

# Chapter IR:I

## I. Introduction

- ❑ Information Retrieval in a Nutshell
- ❑ Examples of Information Retrieval Problems
- ❑ Terminology
- ❑ Delineation
- ❑ Historical Background

# Information Retrieval in a Nutshell

- ❑ Billions of documents.  
Text, images, audio files, videos.

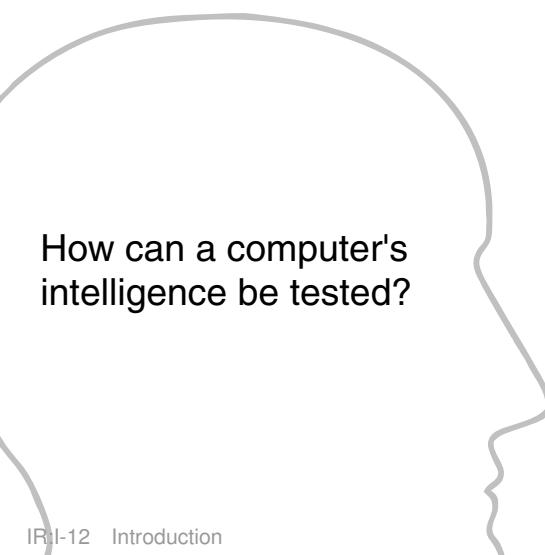
- ❑ A vague request.  
Expression of a complex information need: 1-5 keywords, or a question.

- ❑ High class imbalance.  
Only a handful of documents are relevant to the request.

→ Retrieve the relevant documents in milliseconds.

# Information Retrieval in a Nutshell

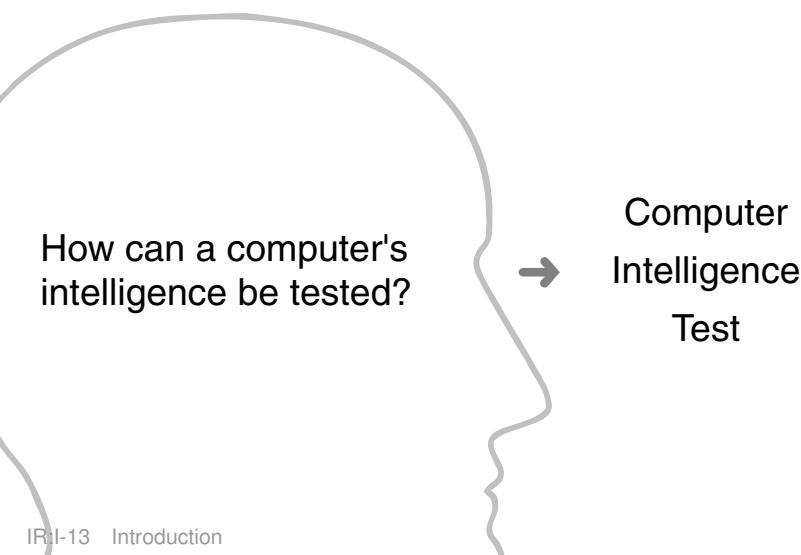
- Billions of documents.  
Text, images, audio files, videos.
  - A vague request.  
Expression of a complex information need: 1-5 keywords, or a question.
  - High class imbalance.  
Only a handful of documents are relevant to the request.
- Retrieve the relevant documents in milliseconds.



How can a computer's  
intelligence be tested?

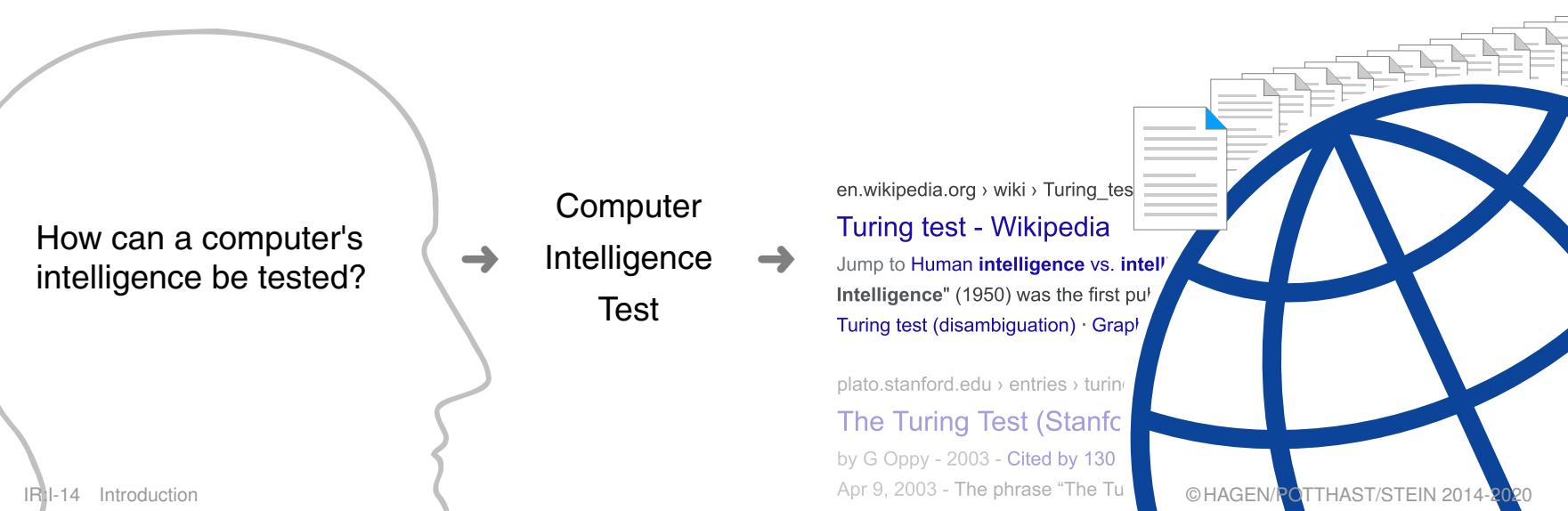
# Information Retrieval in a Nutshell

- Billions of documents.  
Text, images, audio files, videos.
  - A vague request.  
Expression of a complex information need: 1-5 keywords, or a question.
  - High class imbalance.  
Only a handful of documents are relevant to the request.
- Retrieve the relevant documents in milliseconds.



# Information Retrieval in a Nutshell

- ❑ Billions of documents.  
Text, images, audio files, videos.
  - ❑ A vague request.  
Expression of a complex information need: 1-5 keywords, or a question.
  - ❑ High class imbalance.  
Only a handful of documents are relevant to the request.
- Retrieve the relevant documents in milliseconds.



# Chapter IR:I

## I. Introduction

- ❑ Information Retrieval in a Nutshell
- ❑ Examples of Information Retrieval Problems
- ❑ Terminology
- ❑ Delineation
- ❑ Historical Background

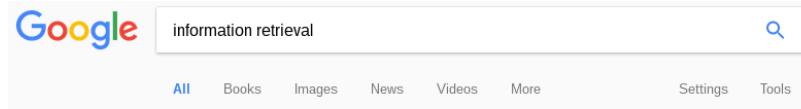
# Examples of Information Retrieval Problems

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

# Examples of Information Retrieval Problems

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

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



About 15,100,000 results (0.38 seconds)

## Information retrieval - Wikipedia

[https://en.wikipedia.org/wiki/Information\\_retrieval](https://en.wikipedia.org/wiki/Information_retrieval) ▾

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing.

[Overview](#) · [History](#) · [Model types](#) · [Performance](#) and ...

## [PDF] Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/pdf/01book.pdf> ▾

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

## Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/> ▾

The book aims to provide a modern approach to information retrieval from a computer science perspective. It is based on a course we have been teaching in ...

[Introduction to Information](#) ... · [Information Retrieval and Web](#) ... · [Boolean retrieval](#)

## Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html> ▾

Introduction to Information Retrieval. By Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze. Website: <http://informationretrieval.org/>. Cambridge ...

## Information Retrieval and Web Search: CS 276

<https://cs276.stanford.edu/> ▾

Information retrieval is the process through which a computer system can respond to a user's query for text-based information on a specific topic. IR was one of ...

## [PDF] Introduction to Information Retrieval - Stanford NLP Group

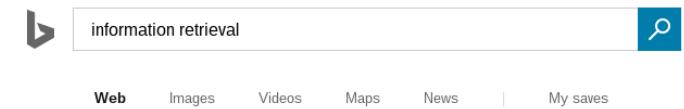
<https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

Aug 1, 2006 - Information. Retrieval. Christopher D. Manning. Prabhakar Raghavan. Hinrich Schütze. Cambridge University Press. Cambridge, England ...

## Information Retrieval Journal - Springer

<https://link.springer.com/journal/10791>

The journal provides an international forum for the publication of theory, algorithms, and experiments across the broad area of information retrieval. Topics of ...



67,200,000 RESULTS

Any time ▼

## Information retrieval - Wikipedia

[https://en.wikipedia.org/wiki/Information\\_retrieval](https://en.wikipedia.org/wiki/Information_retrieval) ▾

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources.

## Introduction to Information Retrieval

<http://nlp.stanford.edu/IR-book/> ▾

Introduction to Information Retrieval. This is the companion website for the following book. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze ...

[Bib](#) · [Errata](#)

## Information Retrieval | Definition of Information ...

[https://www.merriam-webster.com/dictionary/information\\_retrieval](https://www.merriam-webster.com/dictionary/information_retrieval) ▾

Define **information retrieval**: the techniques of storing and recovering and often disseminating recorded data especially through the use of a...

## Information Retrieval | Article about Information ...

<http://encyclopedia2.thefreedictionary.com/information+retrieval> ▾

**information retrieval**[in-fər'mā-shən rē-tē-vəl] (computer science) The technique and process of searching, recovering, and interpreting **information** ...

## CS 276: Information Retrieval and Web Search

<https://web.stanford.edu/class/cs276/> ▾

Information retrieval is the process through which a computer system can respond to a user's query for text-based **information** on a specific topic.

## [PDF] Information Retrieval - Stanford University

<https://web.stanford.edu/class/cs276/handouts/lecture2-dictionary.pdf>

3 Introduction to Information Retrieval Introduction to Information Retrieval Terms The things indexed in an IR system Introduction to Information Retrieval

## Information retrieval - Britannica.com

<https://www.britannica.com/topic/information-retrieval> ▾

**information retrieval**: Recovery of **information**, especially in a database stored in a computer. Two main approaches are matching words in the query against the ...

## 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 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 snippet of Bing’s 6th result is flawed.

# Examples of Information Retrieval Problems

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

# Examples of Information Retrieval Problems

Plan a trip from San Francisco to Paris, France.

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

The screenshot shows a search results page with the query "flight sf paris". The results are categorized under "Web". There are two main sections of results:

- Flights from Paris - Fly from Paris at less**  
Ad · [www.CheapOair.com/Paris](http://www.CheapOair.com/Paris)  
Fly from Paris at less! Grab the Travel Deals & Save Maximum.  
Cheap Price Worldwide · Millions of Cheap Flights · 24/7 Customer Care
- Flights to Flights Paris - eDreams.com**  
Ad · [www.eDreams.com/Flights-Paris](http://www.eDreams.com/Flights-Paris)  
Book Today Cheap Flights to Flight Paris. Book Now!  
Cheap Flights from £19 · More than 750 Airlines · Mobile Friendly · Service 7/7  
Save on Flight+Hotel Deal      Flights to London 75% Off  
Best Offers to/from Paris      Cheap City Breaks Europe

Showing results for **flights paris**.  
No results found for flight sf paris.

Below the results is a large image of the Eiffel Tower and the Paris skyline. At the bottom, there is a form for searching flights from CMF - Chambery - Chambery to CDG - Paris - Paris Charles De Gaulle, departing on Sun, Aug 27 and arriving on Tue, Sep 19. A message at the bottom says: "We couldn't find any flights. Try changing your dates or airports."

The screenshot shows a search results page on Yandex with the query "paris hilton". The results are categorized under "WEB". The top result is an advertisement for Hilton Paris Opera 4\*.

**Hilton Paris Opera 4\* – Онлайн бронирование номеров**  
[hotelhunter.com](http://hotelhunter.com) ad  
Эксклюзивные тарифы напрямую со скидкой до 70%  
Онлайн бронирование · Бесплатный Wi-Fi · Фото и отзывы · Специальные скидки

**Official website**  
[parishilton.com](http://parishilton.com)

**Paris Hilton - Wikipedia**  
[en.wikipedia.org](http://en.wikipedia.org) · Paris Hilton +  
Paris Whitney Hilton (born February 17, 1981) is an American businesswoman, socialite, television and media personality, model, actress, singer, and DJ.

**paris hilton – browse images**  
[yandex.com/images](http://yandex.com/images) > paris hilton

Complain

Four small images of Paris Hilton are shown.

**Paris Hilton - YouTube**  
[youtube.com](http://youtube.com) > user/ParisHilton +  
Paris Hilton - Official YouTube Stars Are Blind - Duration: 4 minutes, 33 seconds... Paris Hilton - Good Time (Explicit) ft. Lil Wayne - Duration: 3 minutes, 47 seconds.

**Paris Hilton - IMDb**  
[imdb.com](http://imdb.com) > name/nm0385296 +  
Socialite Paris Whitney Hilton was born on February 17, 1981 in New York City, into the Hilton family, and has three younger siblings, Nicky Hilton Rothschild...

**Paris Hilton (@ParisHilton) | Твиттер**  
[twitter.com](http://twitter.com) > parishilton +  
Paris Hilton нача(а) читать. .... that time we sent @irin to mexico city with @ParisHilton and all my @marieclaire editor dreams came truehttp...  
Friends 15 million subscribers, 6 thousand friends

**Paris Hilton - Google+**  
[plus.google.com](http://plus.google.com) > ParisHilton +  
Mike Kerwin: I do really like cats +Paris Hilton, I was just sitting her thinking my way through petting your cat and using that as an excuse to . . .

## 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 Information Retrieval Problems

*What were the news of the day?*

# Examples of Information Retrieval Problems

*What were the news of the day?*

Hit the news feeds.

Google News    Search

Headlines Local For You U.S. ▾

### Top Stories

**Bannon's departure is unlikely to calm the turmoil in Trump's White House**  
Washington Post · 9h ago



RELATED COVERAGE  
Steve Bannon, Unrepentant  
Highly Cited · The American Prospect · Aug 16, 2017

**IS conflict: Iraq launches ground offensive in Tal Afar**  
BBC News · 4h ago



RELATED COVERAGE  
Iraqi forces to commence Tal Afar operation 'in the next few days'  
Local Source · Rudaw · 14h ago

**Protesters Flood Streets, and Trump Offers a Measure of Praise**  
New York Times · 6h ago



RELATED COVERAGE  
Protesters face a tricky balance on free speech  
Local Source · The Boston Globe · 11h ago

**Researchers find wreckage of famed Navy cruiser Indianapolis, sunk in 1945**  
Los Angeles Times · 3h ago



RELATED COVERAGE  
Wreckage From USS Indianapolis Located In Philippine Sea | Paul Allen  
Most Referenced · Paul Allen · 3h ago

#news    #news

Top Latest People Photos Videos News Broadcasts

**KRTpro News @KRTpro\_News · 1m**  
#BREXIT  
UK to release tranche of Brexit position papers reut.rs/2wl5my  
#KRTpro #News



**UK to release tranche of Brexit position papers**  
Britain will issue a cluster of new papers this week to outline its strategy positions in divorce talks with the European Union, ranging from regulation reuters.com

**Breaking News India @Golstream · 3m**  
Boston March Against Hate Speech Avoids Charlottesville Chaos - NDTV if.tt/2wltLnv #India #News



# Examples of Information Retrieval Problems

*What were the news of the day?*

Hit the news feeds.

Google News    Search

Headlines Local For You U.S. ▾

### Top Stories

 **Bannon's departure is unlikely to calm the turmoil in Trump's White House**  
Washington Post · 9h ago

RELATED COVERAGE  
Steve Bannon, Unrepentant  
Highly Cited · The American Prospect · Aug 16, 2017

 **IS conflict: Iraq launches ground offensive in Tal Afar**  
BBC News · 4h ago

RELATED COVERAGE  
Iraqi forces to commence Tal Afar operation 'in the next few days'  
Local Source · Rudaw · 14h ago

 **Protesters Flood Streets, and Trump Offers a Measure of Praise**  
New York Times · 6h ago

RELATED COVERAGE  
Protesters face a tricky balance on free speech  
Local Source · The Boston Globe · 11h ago

 **Researchers find wreckage of famed Navy cruiser Indianapolis, sunk in 1945**  
Los Angeles Times · 3h ago

RELATED COVERAGE  
Wreckage From USS Indianapolis Located In Philippine Sea | Paul Allen  
Most Referenced · Paul Allen · 3h ago

LIBERAL i SHOWING POSTS ABOUT: "HEALTH CARE"

 Her emails didn't take away your healthcare.  
Trump and his republican cronies aren't going to stop trying to take away your healthcare and basic human rights. Sign here and vow to fight back => <http://bit.ly/Join-LA>

1 801 49 500

DAILY KOS Daily Kos 12 hours ago

Facts. Pass them on.

 Obamacare still isn't collapsing  
The major concern for the 2018 enrollment period in the Affordable Care...  
DAILYKOS.COM

1 393 14 105

CONSERVATIVE i

 \$ 1.3 Billion Health Care Fraud Just Erupts! Jef...  
Attorney General Jeff Sessions said at a news conference. "We will cont...  
USAPOLITICSTODAY.COM

1.3K 72 558

Allen West about a week ago

Sing it with me: And another one gone, and another one gone. Another one bites the dust.

 Implosion: ANOTHER major healthcare compa...  
Epic FAIL  
ALLENBWEST.COM | BY ALLEN WEST

2.3K 227 1K

Breitbart about a week ago

Sinking like a stone...

CANADIAN MSNBC ANCHOR VS. REPUBLICAN CONGRESSMAN

## 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 Information Retrieval Problems

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

# Examples of Information Retrieval Problems

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

Search for facts

Google search results for "how high can kangaroos jump":

All Shopping Videos Images News More Settings Tools

About 6,340,000 results (0,32 seconds)

Kangaroos Can Jump 30 Feet High. Mar 10, 2014 

Kangaroos Can Jump 30 Feet High - YouTube <https://www.youtube.com/watch?v=1L0YNsVZYNQ>

How can kangaroos jump so high? How high can they jump - Quora <https://www.quora.com/How-can-kangaroos-jump-so-high-How-high-can-they-jump>

Secret of kangaroo's bounce - The Telegraph [www.telegraph.co.uk/News/Science/Science\\_News/](http://www.telegraph.co.uk/News/Science/Science_News/)

How and why do kangaroos hop? | Discover Wildlife [www.discoverwildlife.com/animals/mammals/how-and-why-do-kangaroos-hop/](http://www.discoverwildlife.com/animals/mammals/how-and-why-do-kangaroos-hop/)

How far can a Kangaroo Jump? - AnimalWised [https://www.animalwised.com/Fun\\_facts/Facts\\_about\\_the\\_animal\\_kingdom.html](http://www.animalwised.com/Fun_facts/Facts_about_the_animal_kingdom.html)

Red Kangaroo | National Geographic [www.nationalgeographic.com/animals/mammals/r/red-kangaroo/](http://www.nationalgeographic.com/animals/mammals/r/red-kangaroo/)

How high can a kangaroo jump? | Reference.com [https://www.reference.com/Pets-&Animals/Mammals/Marsupials.html](http://www.reference.com/Pets-&Animals/Mammals/Marsupials.html)

WolframAlpha computational knowledge engine.

height of empire state building in feet

Web Apps Examples Random

Input interpretation: convert Empire State Building total height to feet

Result: 1250 feet

Show details

Open code

Additional conversions:

- 0.2367 miles
- 417 yards
- 0.2057 nmi (nautical miles)
- 381 meters
- 0.381 km (kilometers)

Comparisons as height:

- $\approx 0.69 \times$  height of the CN Tower ( $\approx 553$  m)
- $\approx 0.7 \times$  architectural height of One World Trade Center (1776 ft)
- $\approx 1.2 \times$  Eiffel Tower height ( $\approx 324$  m)

Corresponding quantity:

Distance to horizon (ignoring topography and other obstructions):

- 70 km (kilometers)
- 69 716 meters
- 43 miles

Sources Download page

POWERED BY THE WOLFRAM LANGUAGE

# Examples of Information Retrieval Problems

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

Search for facts

Google  🔍

🔍 All 🔗 Shopping 📹 Videos 📰 News 🖼 Images ⋮ More Settings Tools

About 4,380,000 results (0.65 seconds)

**6 feet high**

Red **kangaroos** hop along on their powerful hind legs and **do** so at great speed. A red kangaroo **can** reach speeds of over 35 miles an hour. Their bounding gait allows them to cover 25 feet in a single leap and to **jump 6 feet high**.

[www.nationalgeographic.com/animals/mammals/red-kangaroo](http://www.nationalgeographic.com/animals/mammals/red-kangaroo) ▾

**Red Kangaroo | National Geographic**

ⓘ About Featured Snippets  ⓘ Feedback

[www.animalwised.com/Fun facts/Facts about the animal kingdom](http://www.animalwised.com/fun-facts-facts-about-the-animal-kingdom) ▾

**How far can a Kangaroo Jump? - AnimalWised**

Jump to **Do you want to know more about Kangaroos?** - How long and how high can kangaroos jump? Do you want to know more about Kangaroos ...

[www.discoverwildlife.com/Animal Facts/Mammals](http://www.discoverwildlife.com/animal-facts/mammals) ▾

**How and why do kangaroos hop? - Discover Wildlife**

How and why do kangaroos hop? BBC Wildlife contributor Ben Phillips answers your wild question.

[www.quora.com/How-can-kangaroos-jump-so-high-How-high-can-they-jump](http://www.quora.com/How-can-kangaroos-jump-so-high-How-high-can-they-jump) ▾

**How can kangaroos jump so high? How high can they jump - Quora**

1 answer

May 25, 2016 - According to Wiki (Red kangaroo) a male Red kangaroo **can** jump up to 3 meters in the air (around 10 ft) though I've seen claims of much **higher** records.

**How high can a kangaroo jump on average?** Mar 10, 2016

An adult kangaroo **can** jump as **high** as 1.93m with what initial ... Dec 6, 2016

**Why can't kangaroos jump backwards?** Jan 8, 2016

**How far can a kangaroo jump?** Jul 11, 2019

More results from [www.quora.com](http://www.quora.com)

**WolframAlpha** computational knowledge engine.

⭐ ✉

ⓘ Web Apps  ⓘ Examples  ⓘ Random

**Input interpretation:**

convert Empire State Building total height to feet

ⓘ Open code

**Result:**

1250 feet

ⓘ Show details

**Additional conversions:**

0.2367 miles

417 yards

0.2057 nmi (nautical miles)

381 meters

0.381 km (kilometers)

**Comparisons as height:**

$\approx 0.69 \times$  height of the CN Tower ( $\approx 553$  m)

$\approx 0.7 \times$  architectural height of One World Trade Center (1776 ft)

$\approx 1.2 \times$  Eiffel Tower height ( $\approx 324$  m)

**Corresponding quantity:**

Distance to horizon (ignoring topography and other obstructions):

70 km (kilometers)

69 716 meters

43 miles

ⓘ Sources  ⓘ Download page

POWERED BY THE WOLFRAM LANGUAGE

# Examples of Information Retrieval Problems

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

Search for facts, or ask the question outright.

Google search results for "how high can kangaroos jump":

Search term: how high can kangaroos jump

Results:

- 6 feet high**  
Red **kangaroos** hop along on their powerful hind legs and **do** so at great speed. A red kangaroo **can** reach speeds of over 35 miles an hour. Their bounding gait allows them to cover 25 feet in a single leap and to **jump** 6 feet **high**.  
  
www.animalwised.com
- How far can a Kangaroo Jump? - AnimalWised**  
Jump to **Do** you want to know more about **Kangaroos?** - How long and how high can kangaroos jump? Do you want to know more about Kangaroos ...  
www.animalwised.com
- How and why do kangaroos hop? - Discover Wildlife**  
How and why do kangaroos hop? BBC Wildlife contributor Ben Phillips answers your wild question.  
www.discoverwildlife.com
- How can kangaroos jump so high? How high can they jump - Quora**  
1 answer  
May 25, 2016 - According to Wiki (Red kangaroo) a male Red kangaroo **can** jump up to 3 meters in the air (around 10 ft) though I've seen claims of much **higher** records.  
**How high can a kangaroo jump on average?** Mar 10, 2016  
An adult kangaroo **can** jump as **high** as 1.93m with what initial ... Dec 6, 2016  
**Why can't kangaroos jump backwards?** Jan 8, 2016  
**How far can a kangaroo jump?** Jul 11, 2019  
More results from www.quora.com

Google search results for "can kangaroos jump higher than the empire state building":

Search term: can kangaroos jump higher than the empire state building

Results:

- Can a kangaroo jump higher than empire state building?**  
A healthy kangaroo **can** jump higher than any rabbit. Kangaroos also jump faster and further than rabbits. This is due, in part, to their larger size.  
[answers.com/Q/Can\\_a\\_kangaroo\\_jump\\_higher\\_than\\_empire...](https://answers.com/Q/Can_a_kangaroo_jump_higher_than_empire...)
- Can a kangaroo jump higher than the Empire State Building ...**  
The Empire State Building **can't** jump. ... Can a kangaroo jump higher than the Empire State Building? ... but it can jump higher than the World Trade Center" ...  
[https://www.reddit.com/r/Jokes/comments/55ndr2/can\\_a\\_kangaroo\\_ju...](https://www.reddit.com/r/Jokes/comments/55ndr2/can_a_kangaroo_ju...)
- The Big Apple: "Can a kangaroo jump higher than the Empire ...**  
"What animal can jump higher than the Empire State building?" was cited in 1939. "What animal has eyes and can't see, legs and can't walk, ...  
[barrypopik.com/index.php/new\\_york\\_city/entry/can\\_a\\_kangaroo...](https://barrypopik.com/index.php/new_york_city/entry/can_a_kangaroo...)
- Can a kangaroo jump higher than the Empire State Building ...**  
The Empire State Building **can't** jump. Jump to content. my subreddits. edit subscriptions. popular-all ... Can a kangaroo jump higher than the Empire State Building?  
[https://www.reddit.com/r/cleanjokes/comments/6lzo2n/can\\_a\\_kangar...](https://www.reddit.com/r/cleanjokes/comments/6lzo2n/can_a_kangar...)
- Can a Kangaroo Jump Higher than the Empire State Building ...**  
Can a Kangaroo Jump Higher than the Empire State Building? ... Kangaroos Can Jump 30 Feet High - Duration: ... guy jumping off Empire State Building ...  
[youtube.com/watch?v=7M-a\\_s\\_pGW4](https://youtube.com/watch?v=7M-a_s_pGW4)
- Can A Kangaroo Jump Higher Than The Empire State Building ...**  
Can A Kangaroo Jump Higher Than The Empire State Building.  
[increaseverticaljump.blogspot.com/download/can-a-kangaroo-jump-higher-than-...](https://increaseverticaljump.blogspot.com/download/can-a-kangaroo-jump-higher-than-...)
- Question: Can a kangaroo jump higher than the Empire State ...**  
Best Answer: Yes, but only on Tuesdays. ... Not at all, My Dear. Empire State Building can't jump. But this is very old matter. ... Yes. Buildings ...  
<https://answers.yahoo.com/question/index?qid=20101223195232AAgjjQ>

## Remarks:

- ❑ Users search for facts; search engines employ various strategies to meet these requests, using knowledge bases like Wikidata, or extracting factual information from web pages.
- ❑ 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. Meanwhile, Google fixed the issue.
- ❑ 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.
- ❑ Search engines lack common sense:

The screenshot shows a Google search results page. At the top, the search bar contains the query "how many legs does a fish have". Below the search bar, there are navigation links for All, News, Images, Shopping, Videos, More, Settings, and Tools. A status message indicates "About 341,000,000 results (0.60 seconds)". The main content area features a large, bold, blue-highlighted snippet with the text "4 Legs". Below this, a smaller snippet reads "Fishes do have 4 Legs, do they? Sep 10, 2018". Further down, there is a link to a YouTube video titled "How Many Legs Does a Fish Have? | STRIVIA | Pulse Live ...". At the bottom of the page, there are links for "About Featured Snippets" and "Feedback".

# Examples of Information Retrieval Problems

*Build a fence.*

# Examples of Information Retrieval Problems

*Build a fence.*

Search for tutorials.

Search results 1-10 for *how to build a fence*

Total results: 2037230 (retrieved in 1025.3ms)

**how to build a fence**  
fence-posts.org/tag/how-to-build-a-fence ▾  
Hi, I'm Alex Barnett and I'd like to welcome you to my web site, **How To Build A Fence**. All right, I know, I've heard all of the jokes before. Why on earth would I want to build a web site about fence building? Is there anything more boring? Well, the truth is that building a good, sturdy, long

**How To Build a Wooden Fence**  
fence-posts.org/how-to-build-a-w... ▾

More results from fence-posts.org

**How to build a fence like a pro**  
www.how-to-build-a-fence-like-a-pro.com/ ▾

you will need to build your fence. We will answer frequently asked questions and provide a photo gallery of pictures for your enjoyment. Take your time in planning out your fence. Have fun Make it an event with friends and family members. © Copyright 2008, Trigon Corporation, All Rights Reserved

**How To Build A Fence**  
siteexpansion.com/tag/how-to-build-a-fence ▾

Gardening information structured to support backyard garden themes. Provide seasonal gardening information and largest garden store on the Web. Finally a single source for the backyard gardener. Tags: a **fence** , annual flowers , bedtime stories , bedtime stories ringtone , bulbs Do it yourself

**How to Build a Fence with Goat Panels**  
feedlotpanels.com/how-to-build-a-fence-with-goat-panels ▾

article, you will learn how to build a fence using goat panels. The first step of building a goat fence includes finding out how large you want your goat enclosure or goat pen to be. Each goat panel is about 16 feet long and 48 inches tall. Consequently, if you want a goat fence with an area of 16

**How to Build a Fence, Garden Fencing**  
www.beestonfencingcompany.co.uk/howtobuildafence.htm ▾

You may be wondering how to erect fencing without digging holes or mixing concrete. Here are some quick and easy solutions for building a fence with timber panels and fence posts. To erect a fence on grass or soil, drive metal spikes into the ground with a sledge hammer and insert the upright fence

**How To Build A Jackleg Fence**  
www.mademan.com/mm/how-build-jackleg-fence.html ▾

If you are looking for a relatively inexpensive fencing option, you may want to explore how to build a jackleg fence. A jackleg fence is not an option for keeping small critters or children in or out; however, it is the perfect choice for livestock or just as a property divider. A jackleg fence has

Search results for *how to build a fence*

About 1,320,000 results

**How to Build a Fence | Mitre 10 Easy As**  
Mitre 10 New Zealand  
5 years ago • 40,654 views  
Knowing how to build a fence is one of the more useful DIY skills to learn. And if you learn how to do it properly, your DIY fence ...

**Building a Picket Fence | Setting the Posts**  
The Restoration Couple  
1 year ago • 181,300 views  
Part 1 - Here is a look at our picket fence installation around the vegetable garden. Enjoy! CONTACT US ...

**What to Consider When Building a New Fence**  
The Home Depot  
8 years ago • 1,236,065 views  
Find our full guide to learn how to build a fence:  
http://thd.co/2ozT1W3j Shop all the products you'll need at The

**93 Building 70 Feet of Wooden Fence**  
Memphis Applegate  
6 months ago • 92,372 views  
I'm using the cool South Carolina "winter" to finish the wooden fence around my backyard. In this edition I install about 70 feet

**DIY: How To Build A Fence (BYOT #12)**  
BYOT  
1 year ago • 109,604 views  
This DIY project is all about how to build a fence. With the right tools you can turn an ugly fence into a beautiful new cedar fence ...

**Building a Board on Board Cedar Fence - Part 2**  
April Wilkerson  
1 year ago • 657,037 views  
If you missed Part 1, here is a link: https://youtu.be/v3BlyCf1YPM  
Looking for Part 3? Here ya go: ...

## 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 Information Retrieval Problems

*Write an essay on video surveillance.*

# Examples of Information Retrieval Problems

*Write an essay on video surveillance.*

Search for other's opinions on video surveillance.

video surveillance

About 443,000,000 results (0.48 seconds)

en.wikipedia.org › wiki › Category:Video\_surveillance ▾

**Category:Video surveillance - Wikipedia**

Pages in category "Video surveillance". The following 29 pages are in this category, out of 29 total. This list may not reflect recent changes (learn more).

www.ifsecglobal.com › video-surveillance ▾

**Video Surveillance and CCTV - IFSEC Global**

Video Surveillance or CCTV (closed circuit television) represents the largest segment of Security technology. Video cameras are used to observe an area, ...

www.amazon.com › Video-Surveillance ▾

**Video Surveillance: Electronics: Surveillance ... - Amazon.com**

YI 1080p Home Camera, Indoor IP Security Surveillance System with Night Vision for Home / Office / Nanny / Pet Monitor with iOS, Android App, Cloud Service Available - Works with Alexa... Ring Floodlight Camera Motion-Activated HD Security Cam Two-Way Talk and Siren Alarm, White.

www.pelco.com ▾

**Pelco Security Cameras and Surveillance Systems**

Pelco offers industry's best security cameras, CCTV, and video surveillance systems designed for exceptional performance in the indoor and outdoor ...

People also ask

What is the best video surveillance system?

How long do stores keep video surveillance?

What is surveillance footage?

How much does home surveillance cost?

Feedback

www.backstreet-surveillance.com ▾

**Security Cameras, HD Camera, Video Surveillance Systems ...**

Backstreet Surveillance offers the best HD security cameras and video surveillance systems in the market. Our DIY HD Surveillance Camera systems are perfect ...

video surveillance

Page 1 of 40 arguments (retrieved in 21.1ms) [Pro vs. Con View](#) | [Overall Ranking View](#)

**[con] Often surveillance camera images are not clear and police...**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
Often surveillance camera images are not clear and police cannot identify the criminal. ... ▾

**[con] Surveillance cameras cannot physically protect the public,...**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
Surveillance cameras cannot physically protect the public, only film what is happening. ... ▾

**[pro] Surveillance cameras are not closely monitored and are only...**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
Surveillance cameras are not closely monitored and are only usually viewed if a crime has taken place. ... ▾

**[con] Crime camera evidence is very rarely used in court cases....**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
Crime camera evidence is very rarely used in court cases. ... ▾

**[pro] There is not much privacy in public places....**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
There is not much privacy in public places. ... ▾

**[con] Filming without consent is actually illegal....**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
Filming without consent is actually illegal. ... ▾

**[pro] It is no different to police monitoring a dangerous area....**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
It is no different to police monitoring a dangerous area. ... ▾

**[pro] Crime cameras offer conclusive, unbiased evidence in court....**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
Crime cameras offer conclusive, unbiased evidence in court. ... ▾

**[pro] Crime cameras help catch criminals and get them off the...**

[http://www.debatepedia.org/en/index.php/Debate:\\_Video\\_surveillance](http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance)  
Crime cameras help catch criminals and get them off the street ... ▾

## 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 Information Retrieval Problems

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



# Examples of Information Retrieval Problems

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

The image is an example of the information sought after.



Google image.jpg new york city

All Images Maps Shopping More Settings Tools

About 25,270,000,000 results (0.90 seconds)



Image size:  
756 × 1014

Find other sizes of this image:  
[All sizes](#) - [Medium](#) - [Large](#)

Best guess for this image: [new york city](#)

**The Official Guide to New York City | nycgo.com**

<https://www.nycgo.com/> ▾  
Find out what to do, where to go, where to stay and what to eat in NYC from the experts who know it best.

**New York City - Wikipedia**

[https://en.wikipedia.org/wiki/New\\_York\\_City](https://en.wikipedia.org/wiki/New_York_City) ▾  
The City of New York, often called New York City or simply New York, is the most populous city in the United States. With an estimated 2017 population of 8,622,698 distributed over a land area of about 302.6 square miles (784 km<sup>2</sup>), New York City is also the most densely populated major city in the United States. Located ...

**Visually similar images**



Report images

**Pages that include matching images**

**Fifth Avenue - Wikipedia**

[https://en.wikipedia.org/wiki/Fifth\\_Avenue](https://en.wikipedia.org/wiki/Fifth_Avenue) ▾  
250 × 335 - Fifth Avenue is a major thoroughfare in the borough of Manhattan in New York City, United States. It stretches from West 143rd Street in Harlem to Washington Square North at Washington Square Park in Greenwich Village. It is considered one of the most expensive and elegant streets in the world.

# Examples of Information 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

## 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"<sup>[14][15]</sup> 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.<sup>[16]</sup> 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.<sup>[17][18]</sup> Mars has two moons, Phobos and Deimos, which are small and irregularly shaped. These may be captured asteroids.<sup>[19][20]</sup> 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.<sup>[21][22][23][24]</sup> Liquid water cannot exist on the surface of Mars due to low atmospheric pressure, which is less than 1% of the Earth's<sup>[25]</sup> except at the lowest elevations for short periods.<sup>[26][27]</sup> The two polar ice caps appear to be made largely of water.<sup>[28][29]</sup> 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<sup>[a]</sup>

**Designations**

<b>Pronunciation</b>	UK English: /ma:z/ US English: /mɔ:r.zə/ ( ⓘ listen)
<b>Adjectives</b>	Martian
<b>Orbital characteristics<sup>[2]</sup></b>	
Epoch	
<b>Aphelion</b>	249 200 000 km (154 800 000 mi; 1.66 AU)
<b>Perihelion</b>	206 700 000 km (128 400 000 mi; 1.382 AU)
<b>Semi-major axis</b>	227 939 200 km (141 634 900 mi; 1.523 679 AU)
<b>Eccentricity</b>	0.0934
<b>Orbital period</b>	686.971 d (1.880 82 yr; 668.5991 sols)
<b>Synodic period</b>	779.96 d (2.1354 yr)
<b>Average orbital</b>	24.007 km/s

# Examples of Information Retrieval Problems

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

The text is an example of the information sought after.



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

Not logged in Talk Contributions Create account Log in

Article

Talk

Read

View source

View history

Search Wikipedia



## 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"<sup>[14][15]</sup> 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.<sup>[16]</sup> 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.<sup>[17][18]</sup> Mars has two moons, Phobos and Deimos, which are small and irregularly shaped. These may be captured asteroids.<sup>[19][20]</sup> 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.<sup>[21][22][23][24]</sup> Liquid water cannot exist on the surface of Mars due to low atmospheric pressure, which is less than 1% of the Earth's<sup>[25]</sup> except at the lowest elevations for short periods.<sup>[26][27]</sup> The two polar ice caps appear to be made largely of water.<sup>[28][29]</sup> The volume of water ice in the south polar ice cap, if melted, would be sufficient to cover



Mars in natural color in 2007<sup>[a]</sup>

### Designations

Pronunciation     UK English: /maɪəz/  
US English: /'mɔ:r.zə/ (listen)

Adjectives     Marian

### Orbital characteristics<sup>[2]</sup>

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 sol)

Synodic period     779.96 d  
(2.1354 yr)

Average orbital     24.007 km/s



New analysis



### Text alignment

Found 6 reused passages.

Detailed comparison of your submitted documents: 6 reused passages, length 600 words, 327 shared words

<https://en.wikipedia.org/wiki/Mars>

Geography of Mars

Although better remembered for mapping the Moon,

Johann Heinrich Mädler and Wilhelm Beer

were the first "areographers". They

began

by establishing

that most of Mars's

surface features were permanent and by

more precisely determining the planet's

rotation period. In 1840, Mädler

combined ten years of observations and

drew the first map of Mars.

Rather than giving names to the various

markings,

Beer and Mädler simply designated them

with letters; Meridian Bay (Sinus

Meridiani) was thus feature "a"

Although better remembered for mapping the Moon,

Johann Heinrich Mädler and Wilhelm Beer

were the first "areographers". They

started off

by establishing once and for all

that most of the

surface features were permanent, and

pinned down Mars'

rotation period. In 1840, Mädler

combined ten years of observations and

drew the first map of Mars ever made.

Rather than giving names to the various

markings they mapped,

Beer and Mädler simply designated them

with letters; Meridian Bay (Sinus

Meridiani) was thus feature "a"

<https://en.wikipedia.org/wiki/Mars>

Geography of Mars

Olympus Mons (Mount Olympus).<sup>[114]</sup> The

surface of Mars as seen from Earth is

divided into two kinds of areas, with

differing albedo. The paler plains

covered with dust and sand rich in

reddish iron oxides were once thought of

as Martian "continents" and given names

like Arabia Terra (land of Arabia) or

The surface of Mars as seen from Earth is

consequently

divided into two kinds of areas, with

differing albedo. The paler plains

covered with dust and sand rich in

reddish iron oxides were once thought of

as Martian "continents" and given names

like Arabia Terra (land of Arabia) or

# Examples of Information Retrieval Problems

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

The text is an example of the information sought after.

## 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](mailto:{ghias,bsmith}@cs.cornell.edu), [logan@ghs.com](mailto:logan@ghs.com), [chamber@engr.sgi.com](mailto: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 melody 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 Information Retrieval Problems

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

The text is an example of the information sought after.

Query By Humming  
Musical Information Retrieval in  
An Audio Database

Asif Ghias      Jonathan Logan      David Chamberlin  
Brian C. Smith  
Cornell University  
[ghias,bsmith](mailto:{ghias,bsmith}@cs.cornell.edu)@cs.cornell.edu, [logan@ghs.com](mailto:logan@ghs.com), [chamber@engr.sgi.com](mailto: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 melody 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 ...  
☆ 99 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 ...  
☆ 99 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 Yamamoto... - 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 ...  
☆ 99 Cited by 240 Related articles All 5 versions »»

**[PDF] A Newapproach To Query By Humming In Music Retrieval.**  
L Lu, H You, H 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 ...  
☆ 99 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 ...  
☆ 99 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 ...  
☆ 99 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 ...  
☆ 99 Cited by 2 Related articles All 3 versions »»

IR:I-42 Introduction

© HAGEN/POTTHAST/STEIN 2014-2020

## 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 Information Retrieval Problems

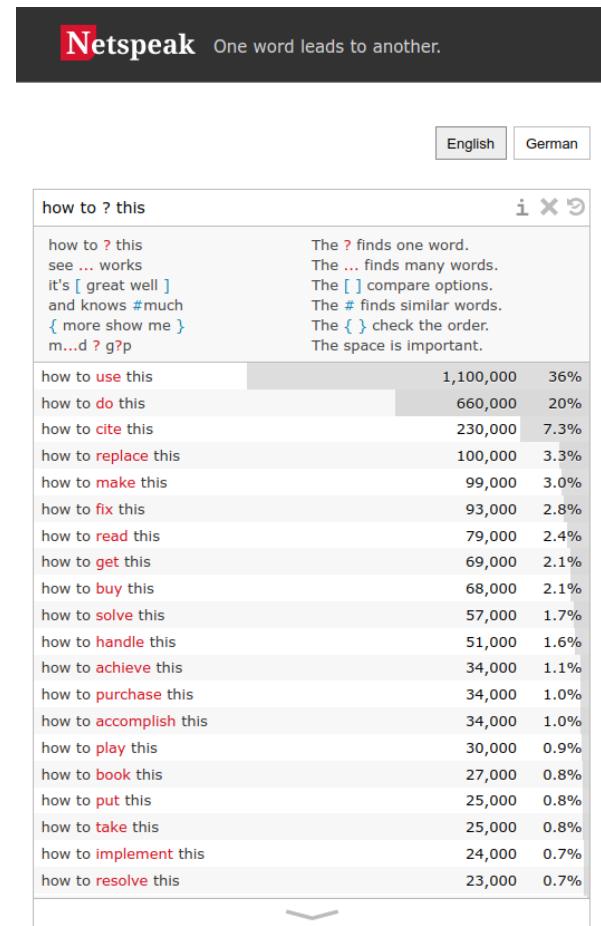
*Find out what people commonly write in the phrase how to ? this.*

# Examples of Information Retrieval Problems

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

Use wildcard search operators to find matching phrases.

The screenshot shows a Google search results page with the query "how to \* this" entered into the search bar. The search results include various links from Stack Exchange, career websites, and general forums, all discussing common ways to phrase questions or instructions. Examples include "How to correctly add a path to PATH? - Unix & Linux Stack Exchange", "How to Answer 'Why Are You Applying for This Position ...'", and "The Galaxy S20 is a great Android phone today, but how long ...". The results are presented in a standard Google search layout with snippets and links.



## 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.

# Chapter IR:I

## I. Introduction

- ❑ Information Retrieval in a Nutshell
- ❑ Examples of Information Retrieval Problems
- ❑ Terminology
- ❑ Delineation
- ❑ Historical Background

# 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 are useful.

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

## Definition 3 (Knowledge)

Justified true beliefs formed on the basis of perceptions, introspection, memory, reason, or testimony.

## 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.
- ❑ The term “epistemology” comes from the Greek words “episteme” and “logos”. “Episteme” can be translated as “knowledge” or “understanding” or “acquaintance”, while “logos” can be translated as “account” or “argument” or “reason”. Just as each of these different translations captures some facet of the meaning of these Greek terms, so too does each translation capture a different facet of epistemology itself. Epistemology seeks to understand one or another kind of cognitive success (or, correspondingly, cognitive failure). Knowledge is among the many kinds of cognitive success that epistemology is interested in understanding.

[\[Stanford Encyclopedia of Philosophy\]](#)

- ❑ Rational thought relies on questioning one’s beliefs and ask for rigorous justification. 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, including hardware, software, operators and 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)

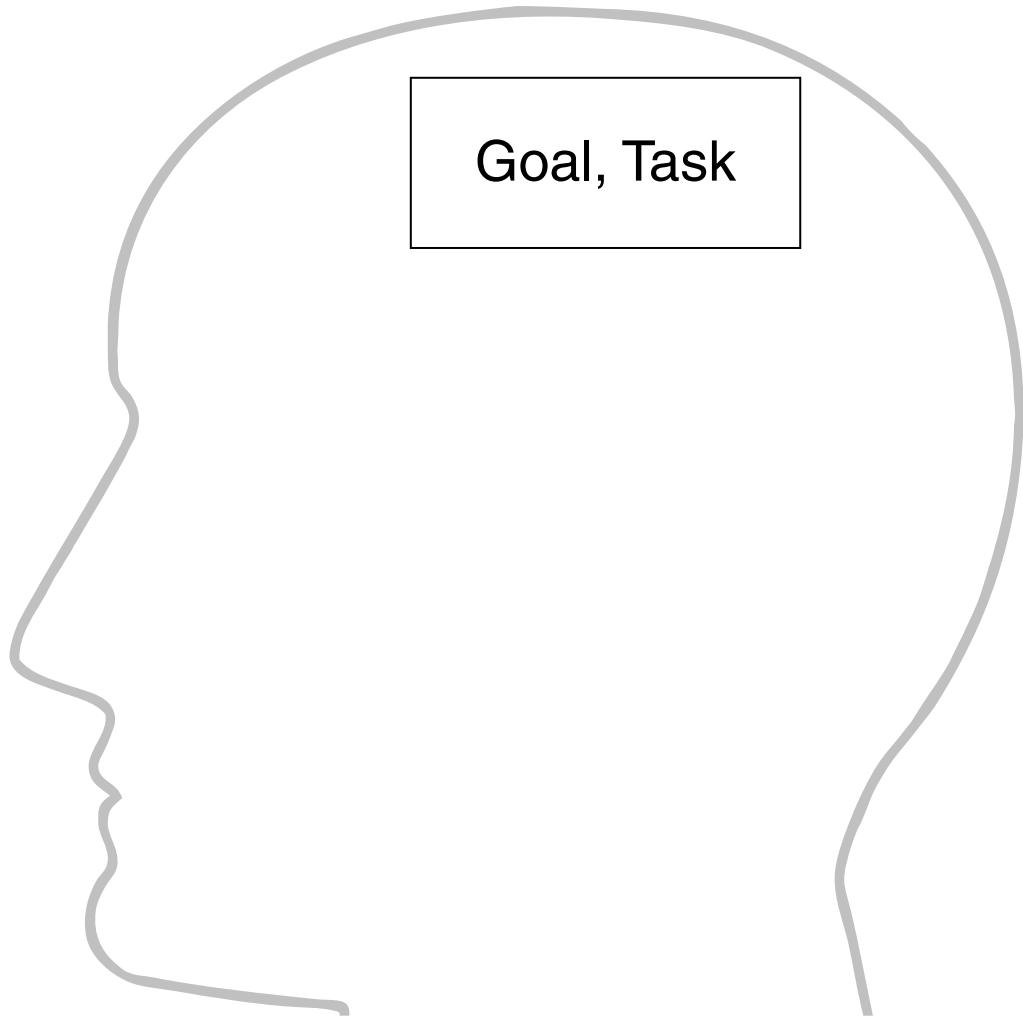
The degree to which a portion of data meets the information need of a user.

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 reach their goal, the more relevant it is.

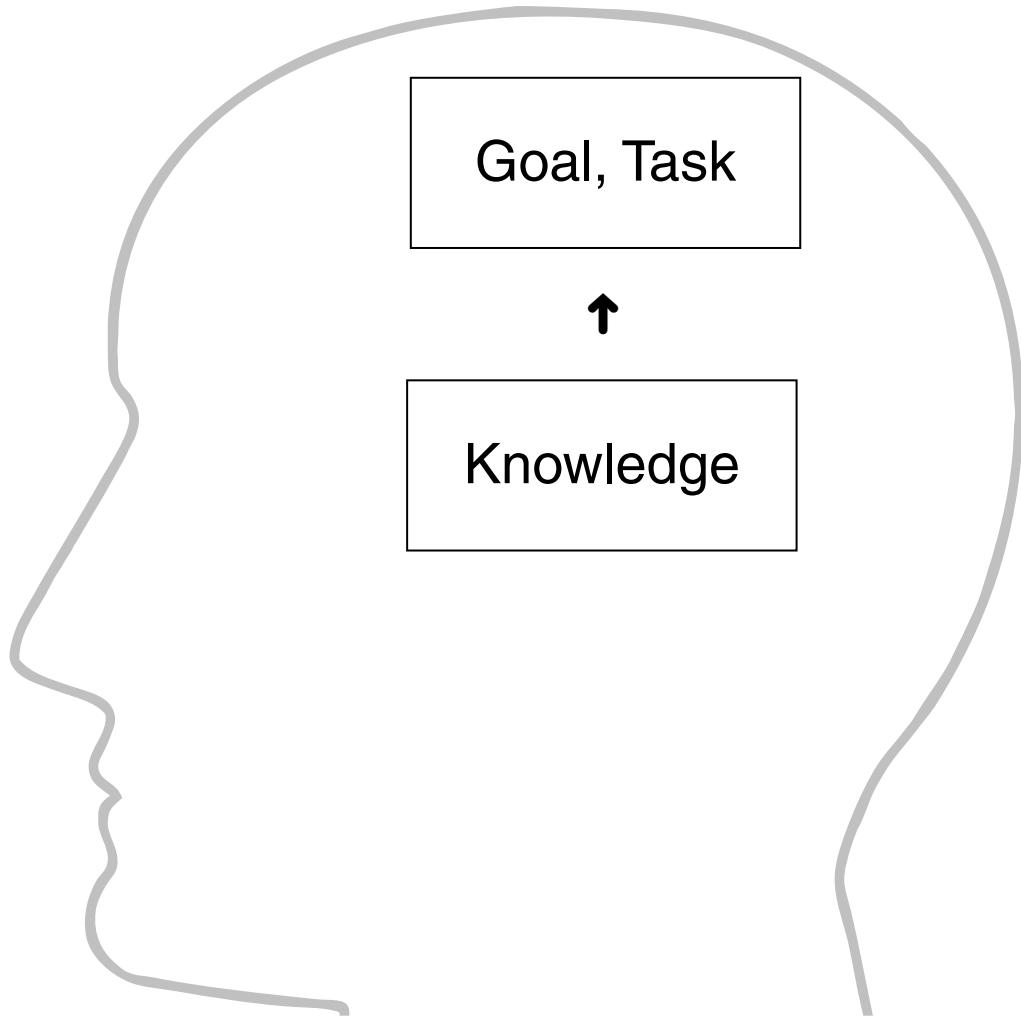
## Remarks:

- ❑ 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.” [\[Wilson 2000\]](#)

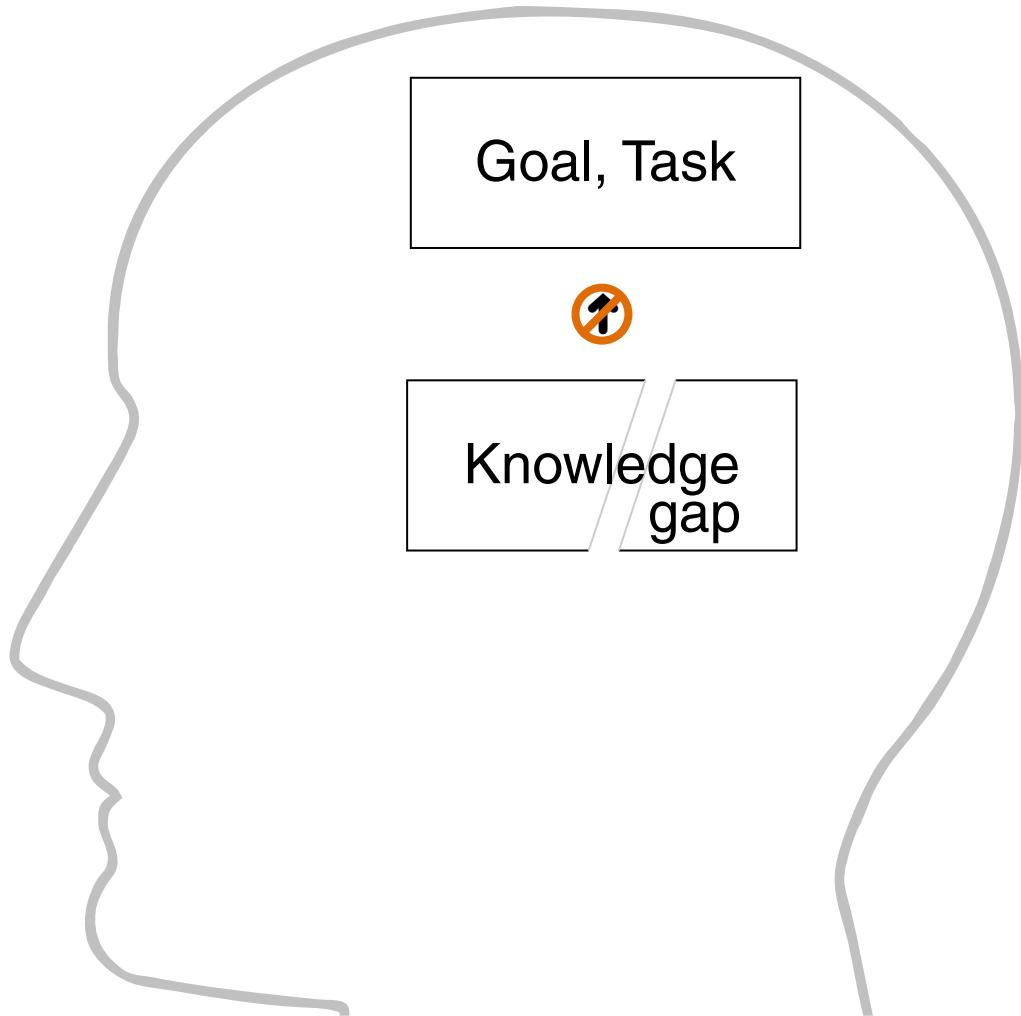
# Terminology



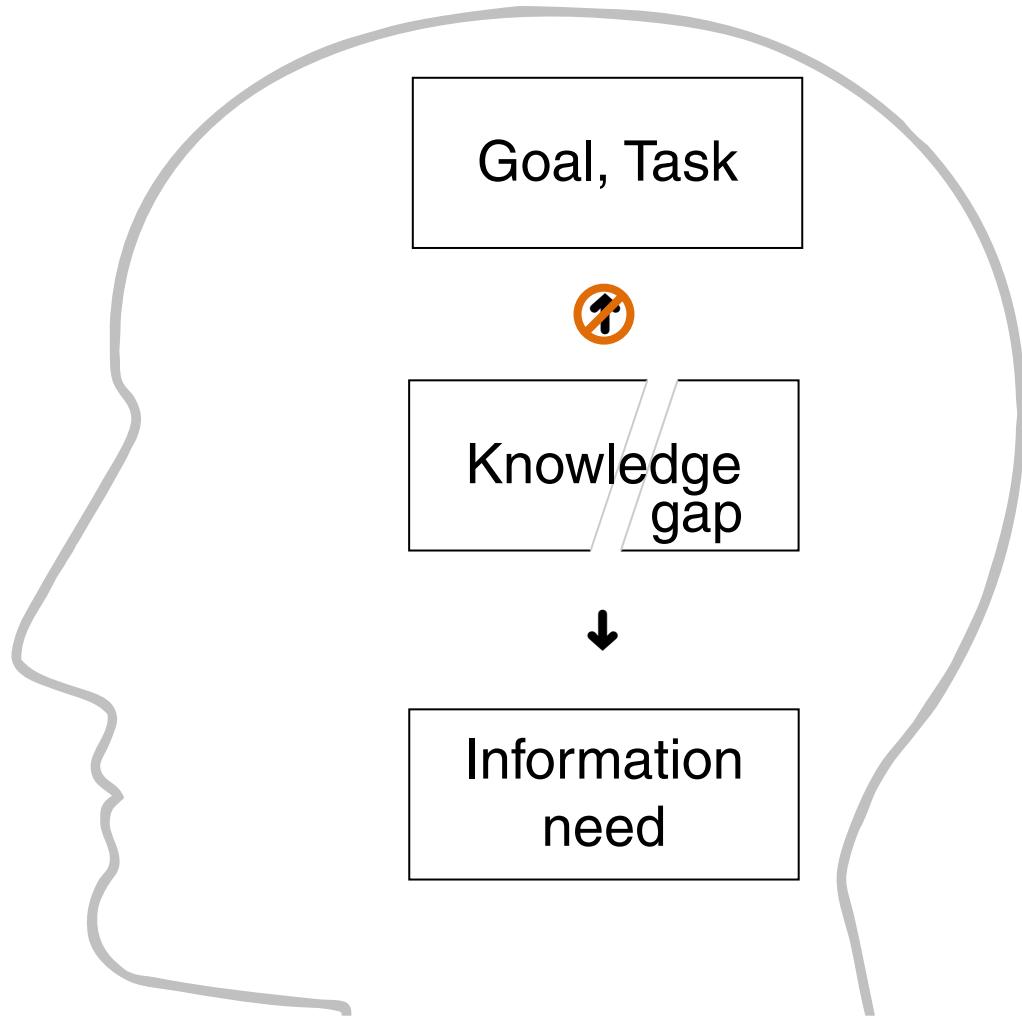
# Terminology



# Terminology



# Terminology

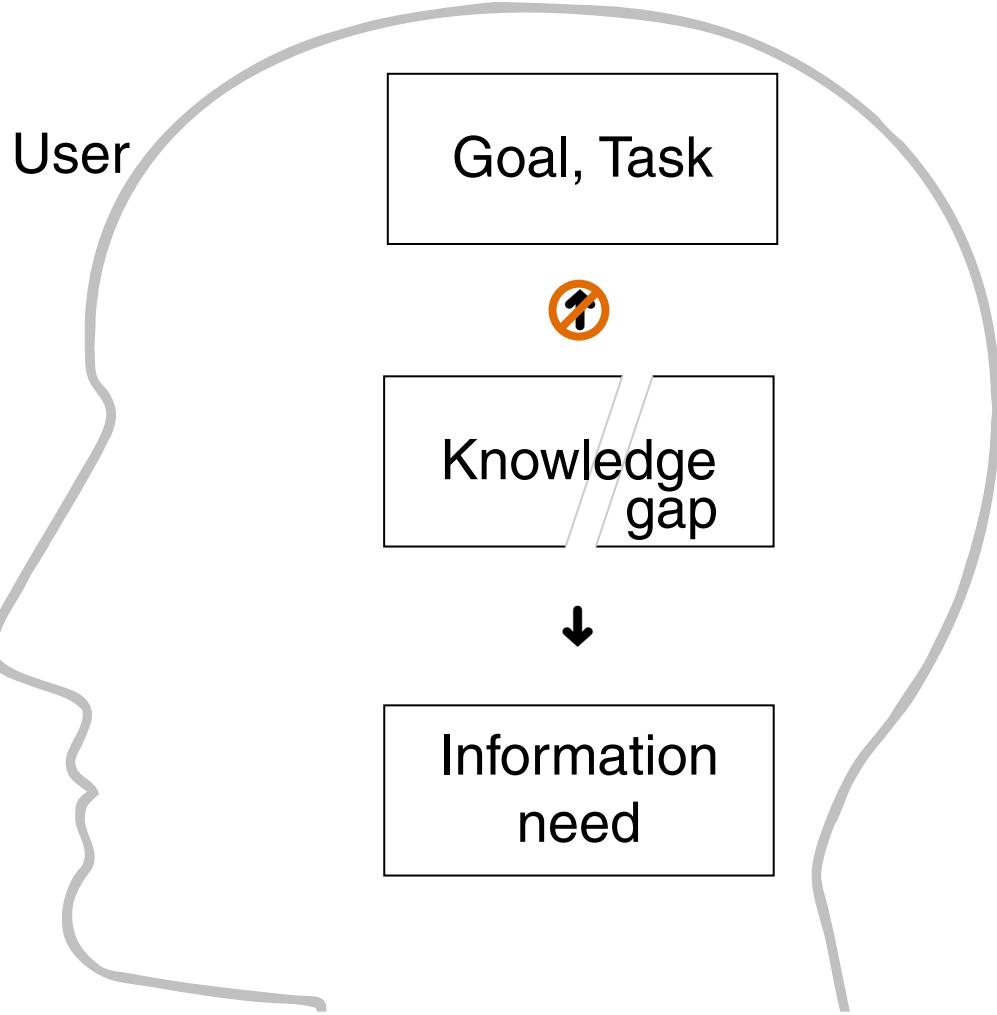


# Terminology

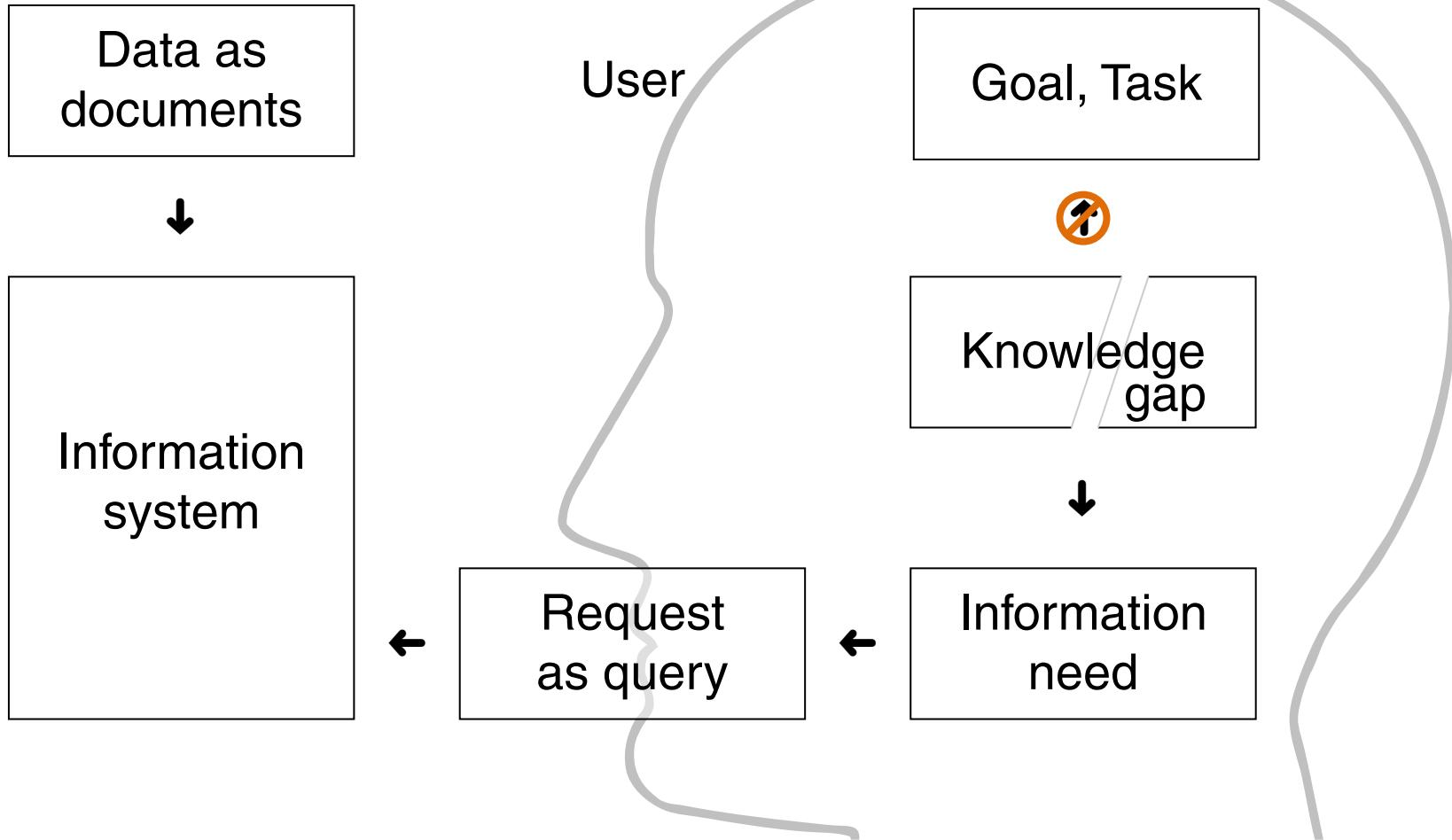
Data as documents



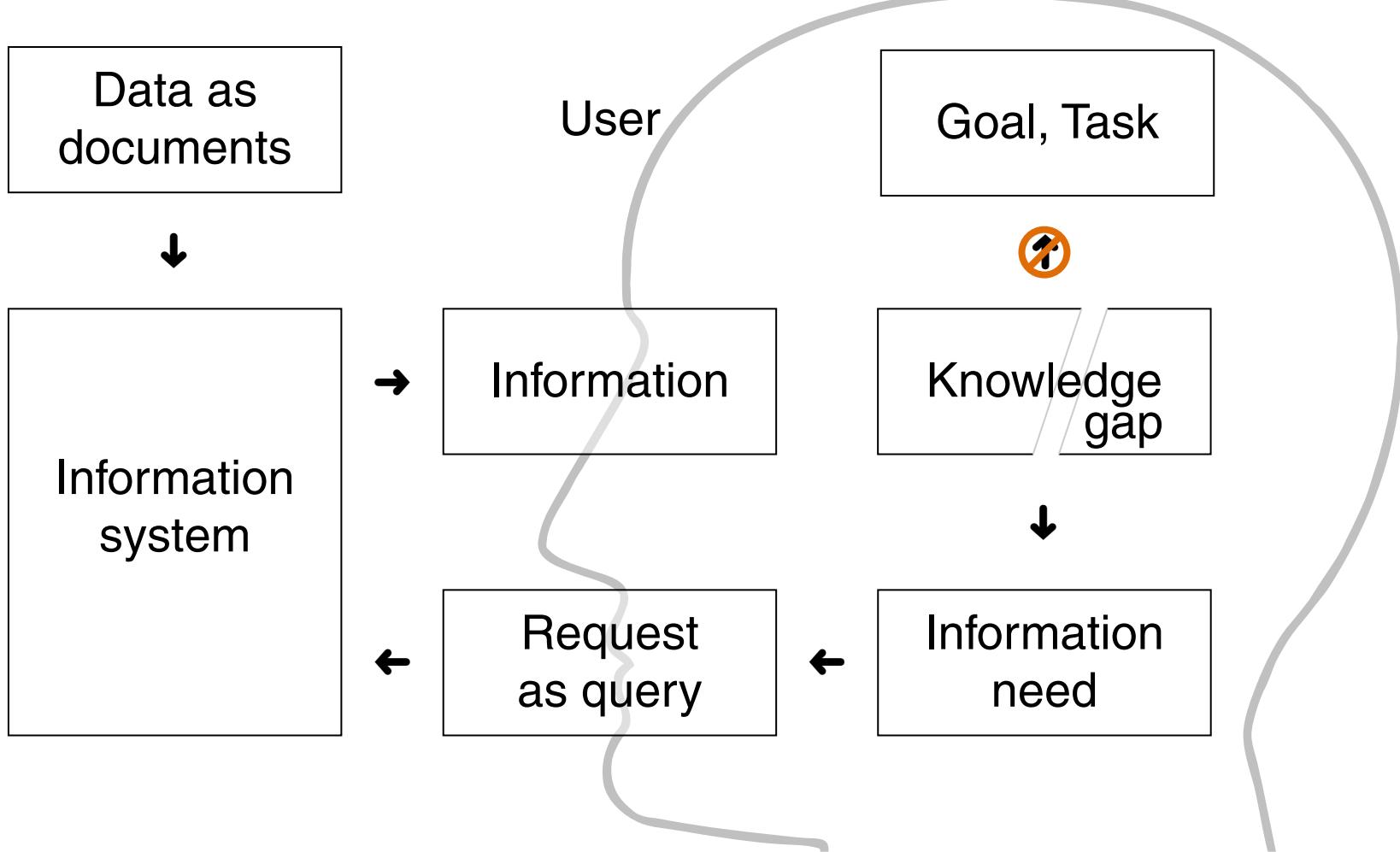
Information system



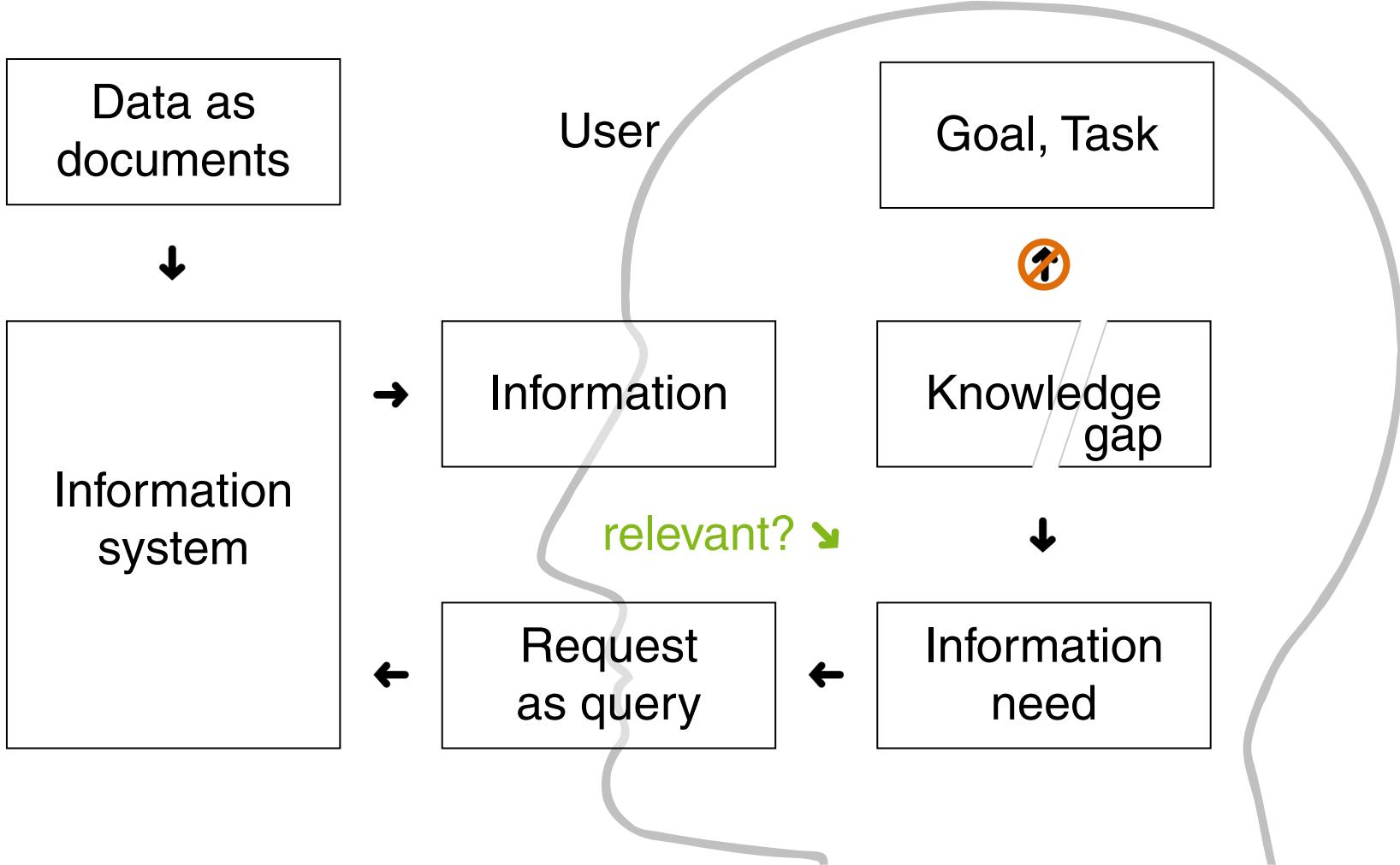
# Terminology



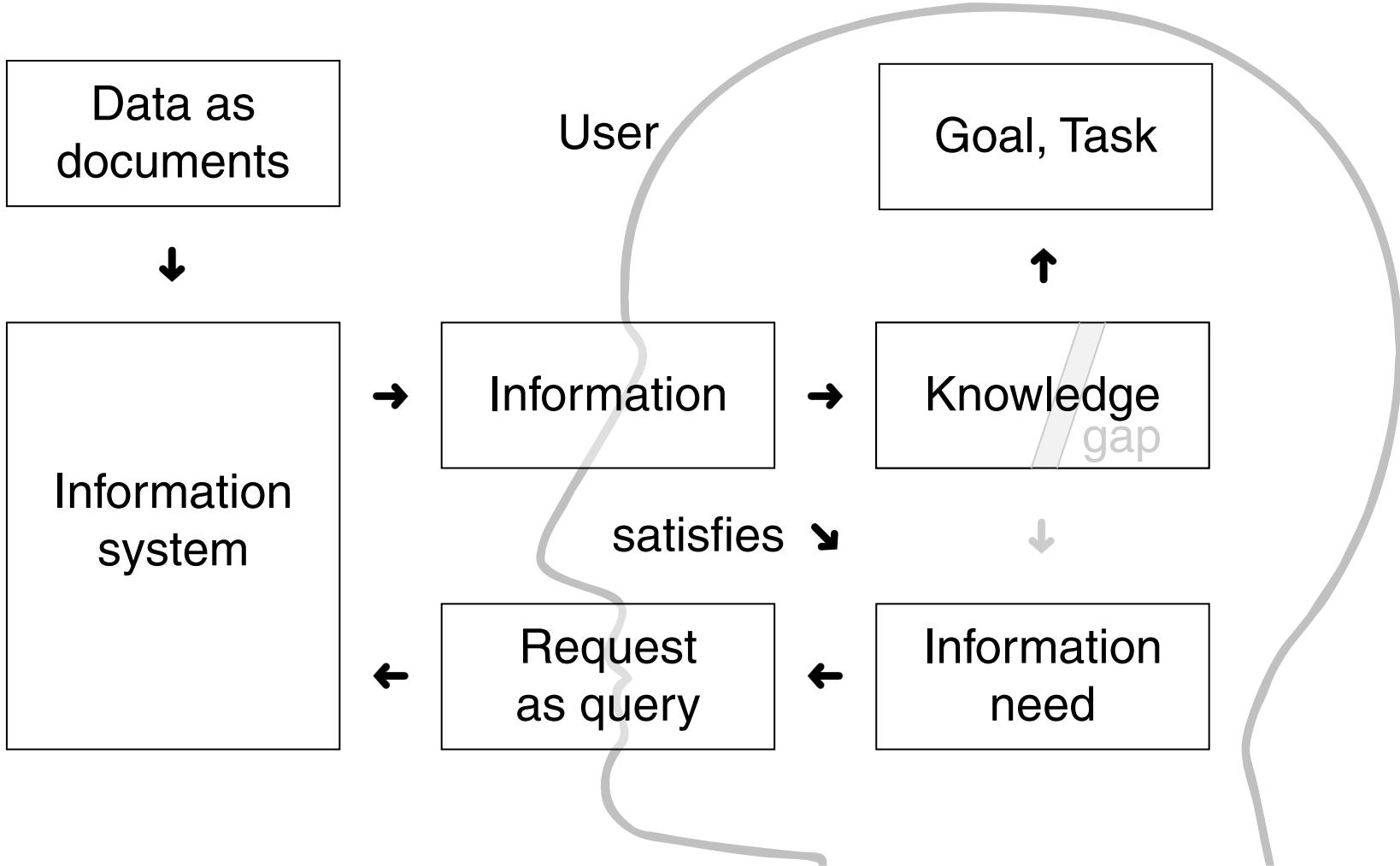
# Terminology



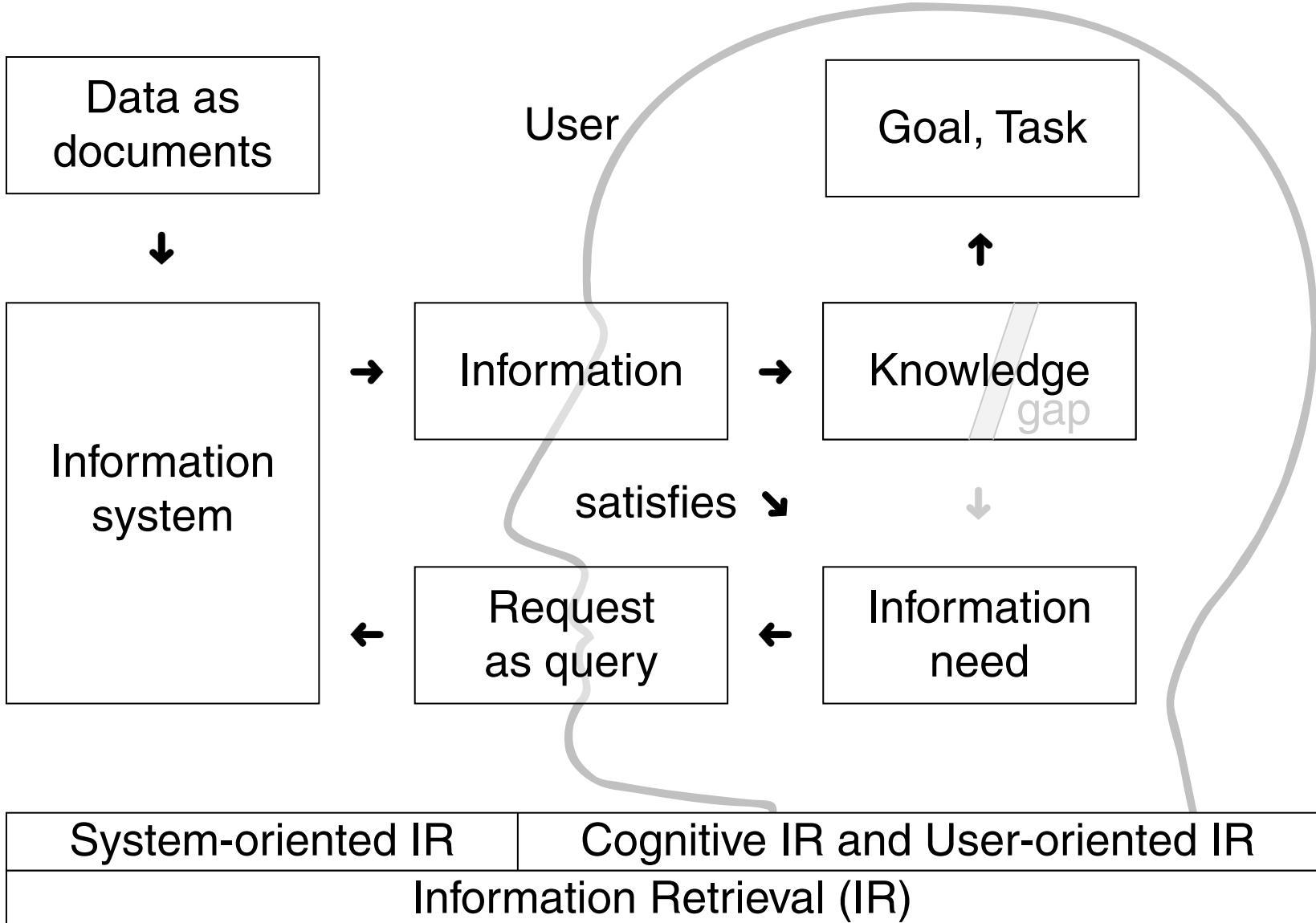
# Terminology



# 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 | 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)

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.
- “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.

[[Toms 2013](#)]

[[Wikipedia](#)]

# Chapter IR:I

## I. Introduction

- ❑ Information Retrieval in a Nutshell
- ❑ Examples of Information Retrieval Problems
- ❑ Terminology
- ❑ Delineation
- ❑ Historical Background

# 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

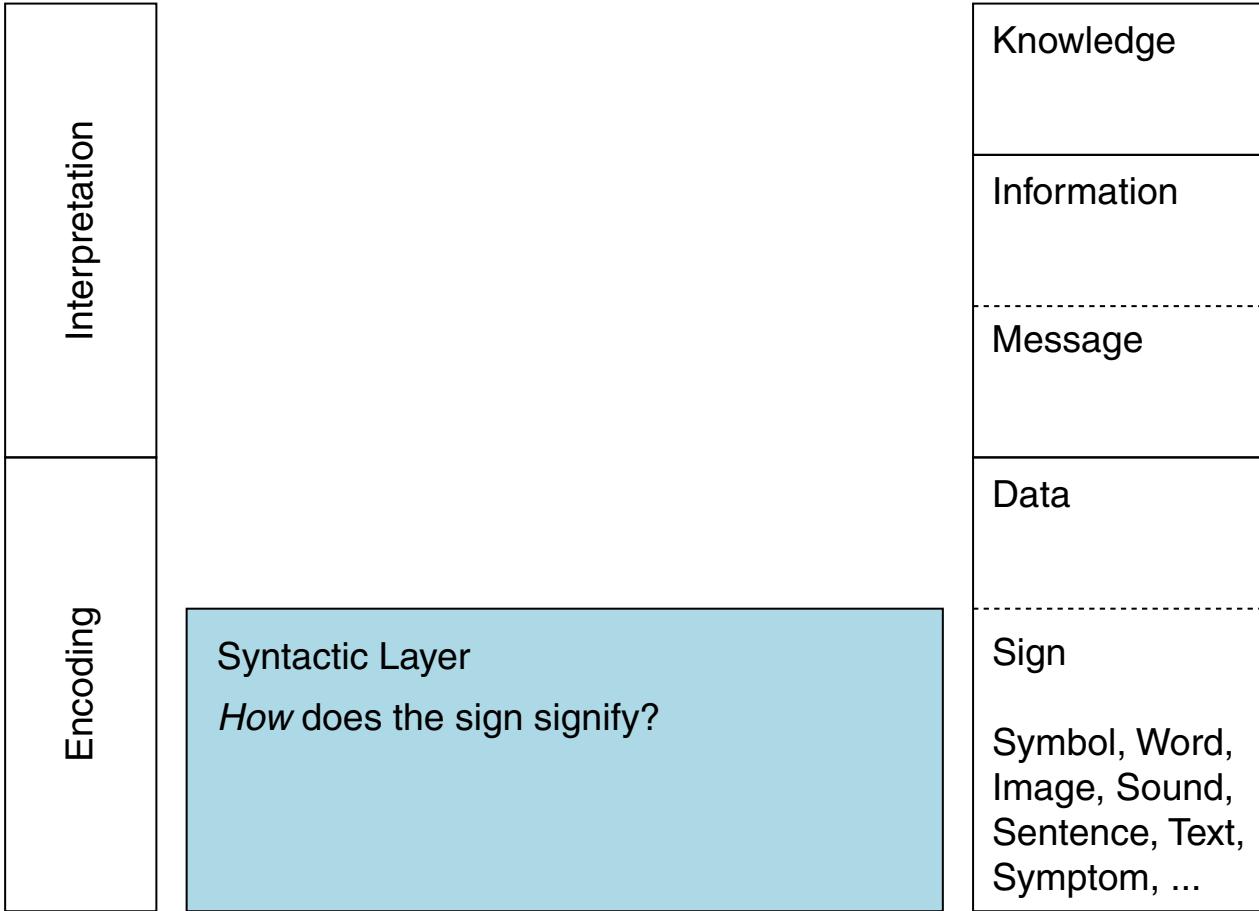
## Semiotics

### Communication

Channel      Semiotics

Information science

Example



10/2/2018

# Delineation

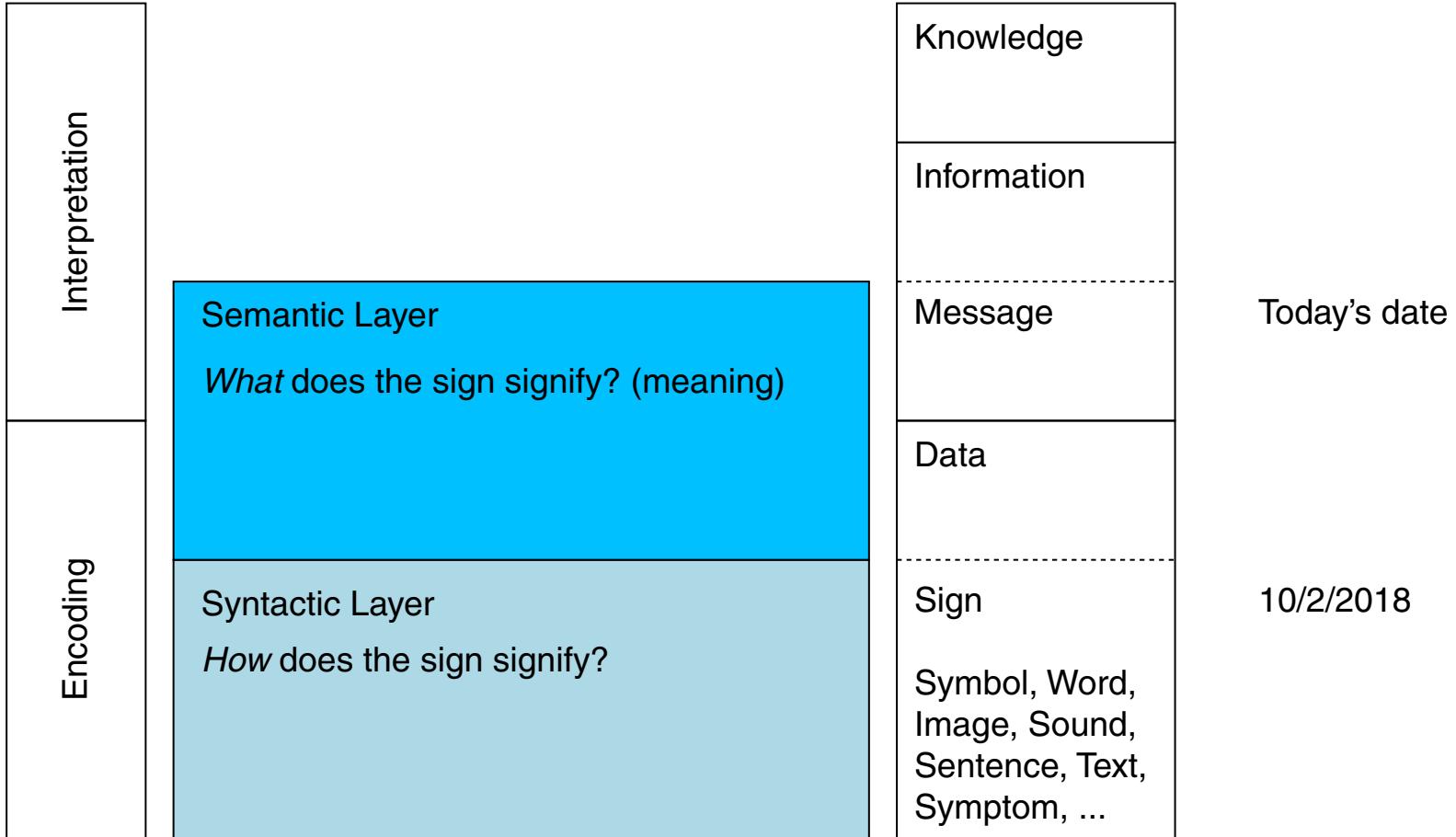
## Semiotics

### Communication

Channel      Semiotics

Information science

Example



# Delineation

## Semiotics

### Communication

Channel      Semiotics

Information science

Example

Interpretation	Semantic Layer <i>What does the sign signify? (meaning)</i>	Knowledge	Today's date
Encoding	Sigmatic Layer <i>What does the sign signify? (object)</i>	Message	Date
	Syntactic Layer <i>How does the sign signify?</i>	Data Sign Symbol, Word, Image, Sound, Sentence, Text, Symptom, ...	10/2/2018

# Delineation

## Semiotics

### Communication

Channel      Semiotics

Information science

Example

Interpretation	Semiotics	Information science	Example
	<p>Pragmatic Layer</p> <p><i>Why (and what for) is the sign signifying?</i></p>	<p>Knowledge</p>	Have I missed my mother's birthday?
	<p>Semantic Layer</p> <p><i>What does the sign signify? (meaning)</i></p>	<p>Information</p> <p>Message</p>	Today's date
	<p>Sigmatic Layer</p> <p><i>What does the sign signify? (object)</i></p>	<p>Data</p>	Date
Encoding	<p>Syntactic Layer</p> <p><i>How does the sign signify?</i></p>	<p>Sign</p> <p>Symbol, Word, Image, Sound, Sentence, Text, Symptom, ...</p>	10/2/2018

# Delineation

## Semiotics

### Communication

Channel      Semiotics

Information science

Example

Pragmatic Layer  
*Why (and what for) is the sign signifying?*

Knowledge

Birthdate of my mother: 10/1/1955

Semantic Layer  
*What does the sign signify? (meaning)*

Information

Have I missed my mother's birthday?

Sigmatic Layer  
*What does the sign signify? (object)*

Message

Today's date

Syntactic Layer  
*How does the sign signify?*

Data

Date

Sign  
Symbol, Word,  
Image, Sound,  
Sentence, Text,  
Symptom, ...

10/2/2018

## 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

OLAP, Online Analytical Processing

KDD, Knowledge Discovery in Databases

**Data mining**, Web mining, Text mining

Scenario: gigabytes, databases, on the  
(semantic) Web, in unstructured text

**Machine learning**

Scenario: in main memory,  
specific deduction model

Statistic analysis

Scenario: clean data,  
hypothesis evaluation

Explorative data analysis

# Delineation

## Machine Learning, Data Mining

Analysis	Information visualization	OLAP, Online Analytical Processing
	Data aggregation ...	KDD, Knowledge Discovery in Databases
	<b>Data mining</b> , Web mining, Text mining Scenario: gigabytes, databases, on the (semantic) Web, in unstructured text	
	<b>Machine learning</b> Scenario: in main memory, specific deduction model	
<b>Statistic analysis</b> Scenario: clean data, hypothesis evaluation		
Descriptive data analysis	Explorative data analysis	

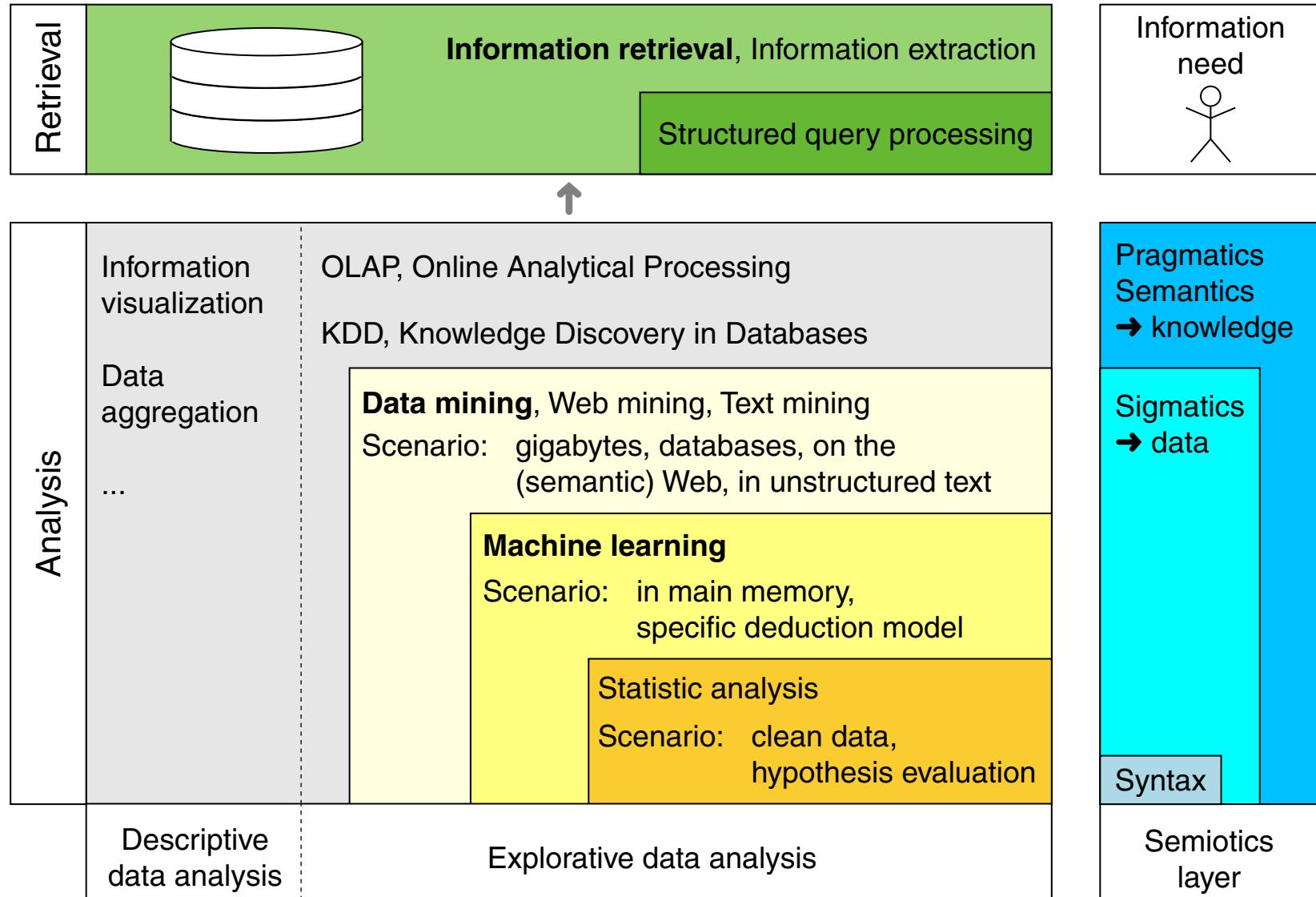
# Delineation

## Machine Learning, Data Mining

Analysis	Information visualization Data aggregation ...	OLAP, Online Analytical Processing KDD, Knowledge Discovery in Databases	Pragmatics Semantics → knowledge
		<b>Data mining</b> , Web mining, Text mining Scenario: gigabytes, databases, on the (semantic) Web, in unstructured text	Sigmatics → data
<b>Machine learning</b> Scenario: in main memory, specific deduction model		Statistic analysis Scenario: clean data, hypothesis evaluation	Syntax
Descriptive data analysis		Explorative data analysis	Semiotics layer

# Delineation

## Machine Learning, Data Mining



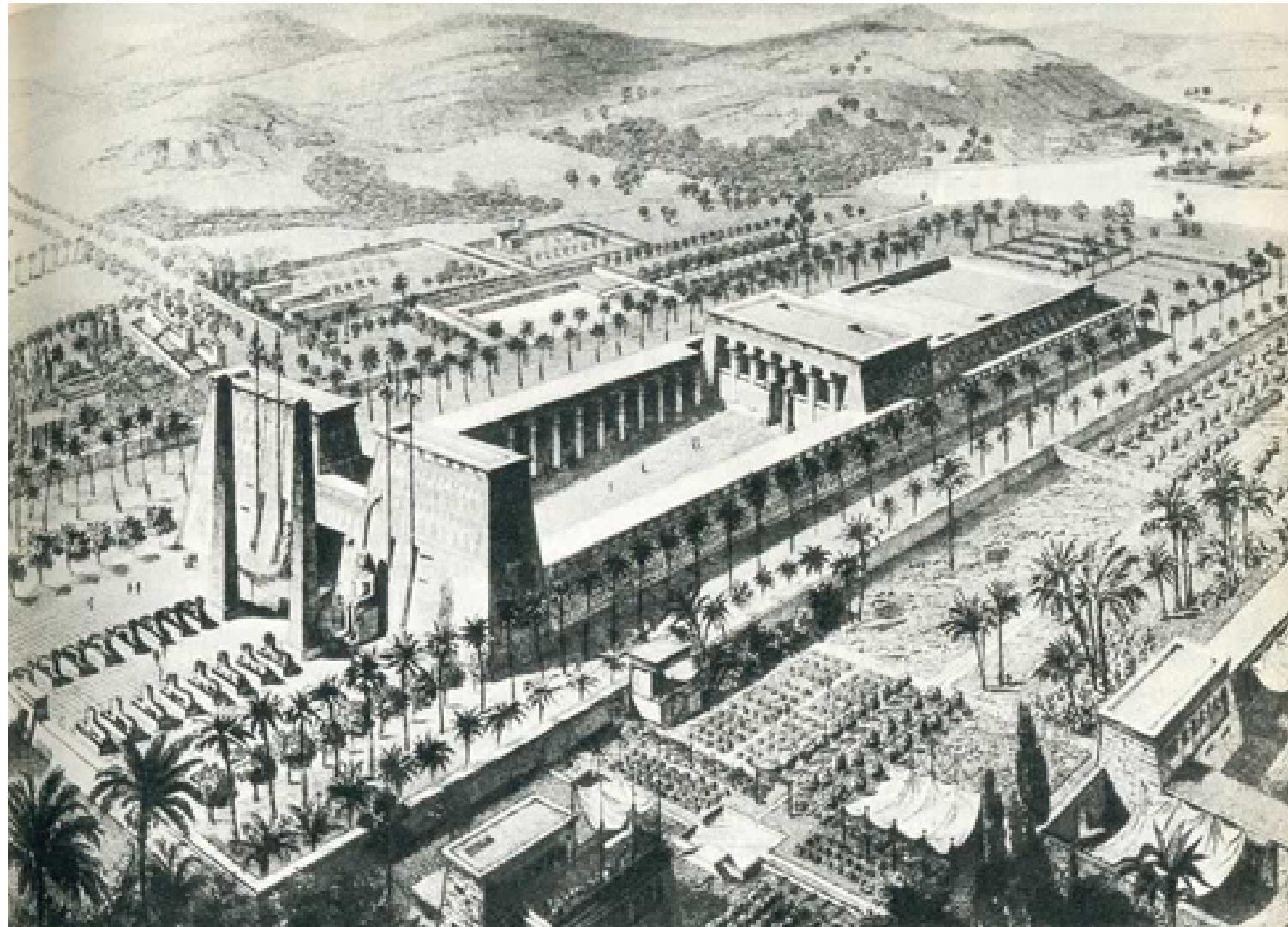
# Chapter IR:I

## I. Introduction

- ❑ Information Retrieval in a Nutshell
- ❑ Examples of Information Retrieval Problems
- ❑ Terminology
- ❑ Delineation
- ❑ Historical Background

# Historical Background

## Manual Retrieval



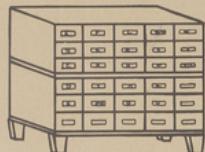
## Remarks:

- ❑ 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

# FROM CARD CATALOG TO THE BOOK ON THE SHELF



**THE CARD CATALOG**  
is an alphabetical list of  
books found in the Library

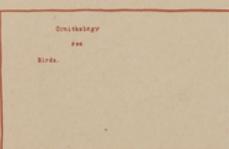
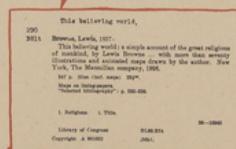
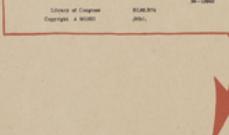
### THE THREE WAYS OF FINDING A BOOK IN THE CATALOG



UNDER AUTHORS SURNAME

UNDER TITLE OF BOOK

UNDER SUBJECT WITH WHICH BOOK DEALS



**PEABODY**  
VISUAL AIDS

PUBLISHED BY  
FOLLETT BOOK COMPANY CHICAGO

Prepared under the direction of Miss Ruby Ethel Gurdif for the Peabody Library School course in Teaching the Use of the Library. Planned by Martha Edmondson, lettered by Mr. McCord.



# Historical Background

## Manual Retrieval



## Remarks:

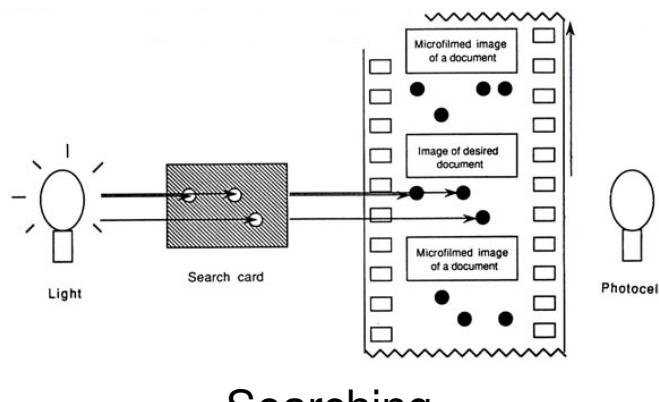
- ❑ Today, [WorldCat](#) is a union catalog that itemizes the collections of 72,000 libraries in 170 countries and territories. It contains more than 447 million records, representing over 2.7 billion physical and digital assets in 484 languages, as of January 2019. [\[OCLC\]](#)
- ❑ What are problems 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

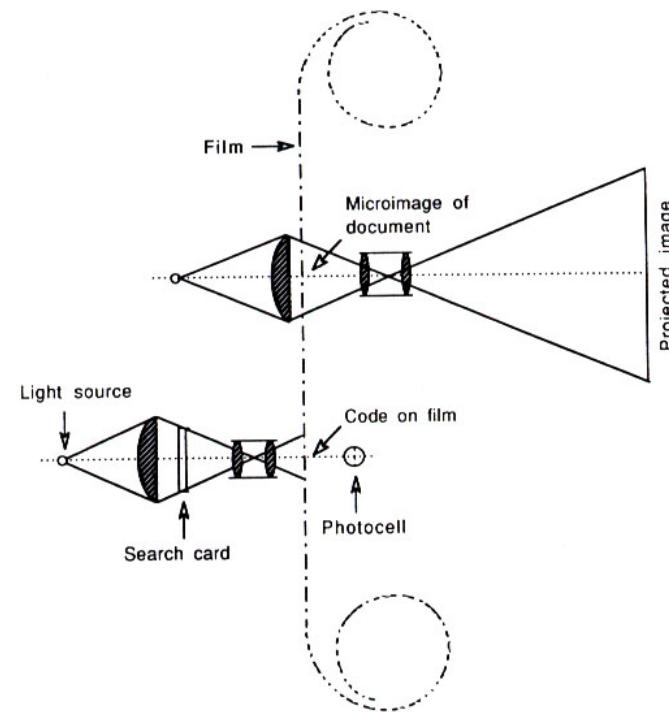
## Mechanical Retrieval

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

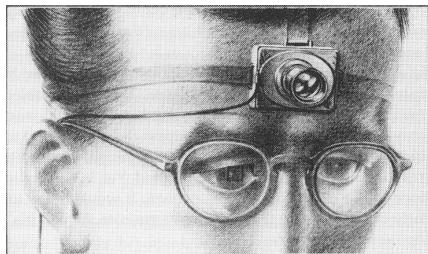


Result presentation

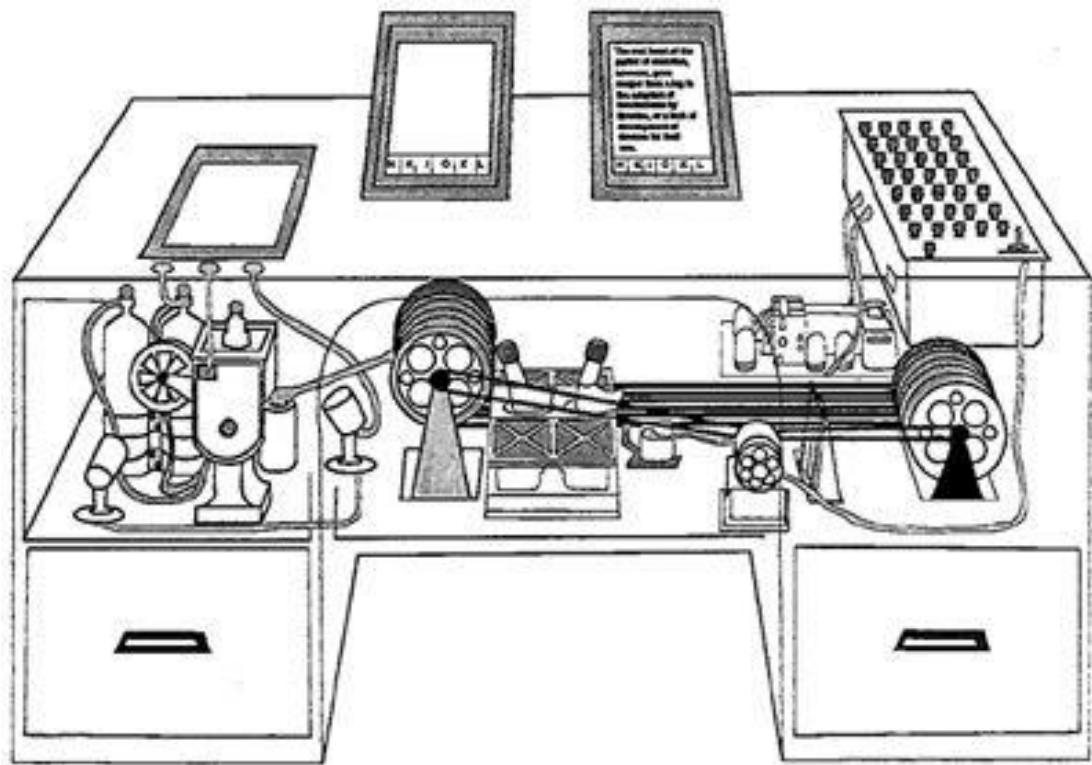
# Historical Background

## 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.

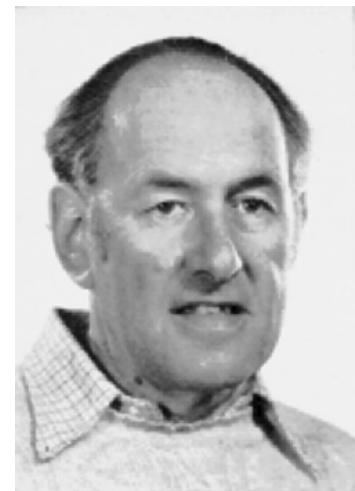
# Historical Background

## 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)

### $tf \cdot idf$ -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  $tf \cdot idf$  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.

# Historical Background

## Information Retrieval (today)

It's been a long way

# Historical Background

## Information Retrieval (today)



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

# Historical Background

## Information Retrieval (today)



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

# Historical Background

## Information Retrieval (today)



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

