- Terminology
- Delineation
- Historical Background
- Architecture of a Search Engine

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:

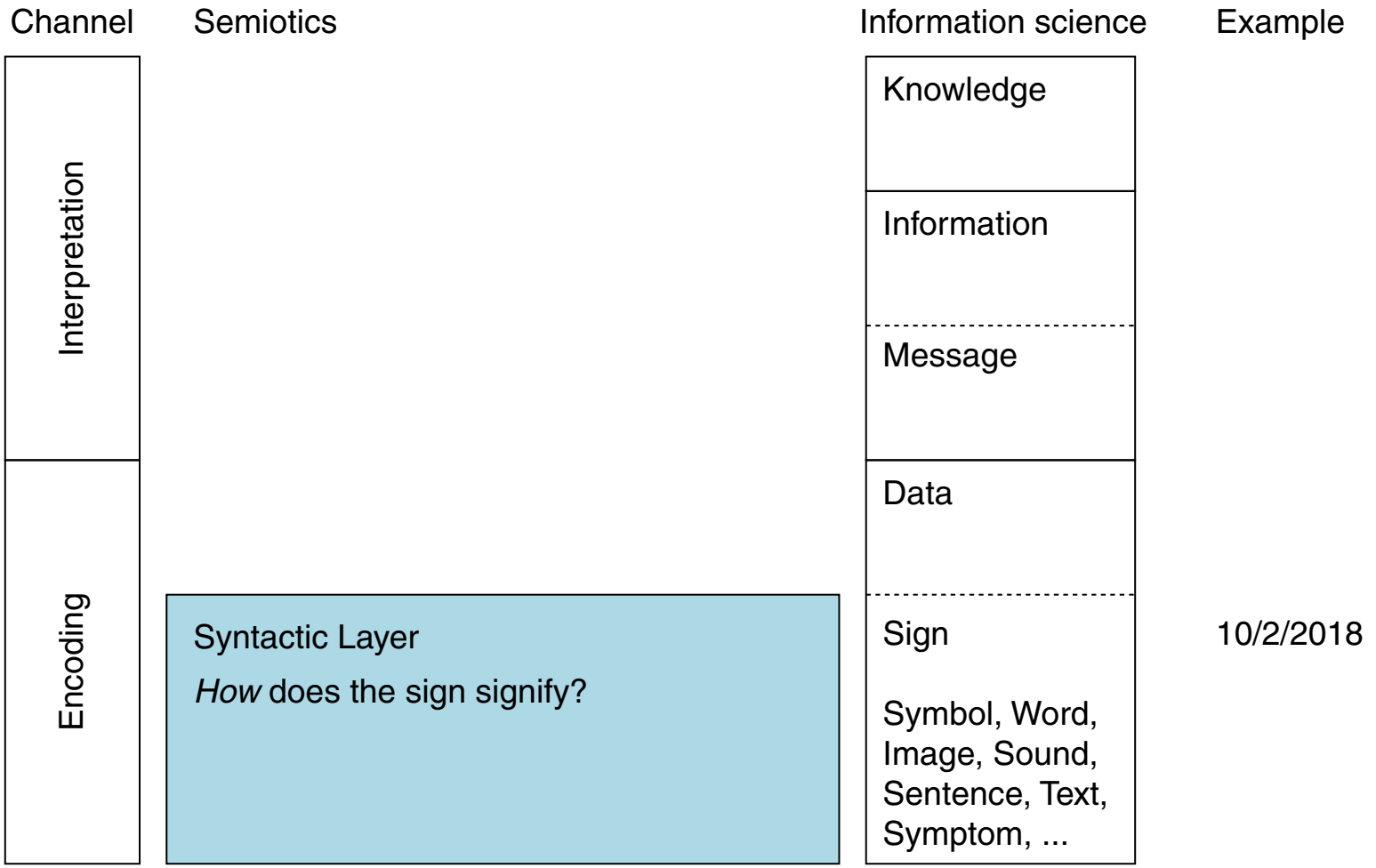
- A major difference between information retrieval (IR) systems and other kinds of information systems is the intrinsic uncertainty of IR. Whereas for database systems, an information need can always (at least for standard applications) be mapped precisely onto a query formulation, and there is a precise 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

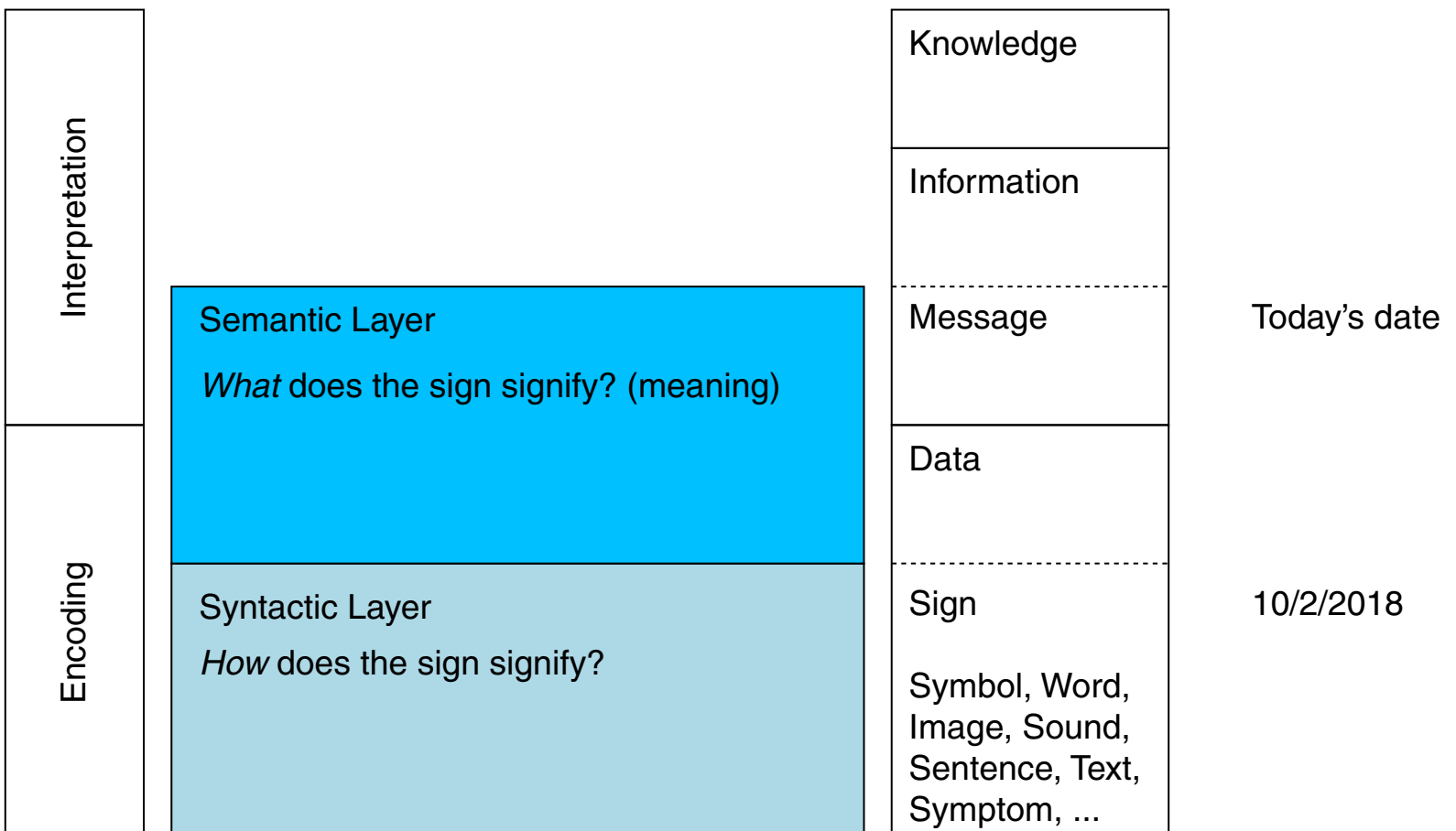
Communication



Delineation

Semiotics

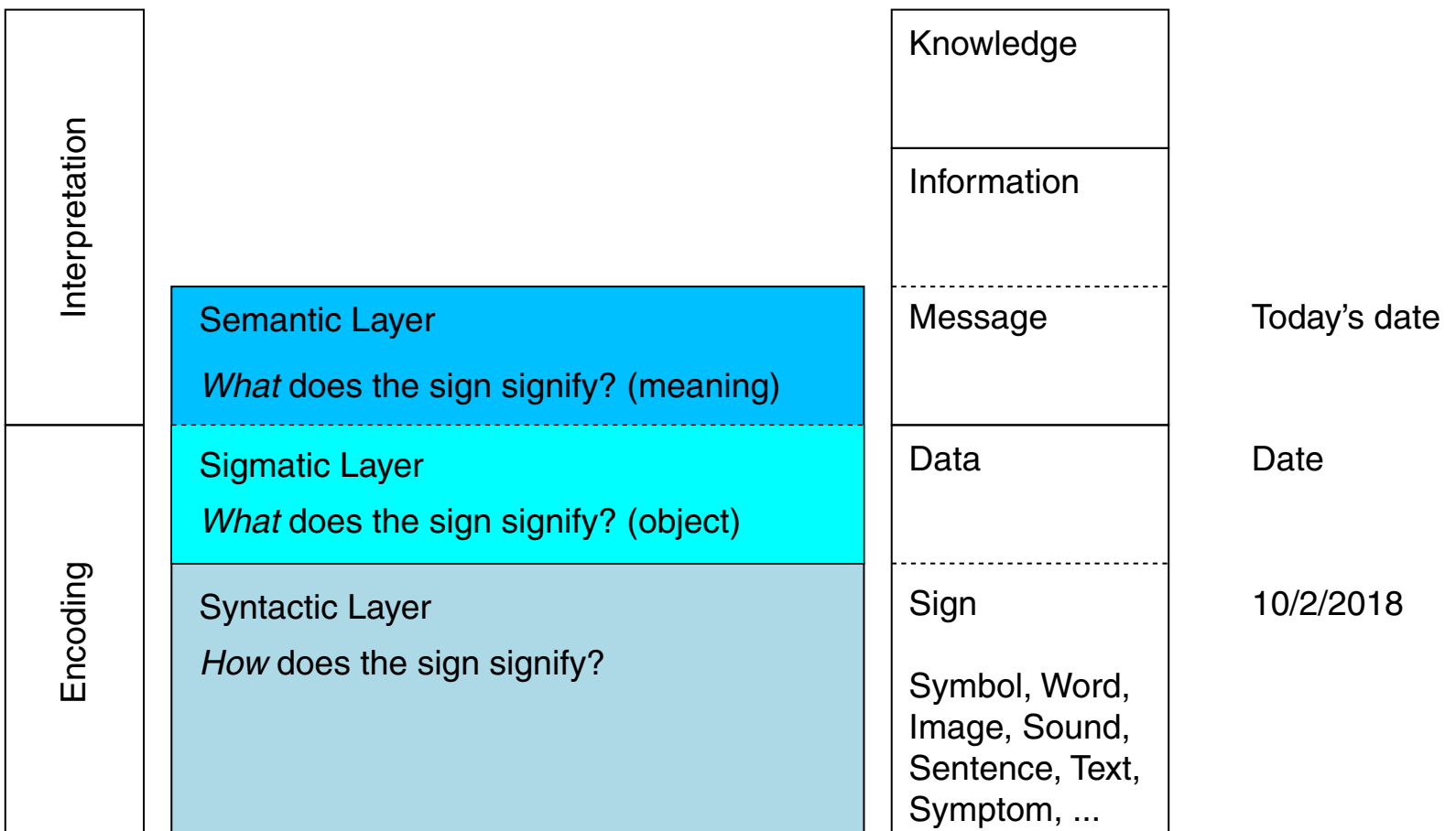
Communication



Delineation

Semiotics

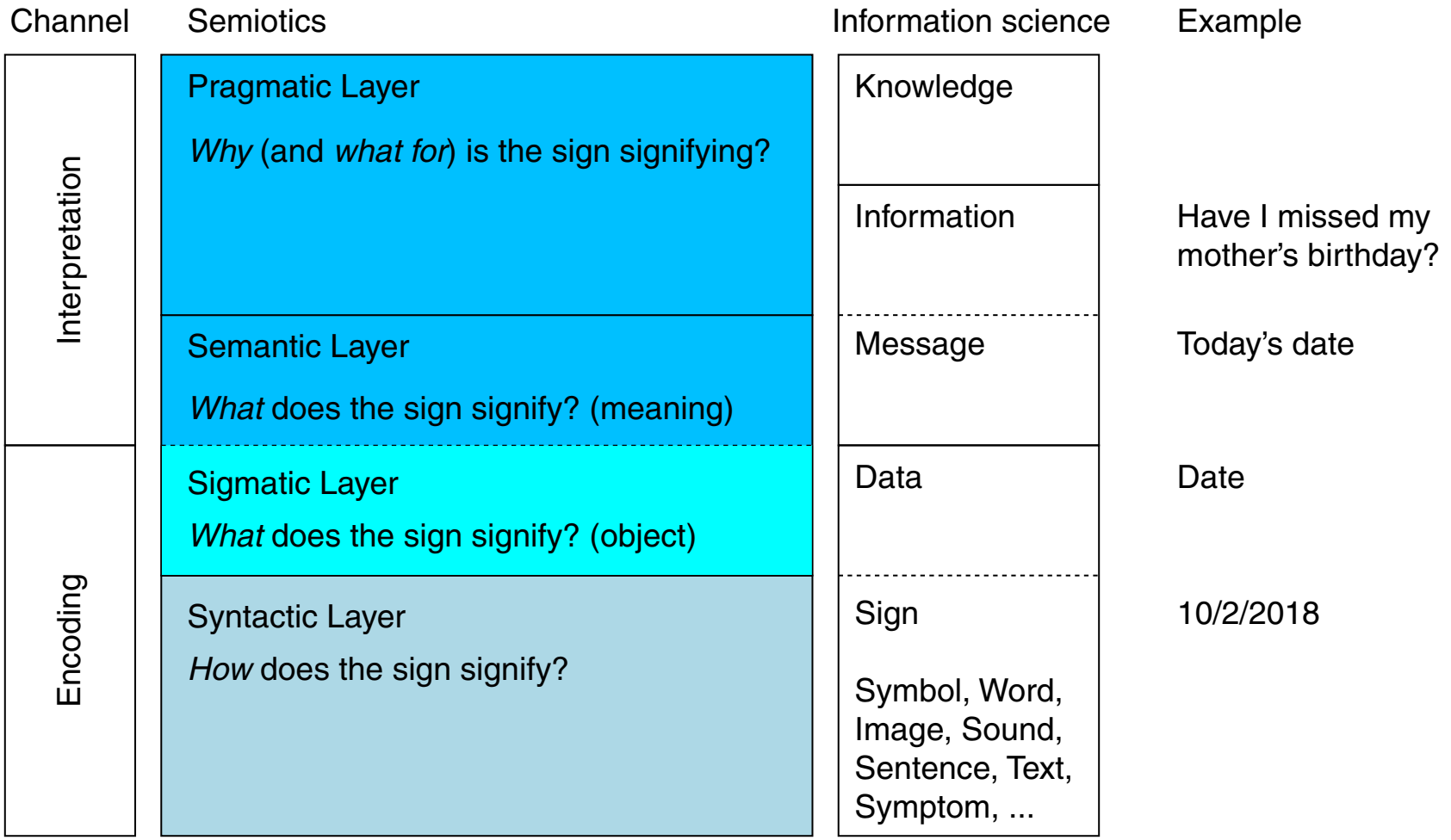
Communication



Delineation

Semiotics

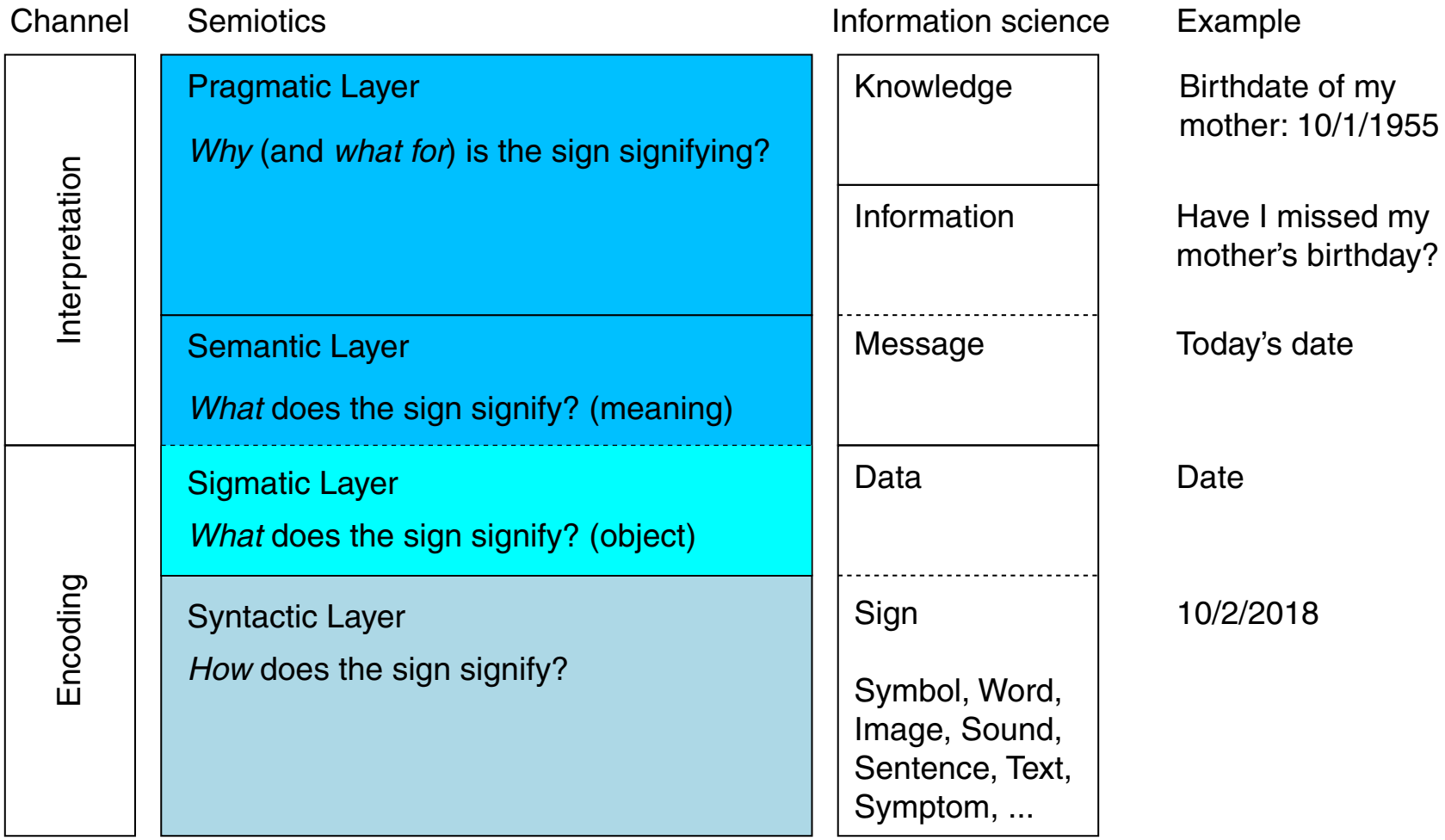
Communication



Delineation

Semiotics

Communication

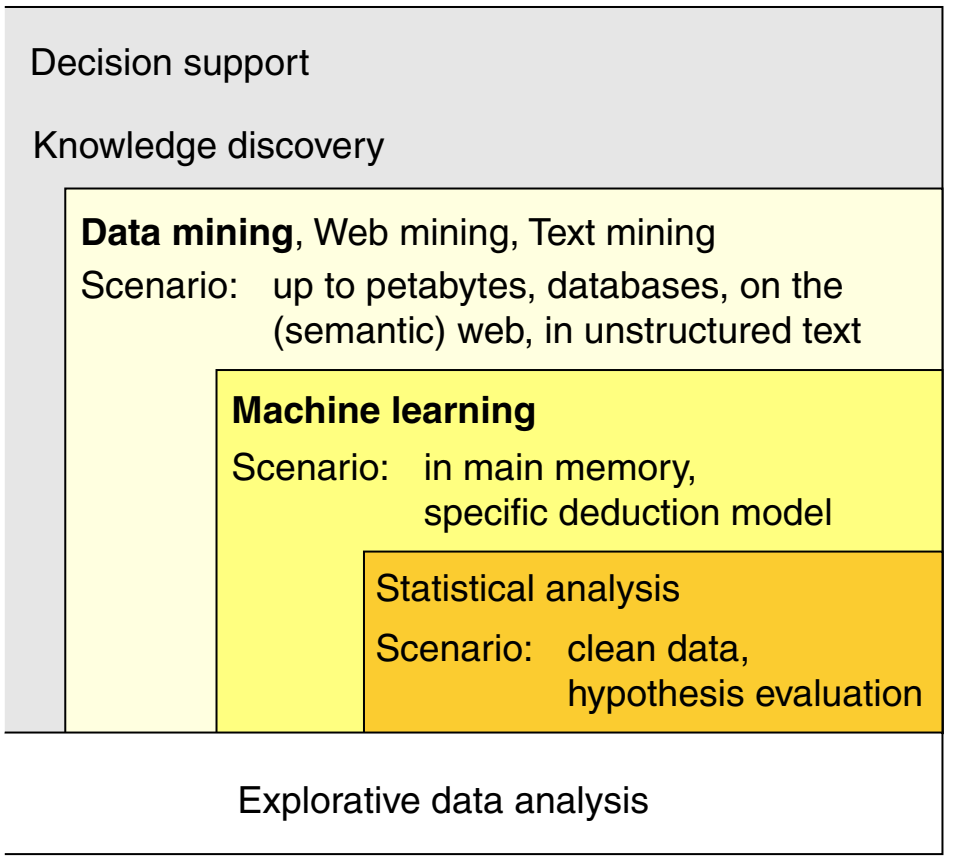


Remarks:

- 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 differentiates the semantic layer by distinguishing the relations between signs and the object to which they belong (sigmatics), and signs and their meaning (strict semantics). [Wikipedia]
- Information retrieval is an associative search that particularly addresses the semantics and pragmatics of documents.

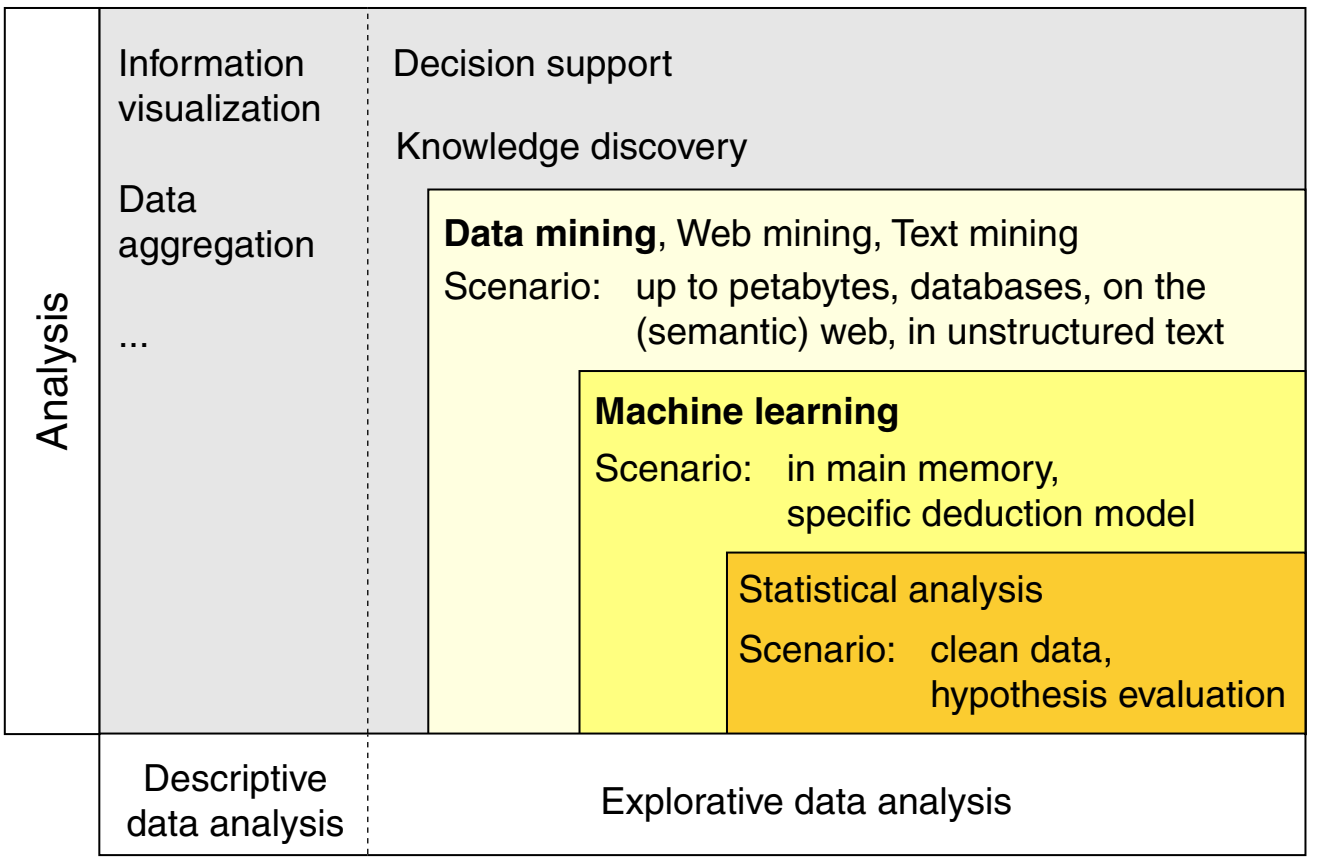
Delineation

Machine Learning, Data Mining



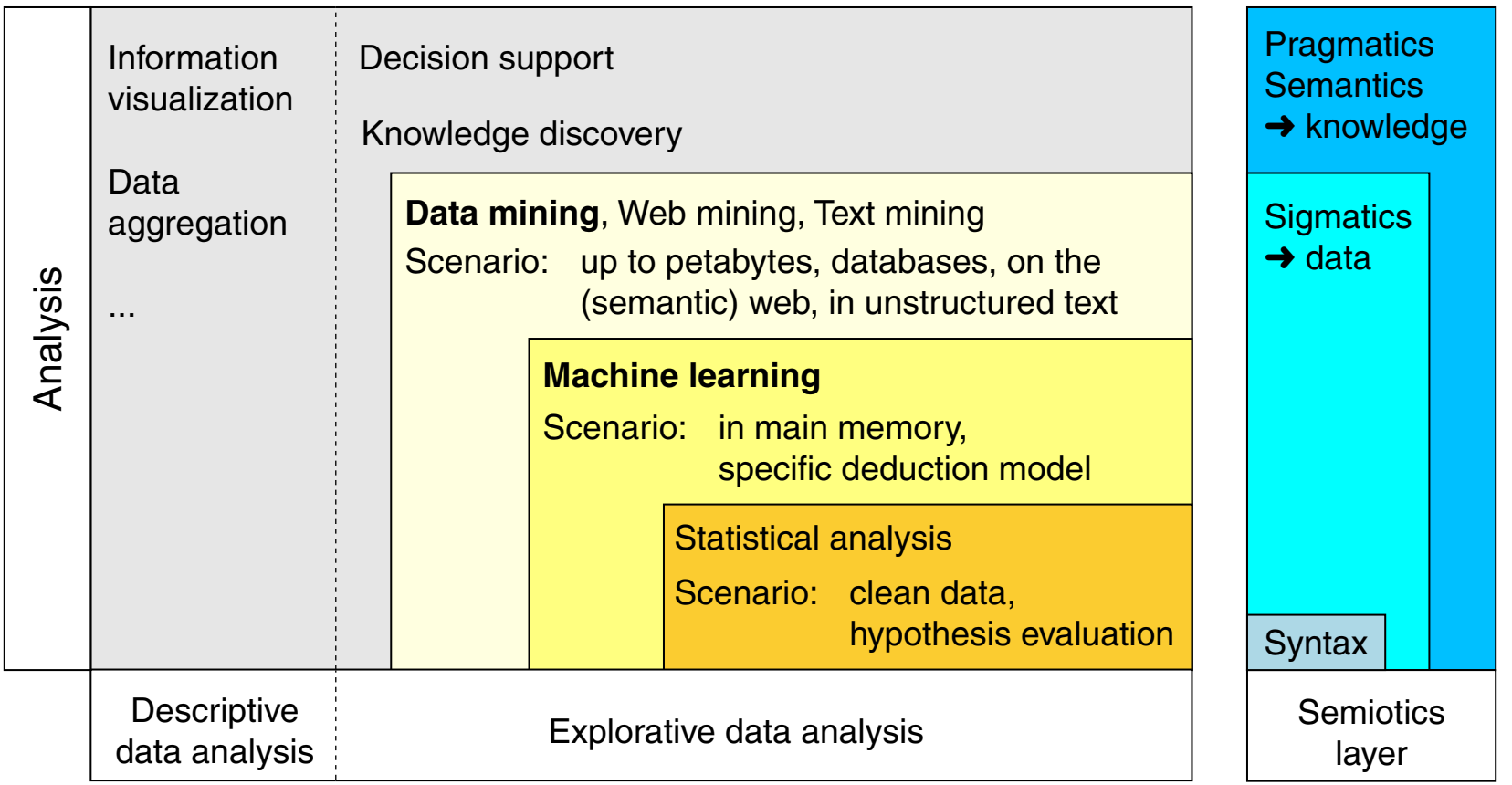
Delineation

Machine Learning, Data Mining



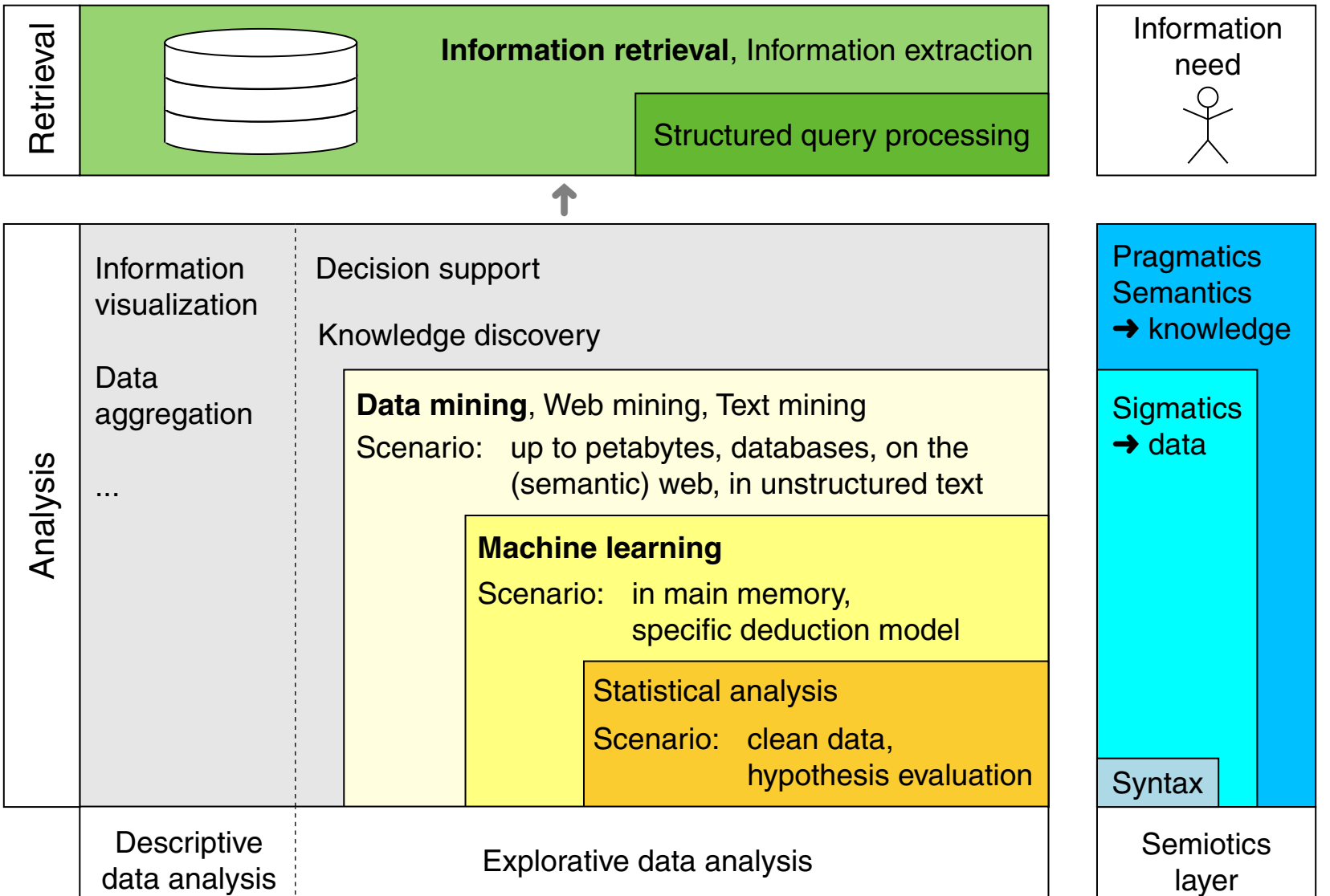
Delineation

Machine Learning, Data Mining



Delineation

Machine Learning, Data Mining



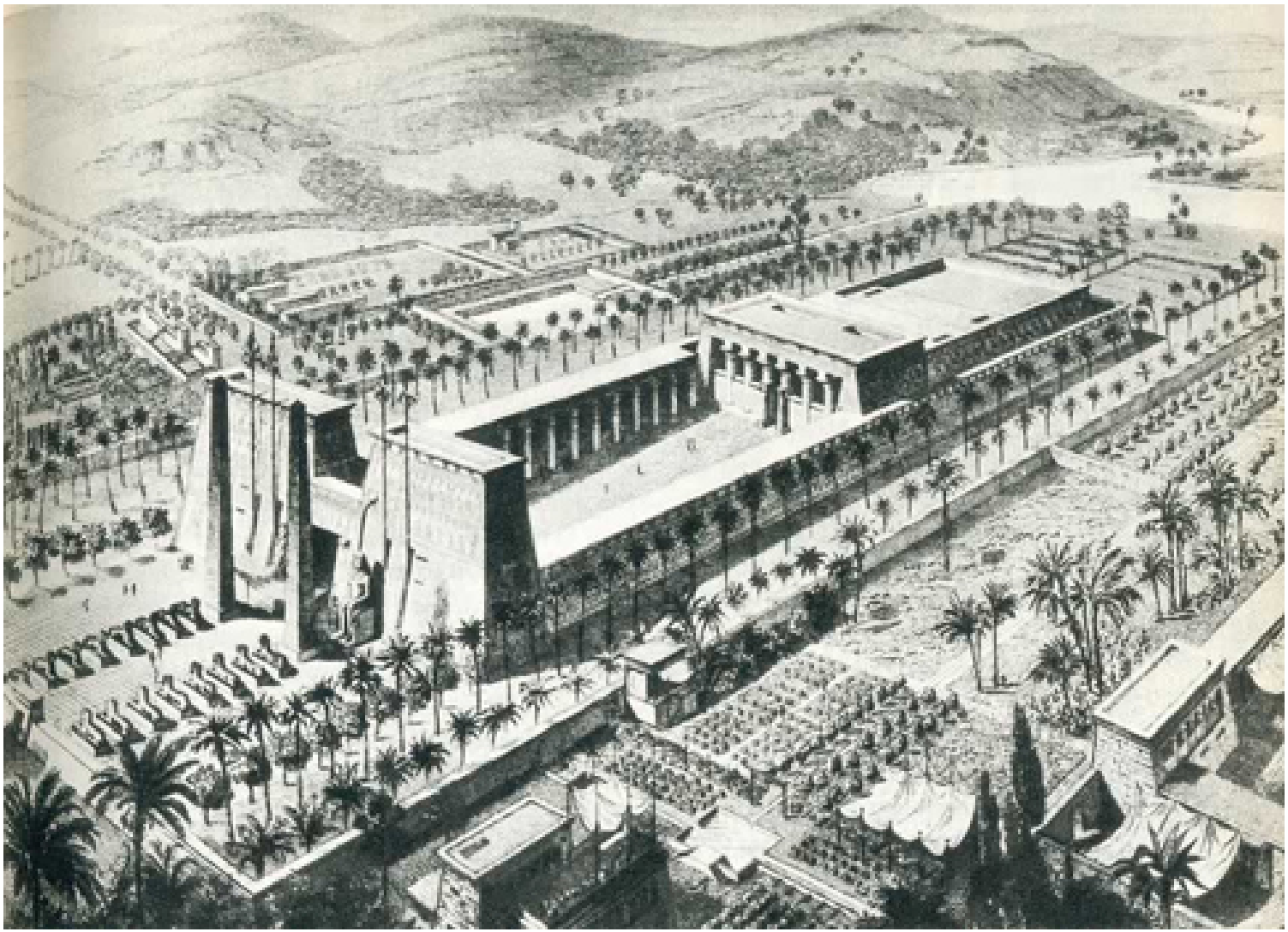
Chapter IR:I

I. Introduction

- Information Retrieval in a Nutshell
- Examples of Retrieval Problems
- Terminology
- Delineation
- Historical Background
- Architecture of a Search Engine

Historical Background

Manual Retrieval

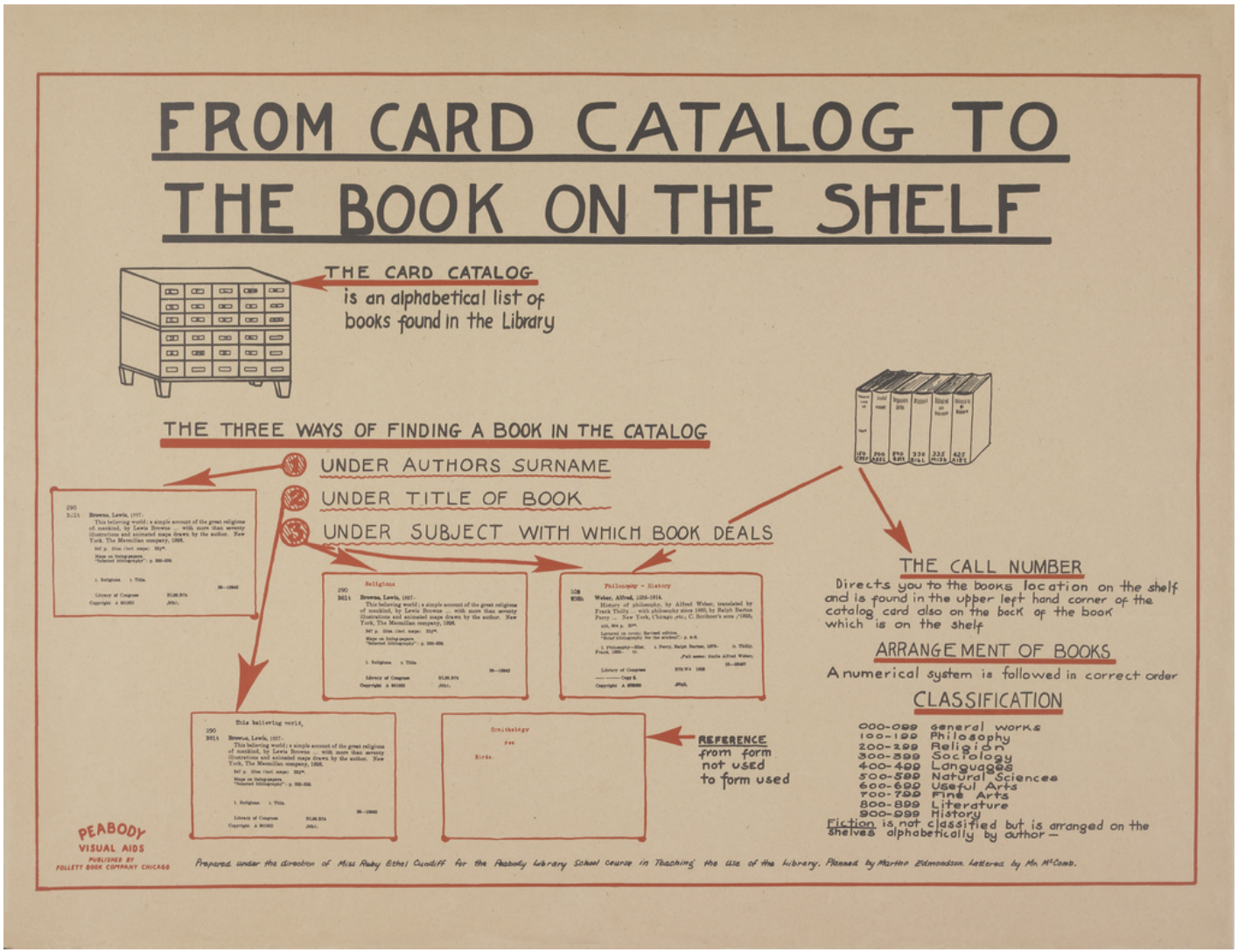


Remarks:

- 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 at Alexandria called the Mouseion, 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:

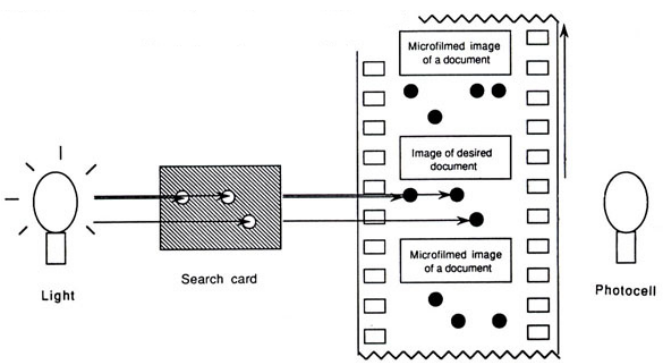
- Today, WorldCat is a union catalog that itemizes the collections of 72,000 libraries in 170 countries and territories. It contains more than 521 million records, representing over 3.2 billion physical and digital assets in 483 languages, as of February 2021. [OCLC]
- What are problems when sorting 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

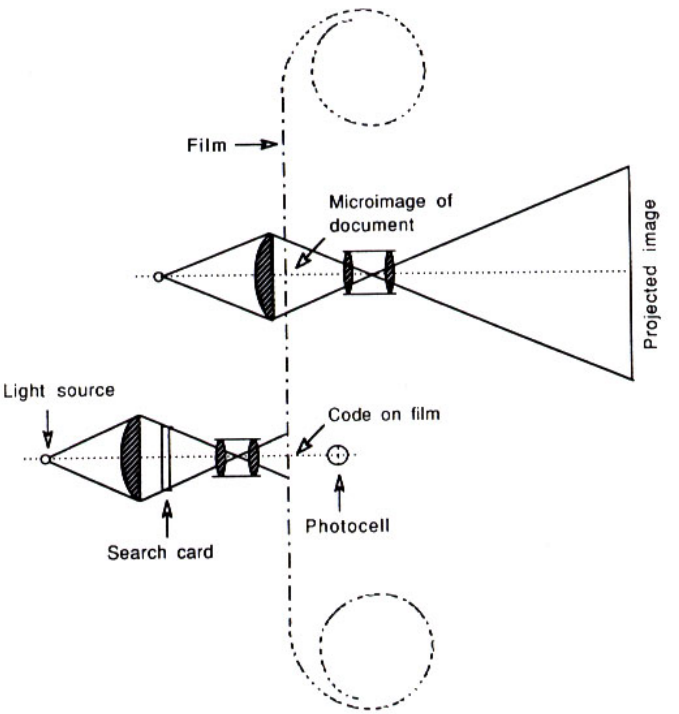
Mechanical Retrieval

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

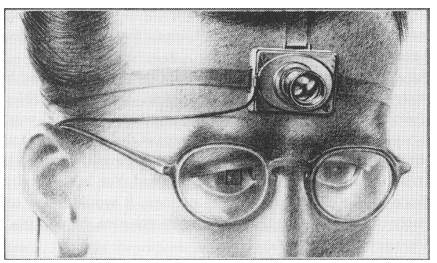


Result presentation

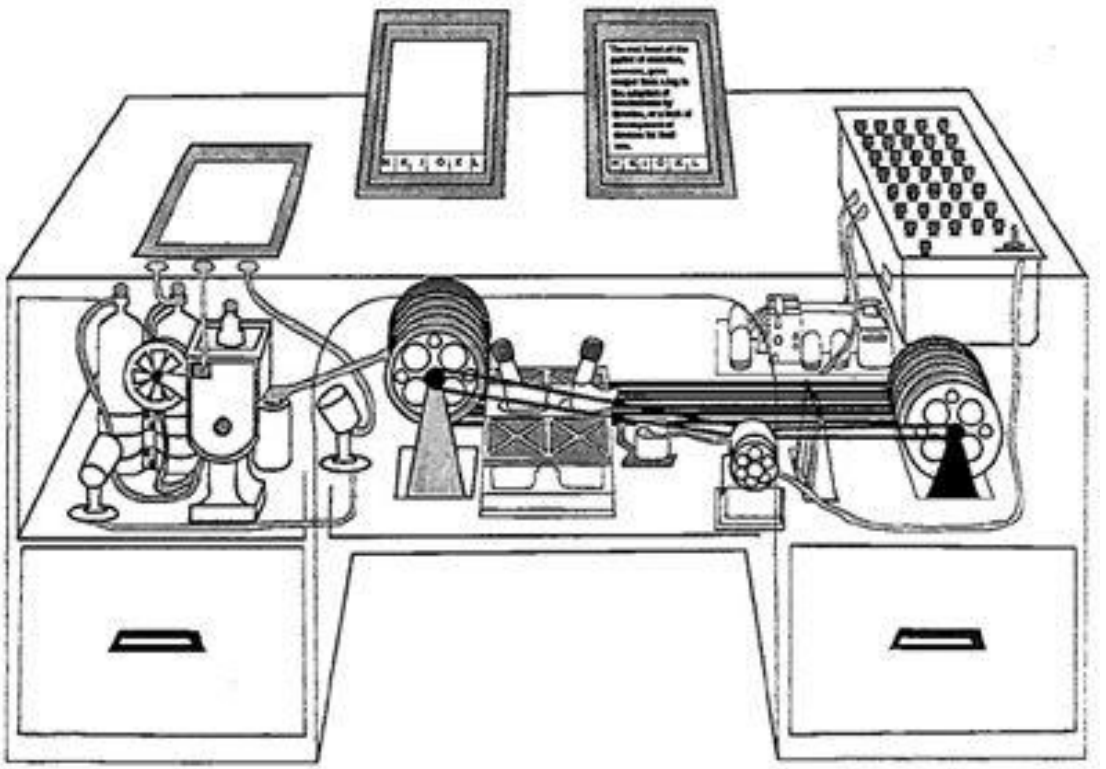
Historical Background

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.

On information growth (later called “information overload”) [Bagley, 1951]:

> [...] 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**.

Remarks:

- Serious research on information retrieval began after the end of World War II, when scientists in the Allied forces turned their attention away from warfare and found that the large amounts of scientific results and other information on a field of research that had accumulated during the war were too much for a single scientist to handle.
- Sanderson and Croft 2012 and Harman 2019 review the history of information retrieval.

Historical Background

Information Retrieval (1950s)

Indexing and ranked retrieval:

1951 **Coordinate Indexing**

Mortimer Taube proposes a “coordinate indexing” of documents based on a selection of independent “uniterms” (now called (index) terms or keywords) that departs from traditional schemes for categorizing subjects.

Assigning uniterms to documents is called indexing. Adding a reference to a document to the catalog cards for its uniterms is called posting. Retrieval is done by searching for a set of uniterms, collecting documents to which at least a subset of them have been assigned.

1957 **Term frequency-based ranking**

Hans Peter Luhn proposes to rank documents based on their relevance to a search query via the frequency of terms in a document as a measure of the importance of the terms.

1958 **Cranfield paradigm**

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

Subject indexing ➜ Uniterm indexing ➜ Full text indexing

Remarks:

- The 1950s and early 1960s saw what amounts to a disillusionment of how information is to be indexed for effective retrieval. The traditional method of indexing by subject (among others) was put into question via Taube's Coordinate Indexing approach. Yet both still rely on controlled vocabularies, respectively, and experts tasked with predicting (but at the same time also limiting) the terms users search for. The Cranfield projects, however, showed that "simply" using every term (e.g., lemmatized noun) of a given text for indexing is superior to subject and uniterm indexing. This way, the user is left with predicting what words a relevant document contains. Computer assistance, however, is a prerequisite for scaling up full text indexing and retrieval.

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.

1965 Integration of the state of the art into the SMART retrieval system.

1983 First laureate of the Gerard Salton Award, 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?
- Commercial applications that emerged at this time were largely developed without exploiting the findings of IR research. Not even relevance-based ranking was adopted. This situation did not change until the mid-1990s with the success of the Web.
  The simple Boolean retrieval models used instead still serve a very important purpose today in some areas such as patent search or prior art search, systematic literature reviews, etc.

Historical Background

Information Retrieval (1970s)

tf · *idf*-weighted Vector Space Model:

1972 **Inverse document frequency**

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

1975 **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:

1977 **Probability ranking principle**

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

1979 C.J. "Keith" van Rijsbergen proposes to incorporate term dependency into probabilistic retrieval models.

Historical Background

Information Retrieval (1980s – mid-1990s)

1990 **BM25**

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

1990 **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.

1991 **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 (non-)relevant documents.

1992 **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.

Historical Background

Information Retrieval (mid-1990s – 2000s)

Web search:

1994 Web crawlers are developed for the rapidly growing Web.

1994 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.

1997 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. At the same time, John M. Kleinberg proposes HITS.

1998 Maximum marginal relevance for diversity

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

1998 Language modeling for IR $\rightsquigarrow$ Neural IR, BERT

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

2002 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.

It’s been a long way

Historical Background

Information Retrieval (today)

DuckDuckGo

bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

Historical Background

Information Retrieval (today)

DuckDuckGo

bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

Computer Intelligence Test

Historical Background

Information Retrieval (today)

Baidu 百度

DuckDuckGo

bing

Google

amazon

Яндекс
Найдётся всё

YAHOO!

Computer Intelligence Test