

Information Retrieval

01 – Introduction

Philipp Schaer

2025-04-03 – Cologne, Germany
<https://ir.web.th-koeln.de>

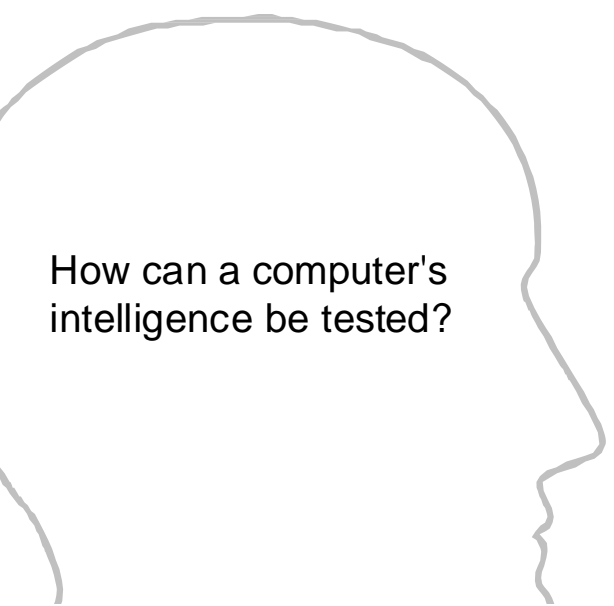


Technology
Arts Sciences
TH Köln



Information Retrieval in a Nutshell

- **Eine vage Anfrage.**
 - Ausdruck eines komplexen Informationsbedürfnisses: Meist eine Frage
- **Milliarden von Dokumenten.**
 - Texte, Bilder, Audio-Dateien, Videos, ...




How can a computer's intelligence be tested?



Information Retrieval in a Nutshell

- **Eine vage Anfrage.**
 - Ausdruck eines komplexen Informationsbedürfnisses: Meist eine Frage
- Finde relevante Dokumente in Millisekunden!
- **Milliarden von Dokumenten.**
 - Texte, Bilder, Audio-Dateien, Videos, ...
- **Hohes Ungleichgewicht.**
 - Nur ein winziger Bruchteil aller Dokumente ist für die Anfrage relevant.



How can a computer's intelligence be tested?

→ „Computer Intelligence Test“ →

[plato.stanford.edu > entries > turing-test](https://plato.stanford.edu/entries/turing-test/) ▼

[The Turing Test \(Stanford Encyclopedia of Philosophy\)](https://plato.stanford.edu/entries/turing-test/)

by G Oppy - 2003 - Cited by 130 - [Related articles](#)

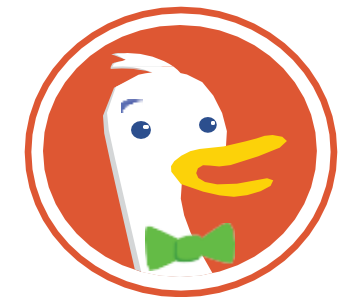
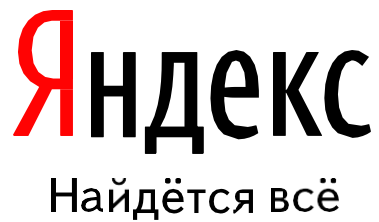
Apr 9, 2003 - The phrase “The Turing Test” is most properly used to refer to a propos

Im Information Retrieval suchen wir also die Nadel im ...



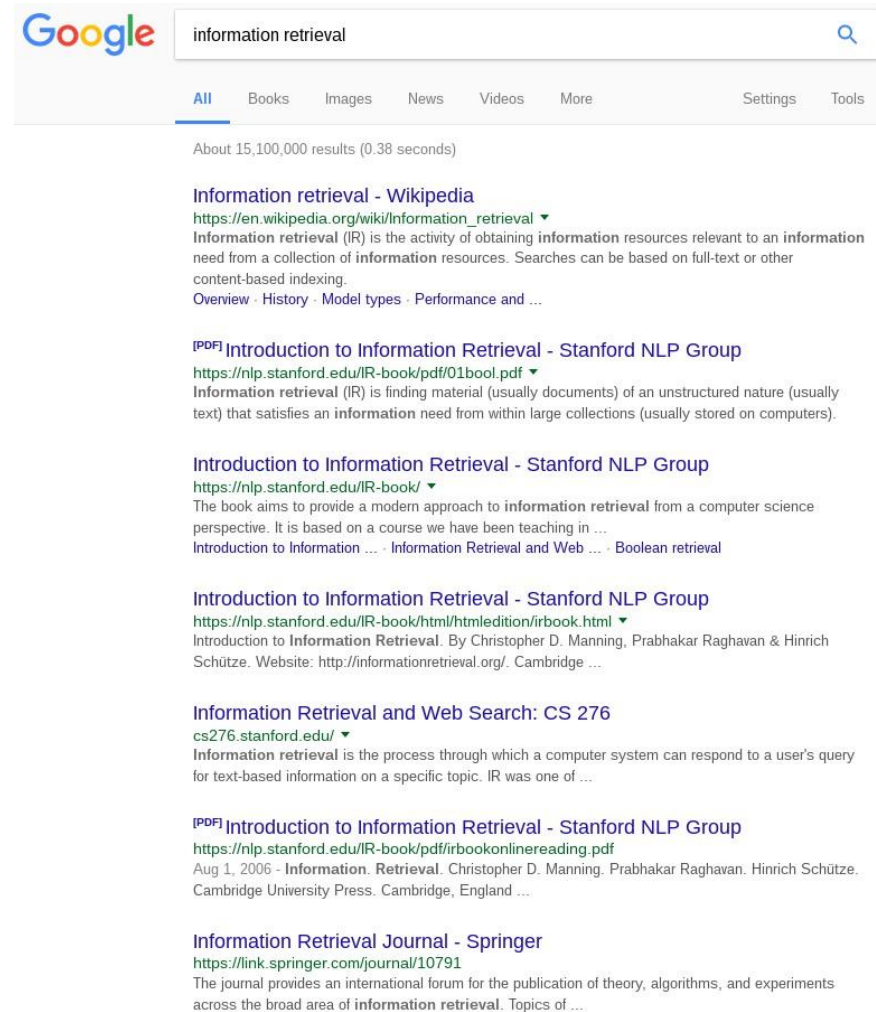
... und die zu finden, ist ganz
schön **schwierig!**

Und das macht nicht nur Google...!



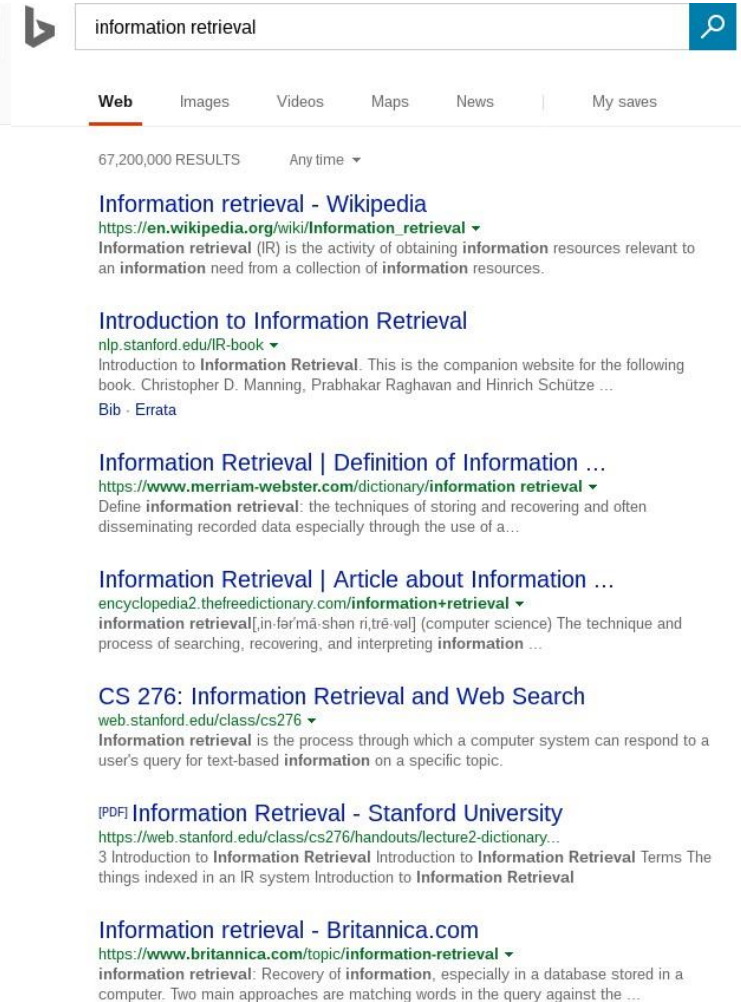
Ein Beispiel...

- Sie möchten alles wissen, was es zum Thema Information Retrieval gibt.
- Sie suchen Texte, die 'information' und 'retrieval' beinhalten.



Google search results for "information retrieval". The search bar shows "information retrieval" and the results are filtered by "All". The results include:

- Information retrieval - Wikipedia**
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/01bool.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**
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 ...

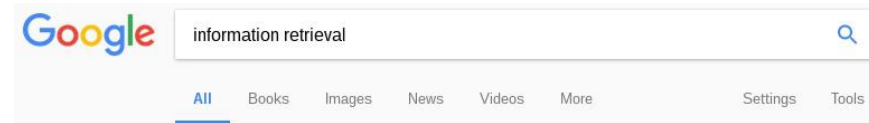


Bing search results for "information retrieval". The search bar shows "information retrieval" and the results are filtered by "Web". The results include:

- Information retrieval - Wikipedia**
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**
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
 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 ...**
encyclopedia2.thefreedictionary.com/information+retrieval
 information retrieval[in-fər-mā-shən rī-tré-vəl] (computer science) The technique and process of searching, recovering, and interpreting information ...
- CS 276: Information Retrieval and Web Search**
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...>
 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 ...

Ein Beispiel...

- Sie möchten alles wissen, was es zum Thema Information Retrieval gibt.
- Sie suchen Texte, die 'information' und 'retrieval' beinhalten.



About 15,100,000 results (0.38 seconds)

Information retrieval - Wikipedia

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

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

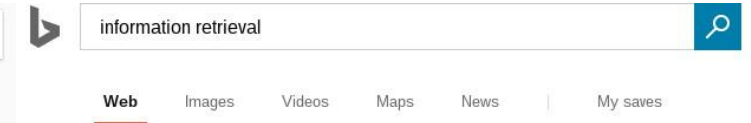
<https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

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

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

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

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 ...

encyclopedia2.thefreedictionary.com/information+retrieval

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

CS 276: Information Retrieval and Web Search

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...>

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

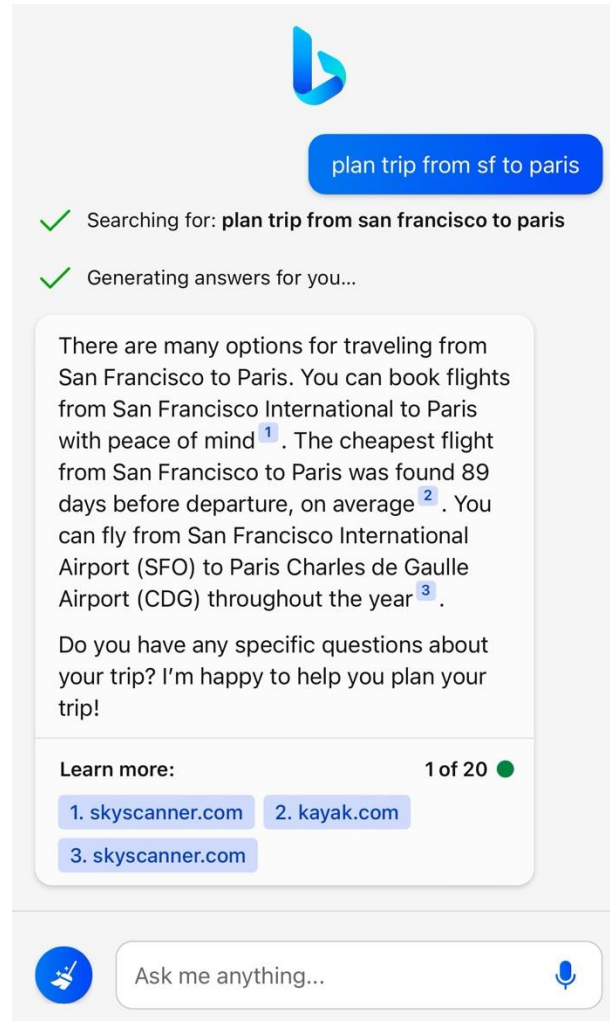
Information retrieval - Britannica.com

<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 ...

Ein weiteres Beispiel...

- Müssen wir wirklich so viele Ergebnisse überprüfen?
- Das geht doch heute viel einfacher... z.B. mit ChatGPT oder Bing oder Perplexity oder ...



plan trip from sf to paris

✓ Searching for: **plan trip from san francisco to paris**

✓ Generating answers for you...

There are many options for traveling from San Francisco to Paris. You can book flights from San Francisco International to Paris with peace of mind ¹. The cheapest flight from San Francisco to Paris was found 89 days before departure, on average ². You can fly from San Francisco International Airport (SFO) to Paris Charles de Gaulle Airport (CDG) throughout the year ³.

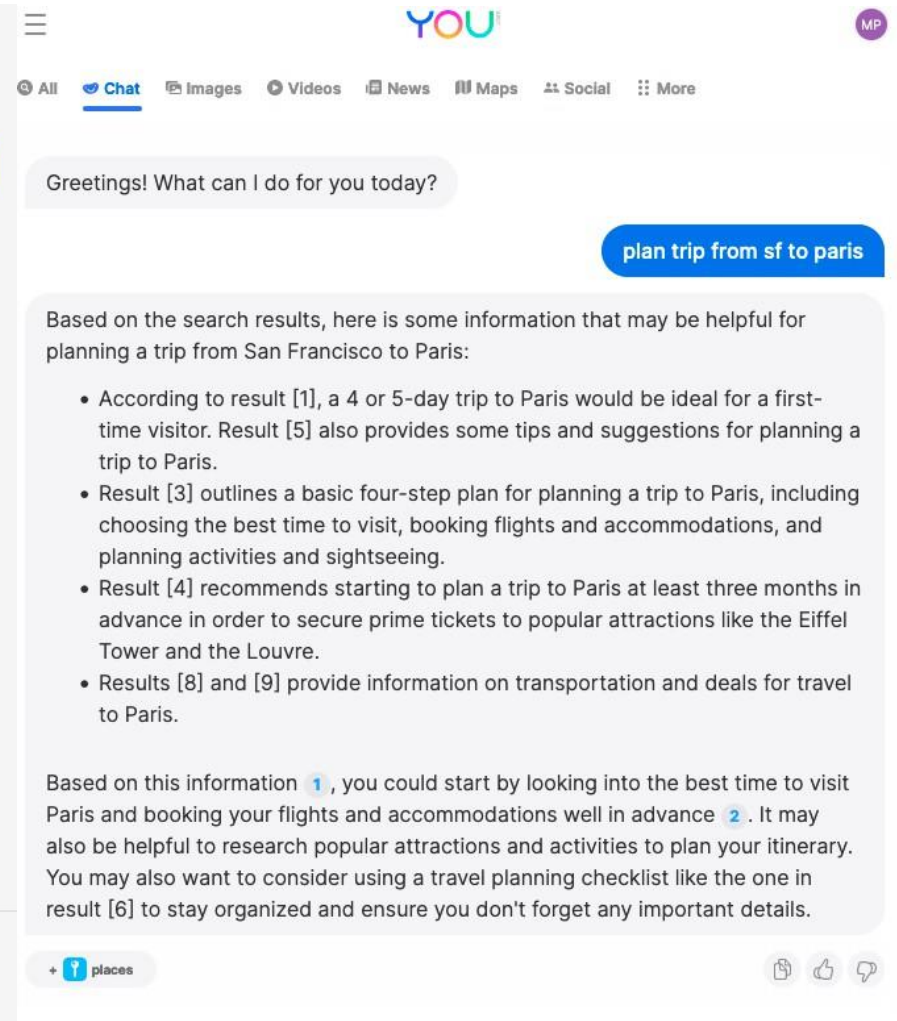
Do you have any specific questions about your trip? I'm happy to help you plan your trip!

Learn more: 1 of 20 ●

1. skyscanner.com 2. kayak.com

3. skyscanner.com

Ask me anything...



YOU

All Chat Images Videos News Maps Social More

Greetings! What can I do for you today?

plan trip from sf to paris

Based on the search results, here is some information that may be helpful for planning a trip from San Francisco to Paris:

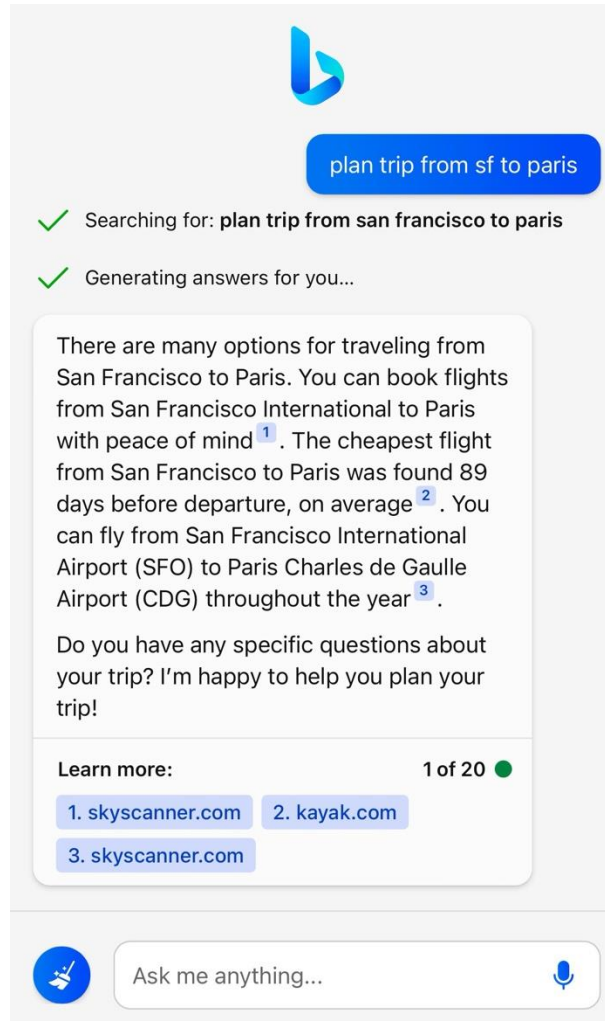
- According to result [1], a 4 or 5-day trip to Paris would be ideal for a first-time visitor. Result [5] also provides some tips and suggestions for planning a trip to Paris.
- Result [3] outlines a basic four-step plan for planning a trip to Paris, including choosing the best time to visit, booking flights and accommodations, and planning activities and sightseeing.
- Result [4] recommends starting to plan a trip to Paris at least three months in advance in order to secure prime tickets to popular attractions like the Eiffel Tower and the Louvre.
- Results [8] and [9] provide information on transportation and deals for travel to Paris.

Based on this information ¹, you could start by looking into the best time to visit Paris and booking your flights and accommodations well in advance ². It may also be helpful to research popular attractions and activities to plan your itinerary. You may also want to consider using a travel planning checklist like the one in result [6] to stay organized and ensure you don't forget any important details.

+ places

Ein weiteres Beispiel...

- Müssen wir wirklich so viele Ergebnisse überprüfen?
- Das geht doch heute viel einfacher... z.B. mit ChatGPT oder Bing oder Perplexity oder ...



The screenshot shows the Bing AI interface. At the top, there's a blue 'b' logo and a search bar with the text 'plan trip from sf to paris'. Below the search bar, there are two green checkmarks indicating the search status: 'Searching for: plan trip from san francisco to paris' and 'Generating answers for you...'. The main content area displays a paragraph of text about flight options from San Francisco to Paris, mentioning Skyscanner and Kayak. Below the text, there's a 'Learn more:' section with three links: '1. skyscanner.com', '2. kayak.com', and '3. skyscanner.com'. At the bottom, there's a blue microphone icon and a text input field with the placeholder 'Ask me anything...'.

plan trip from sf to paris

✓ Searching for: plan trip from san francisco to paris

✓ Generating answers for you...

There are many options for traveling from San Francisco to Paris. You can book flights from San Francisco International to Paris with peace of mind ¹. The cheapest flight from San Francisco to Paris was found 89 days before departure, on average ². You can fly from San Francisco International Airport (SFO) to Paris Charles de Gaulle Airport (CDG) throughout the year ³.

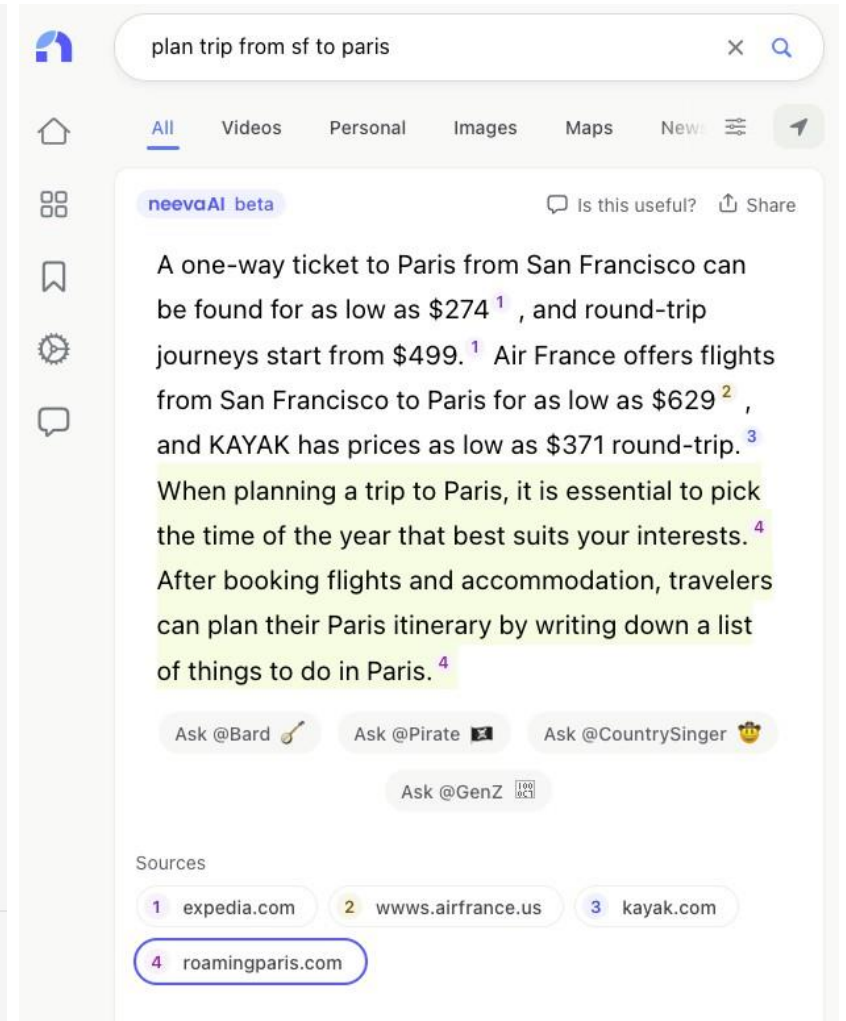
Do you have any specific questions about your trip? I'm happy to help you plan your trip!

Learn more: 1 of 20 ●

1. skyscanner.com 2. kayak.com

3. skyscanner.com

Ask me anything...



The screenshot shows the Neeva AI interface. At the top, there's a search bar with the text 'plan trip from sf to paris'. Below the search bar, there are tabs for 'All', 'Videos', 'Personal', 'Images', 'Maps', and 'News'. The main content area displays a paragraph of text about flight options from San Francisco to Paris, mentioning Air France and Kayak. Below the text, there's a 'Sources' section with four links: '1. expedia.com', '2. www.airfrance.us', '3. kayak.com', and '4. roamingparis.com'. At the bottom, there's a blue microphone icon and a text input field with the placeholder 'Ask me anything...'.

plan trip from sf to paris

All Videos Personal Images Maps News

neevaAI beta Is this useful? Share

A one-way ticket to Paris from San Francisco can be found for as low as \$274 ¹, and round-trip journeys start from \$499. ¹ Air France offers flights from San Francisco to Paris for as low as \$629 ², and KAYAK has prices as low as \$371 round-trip. ³ When planning a trip to Paris, it is essential to pick the time of the year that best suits your interests. ⁴ After booking flights and accommodation, travelers can plan their Paris itinerary by writing down a list of things to do in Paris. ⁴

Ask @Bard Ask @Pirate Ask @CountrySinger

Ask @GenZ

Sources

1 expedia.com 2 www.airfrance.us 3 kayak.com

4 roamingparis.com

Ask me anything...

Begrifflichkeiten

In der Informationswissenschaft unterscheiden wir **Daten**, **Information** und **Wissen**.

Definition 1 (Daten, data)

- Eine auf einem Speichermedium aufgezeichnete Folge von Symbolen.

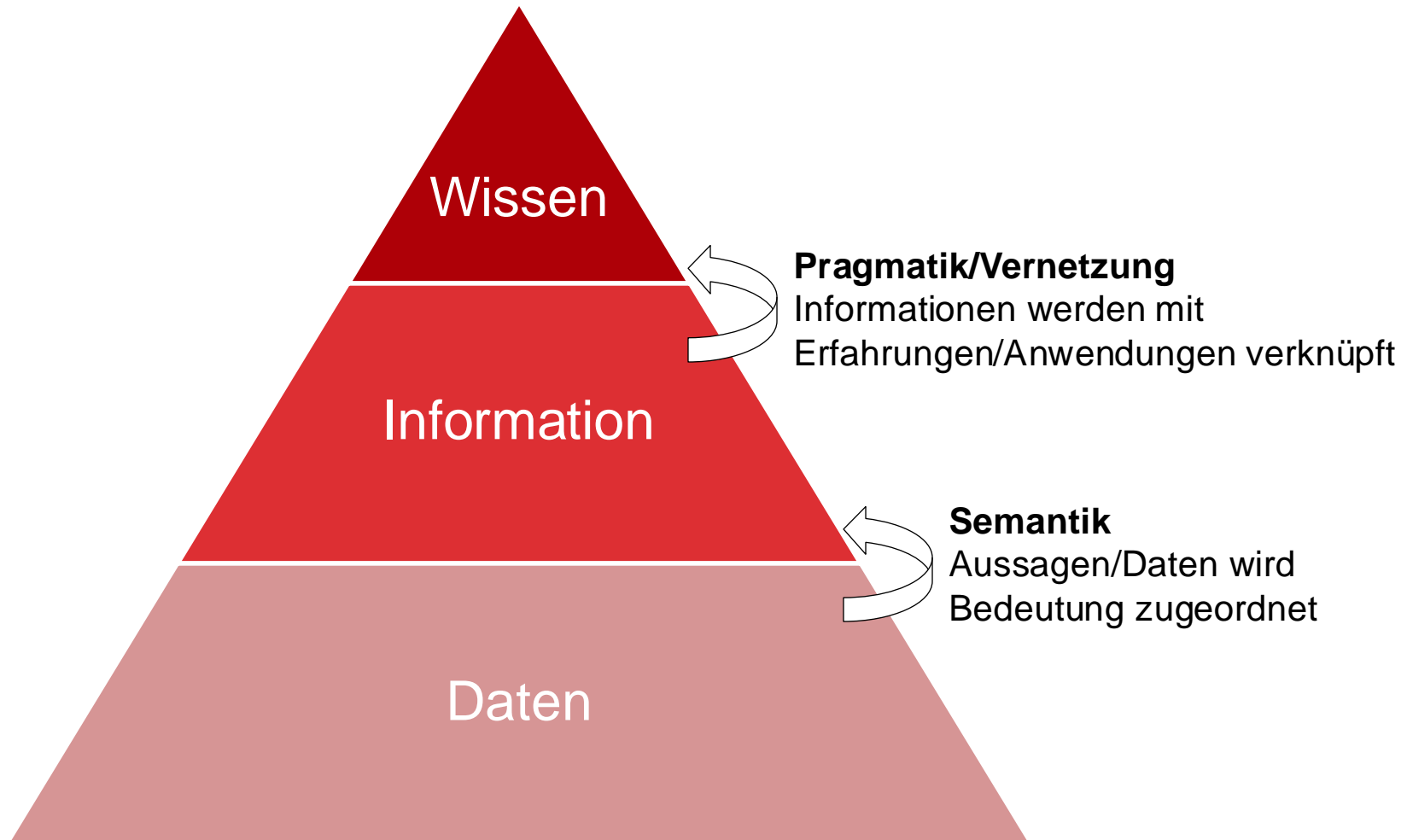
Definition 2 (Information)

- Daten, die nützlich sind, wobei “nützlich” z.B. aussagekräftig, interpretierbar, sachlich, usw. ist.

Definition 3 (Wissen, knowledge)

- Wissen ist ein Gedanke, der (unserer begründeten Überzeugung nach) wahr ist.
- Wissen entsteht aus Informationen und befähigt den Wissenden zum Handeln.

Daten – Information – Wissen



Weitere Begrifflichkeiten

Definition 4 (Informationssystem, information system)

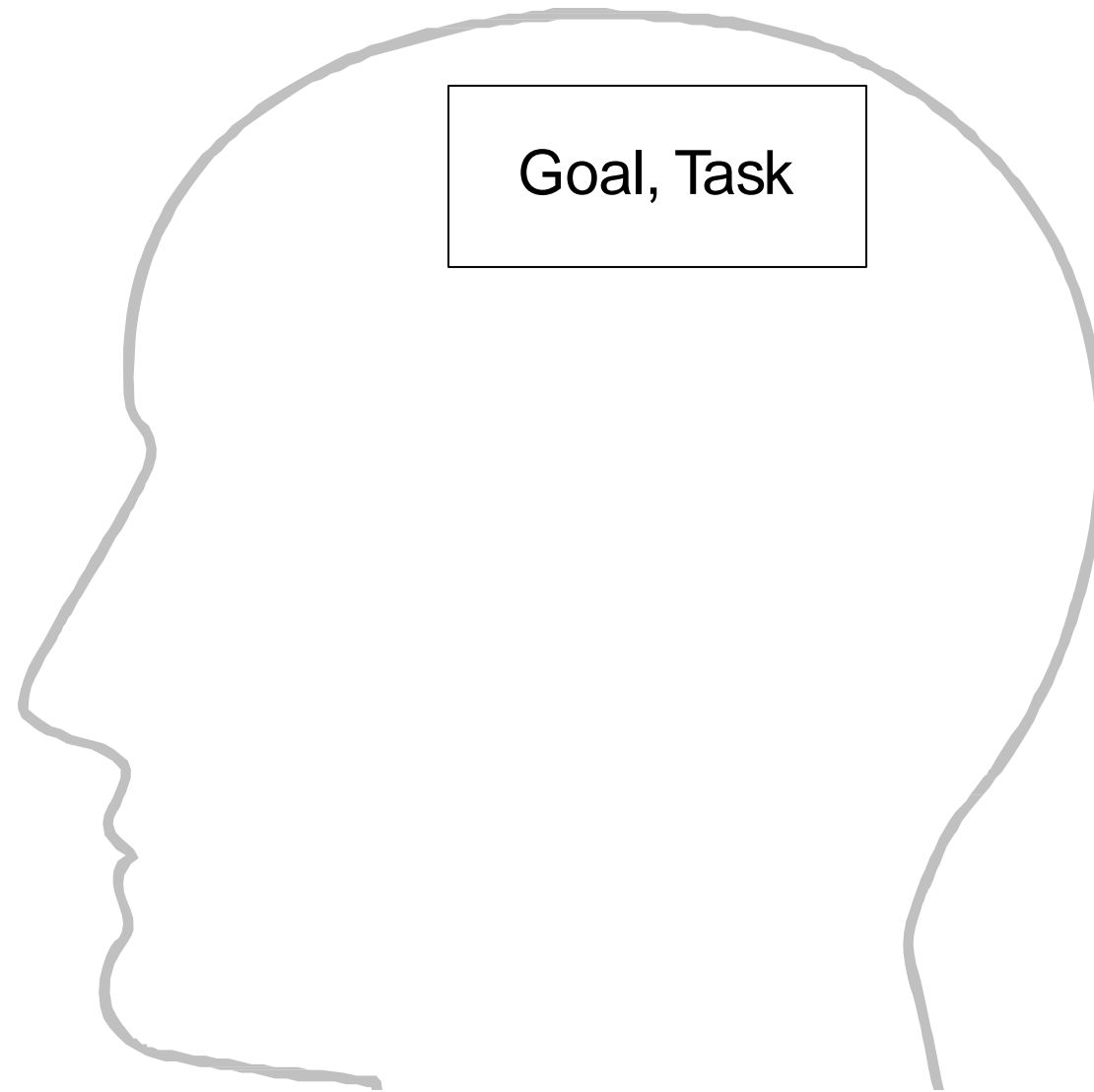
- Ein System zum Sammeln, Erstellen, Speichern, Verarbeiten und Verteilen von Informationen, einschließlich Hard-/Software, Betreiber, Nutzer und der Daten selbst.

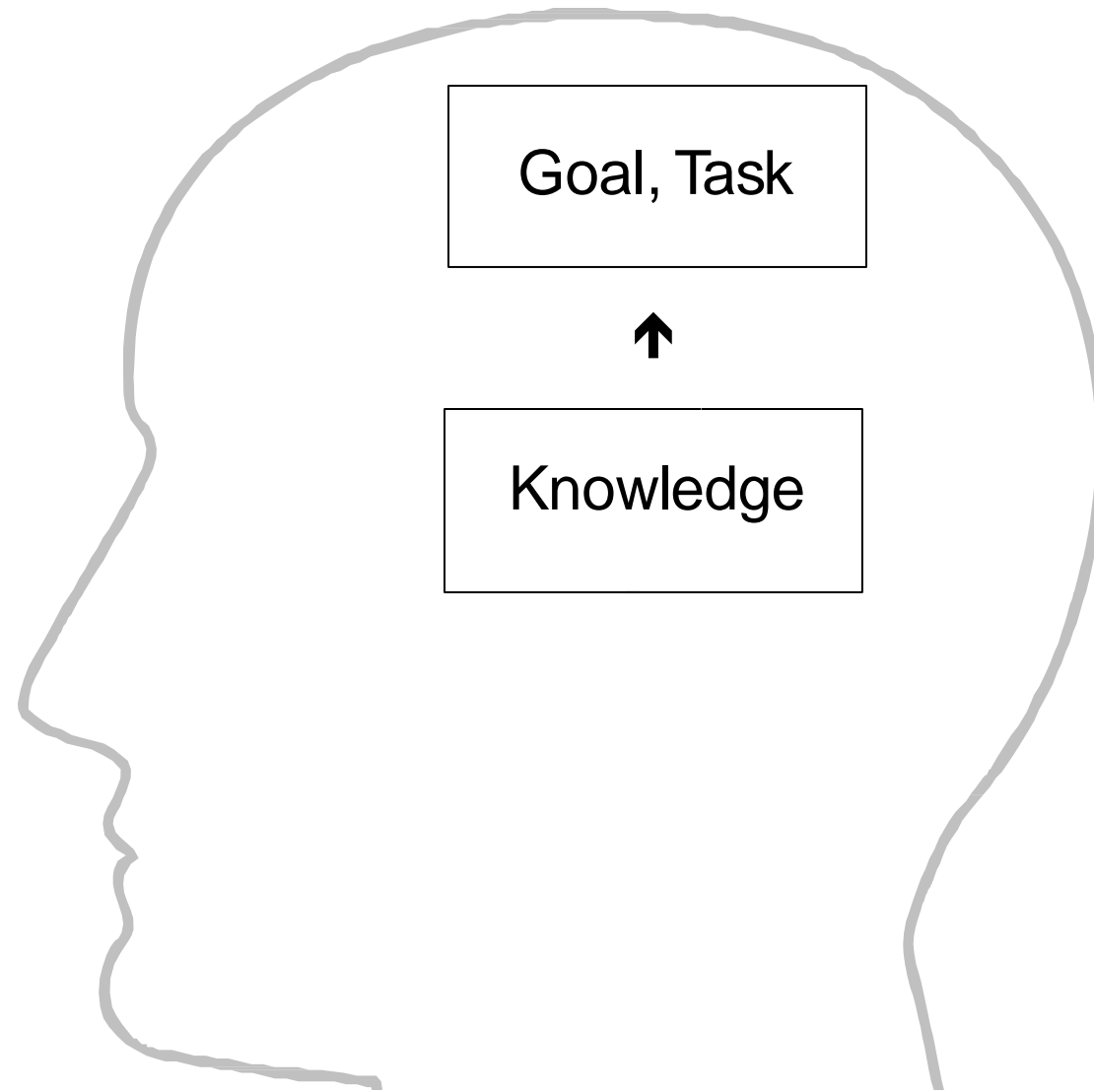
Definition 5 (Informationsbedarf, information need)

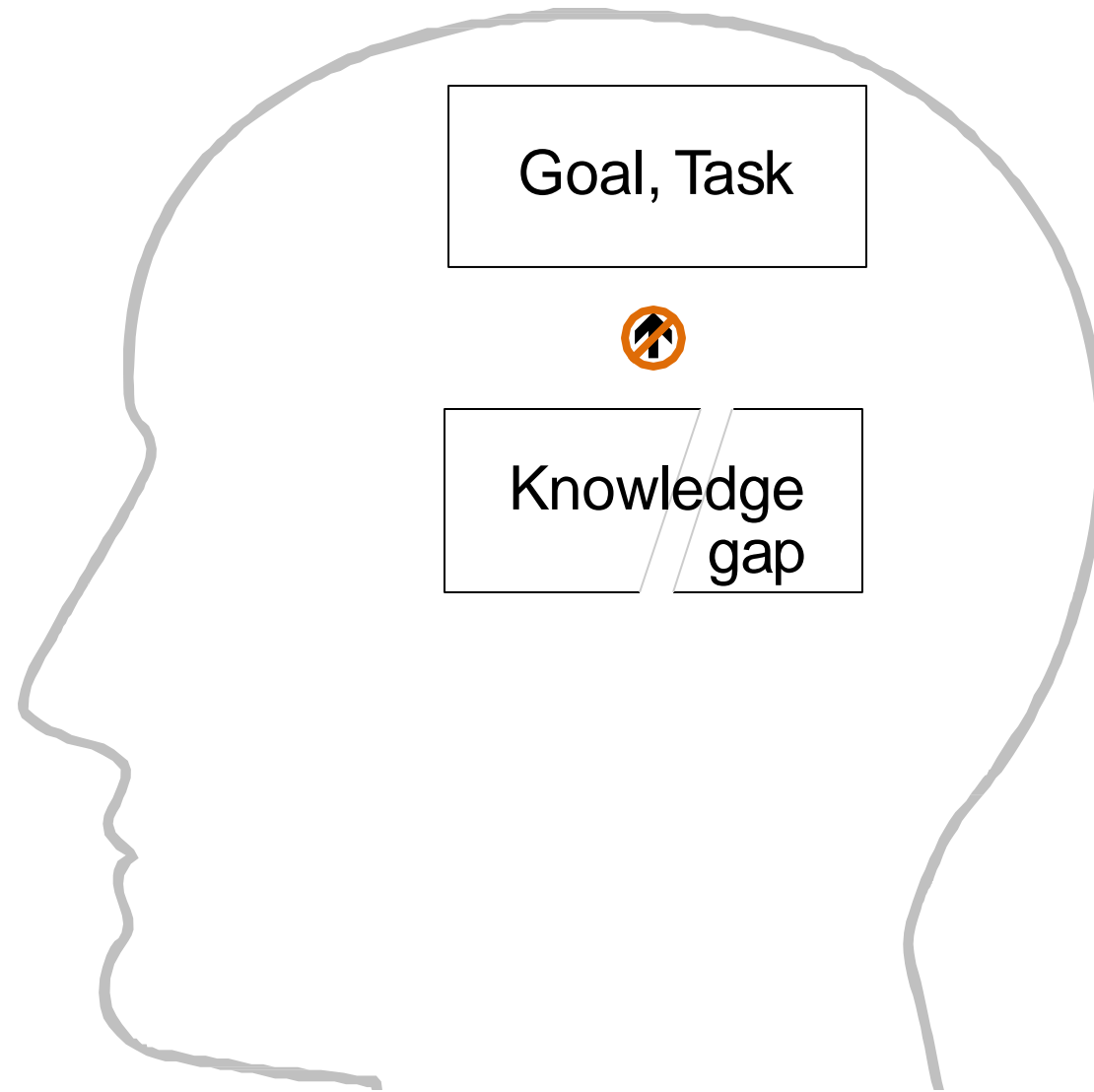
- Der Wunsch eines Benutzers, Informationen zu finden und zu erhalten, um ein bewusstes oder unbewusstes Ziel zu erreichen.

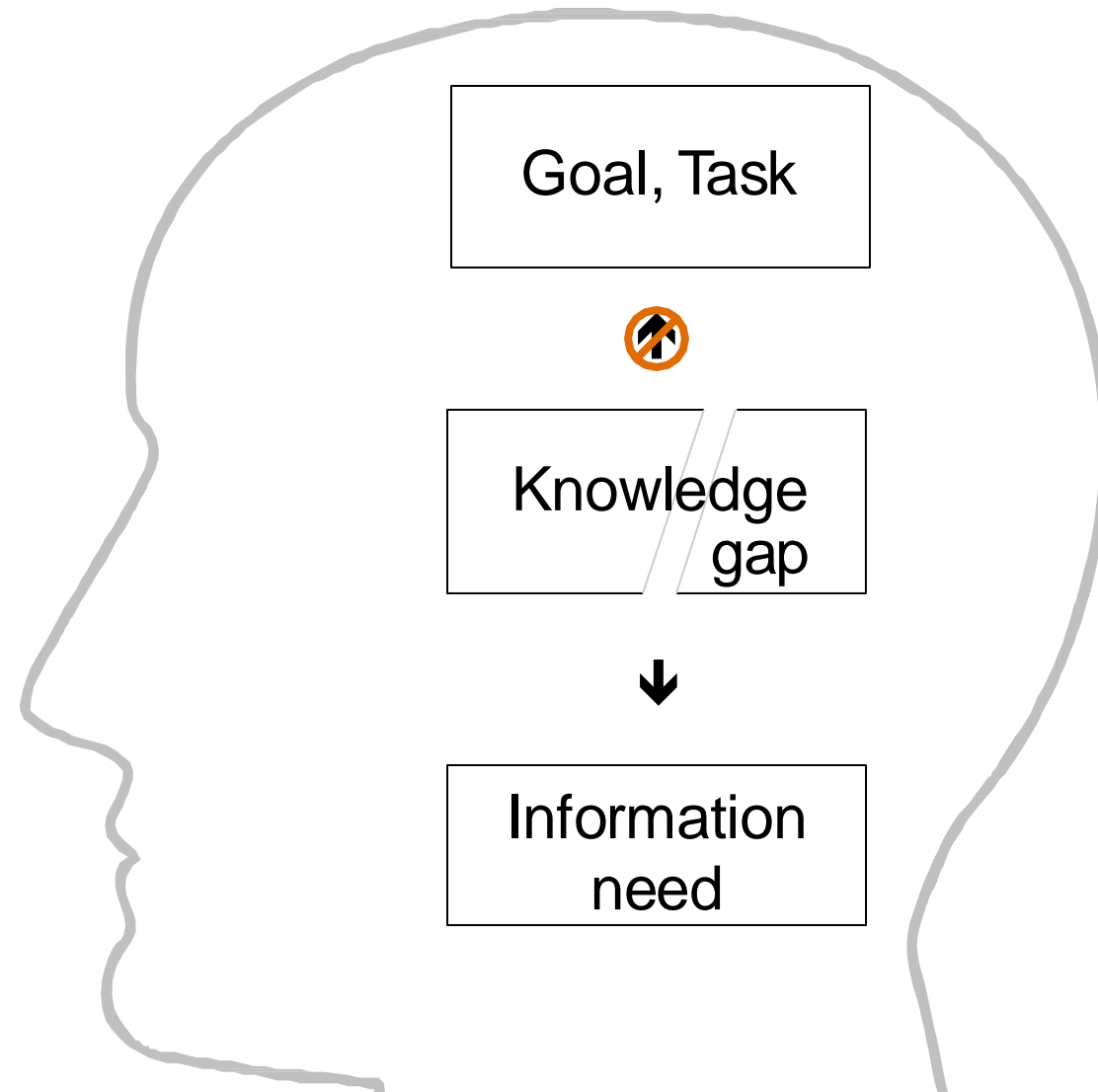
Definition 6 (Relevanz, relevance)

- Der Grad zu dem Ergebnisse das Informationsbedürfnis des Suchenden befriedigen.
- Ein Teil der Daten wird als relevant bezeichnet, wenn er zur Befriedigung eines bestimmten Informationsbedürfnisses (teilweise) nützlich ist. Je näher sie der Befriedigung des Nutzers kommen, desto relevanter sind sie.









Data as documents



Information system

User

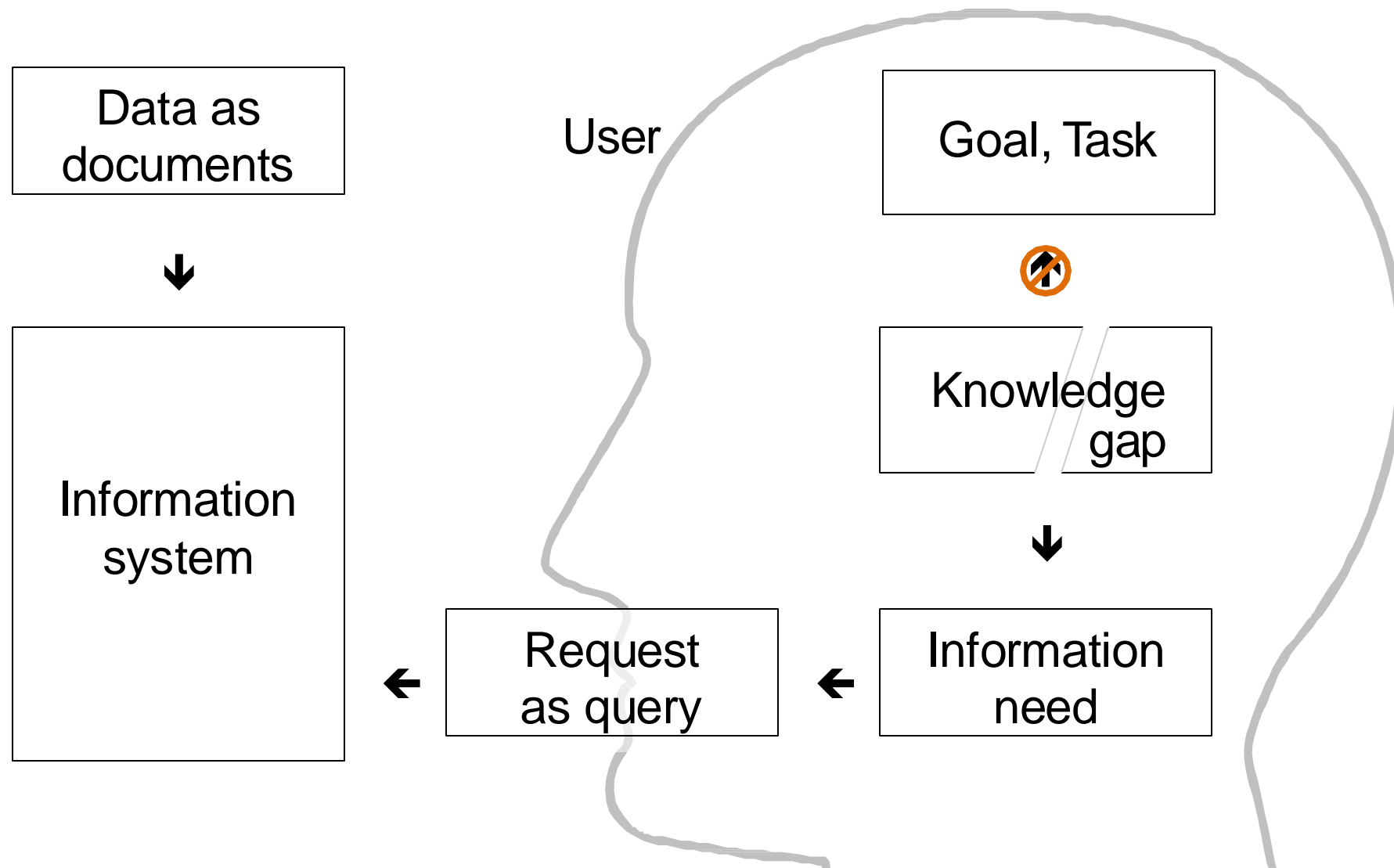
Goal, Task

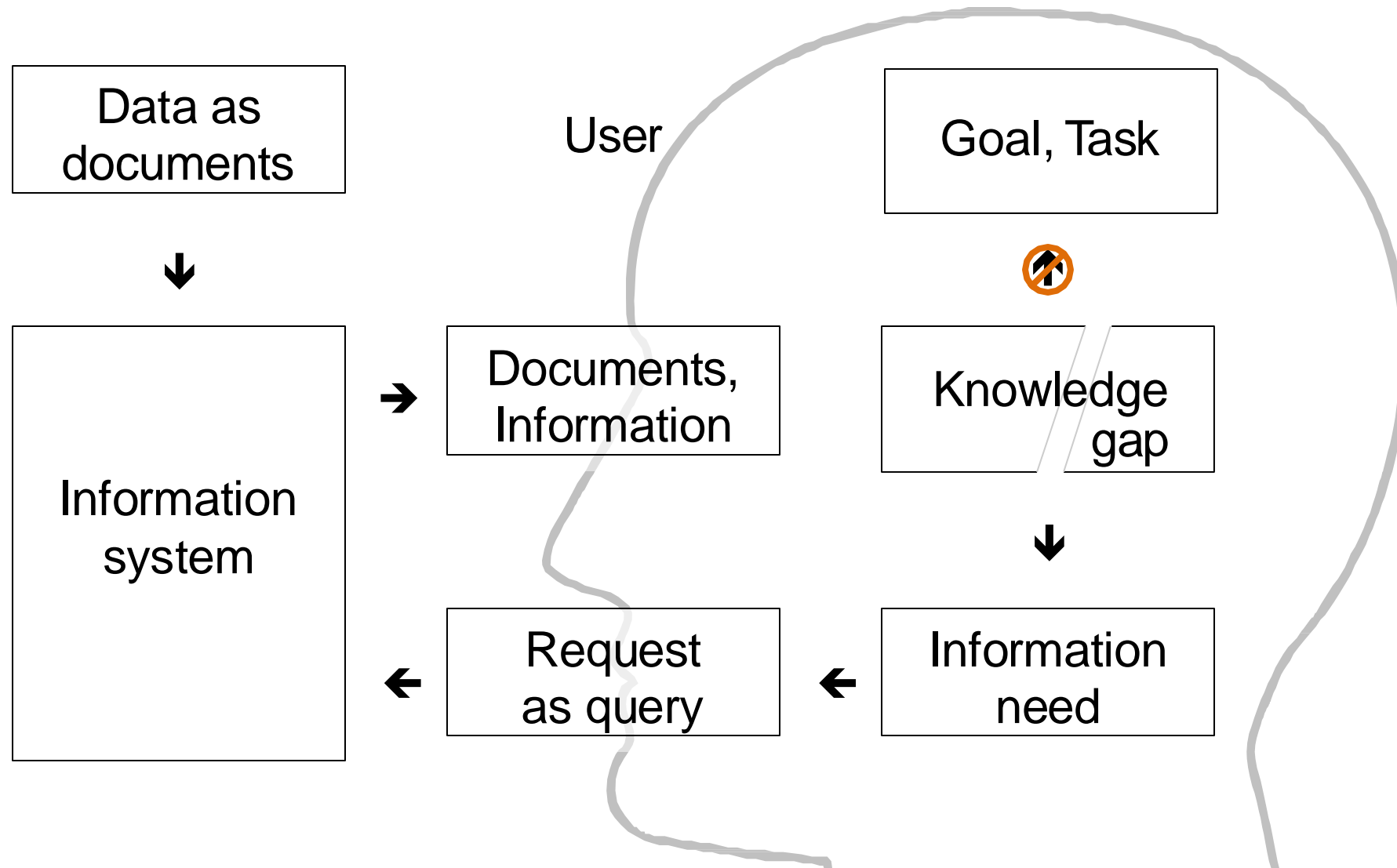


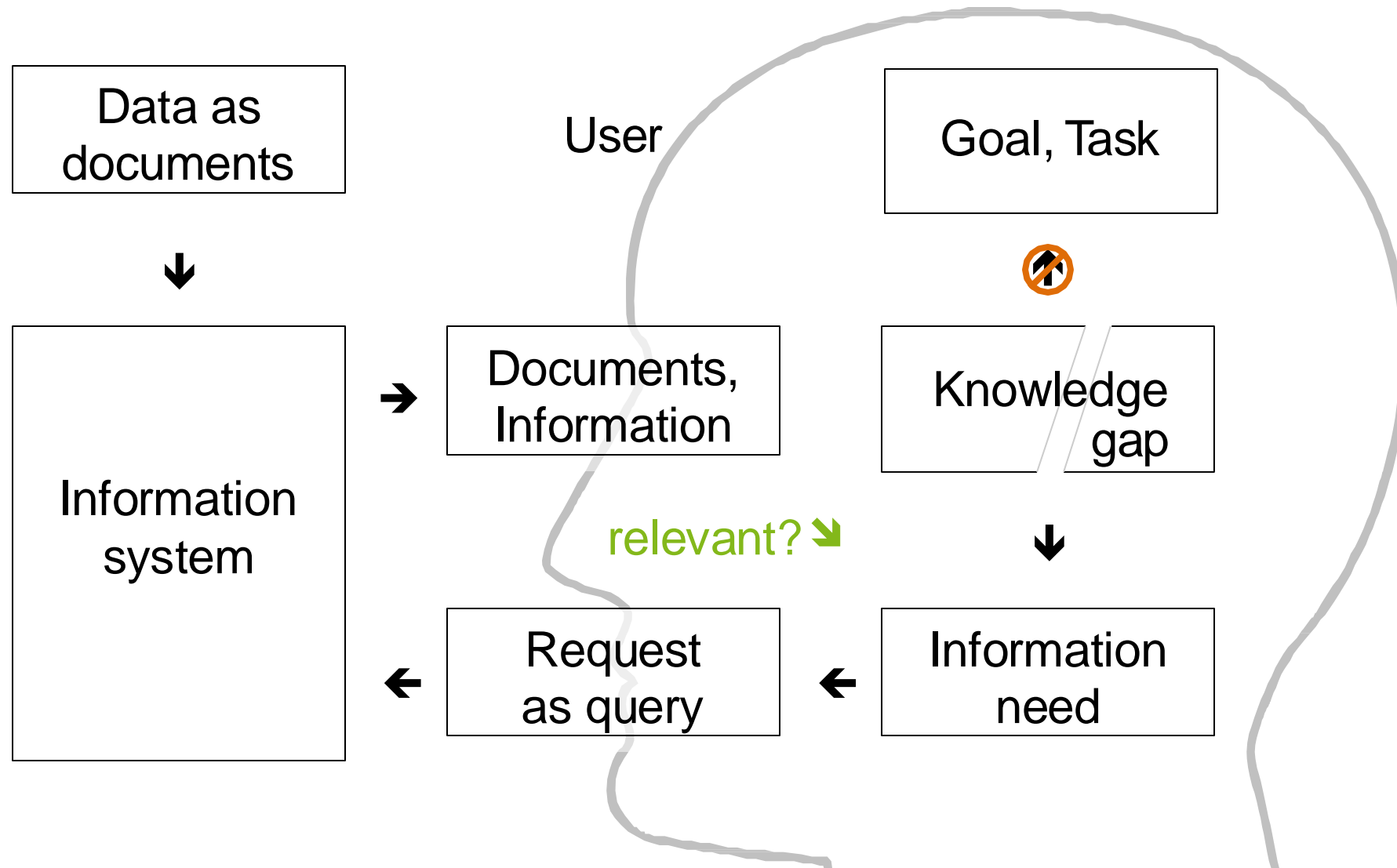
Knowledge gap

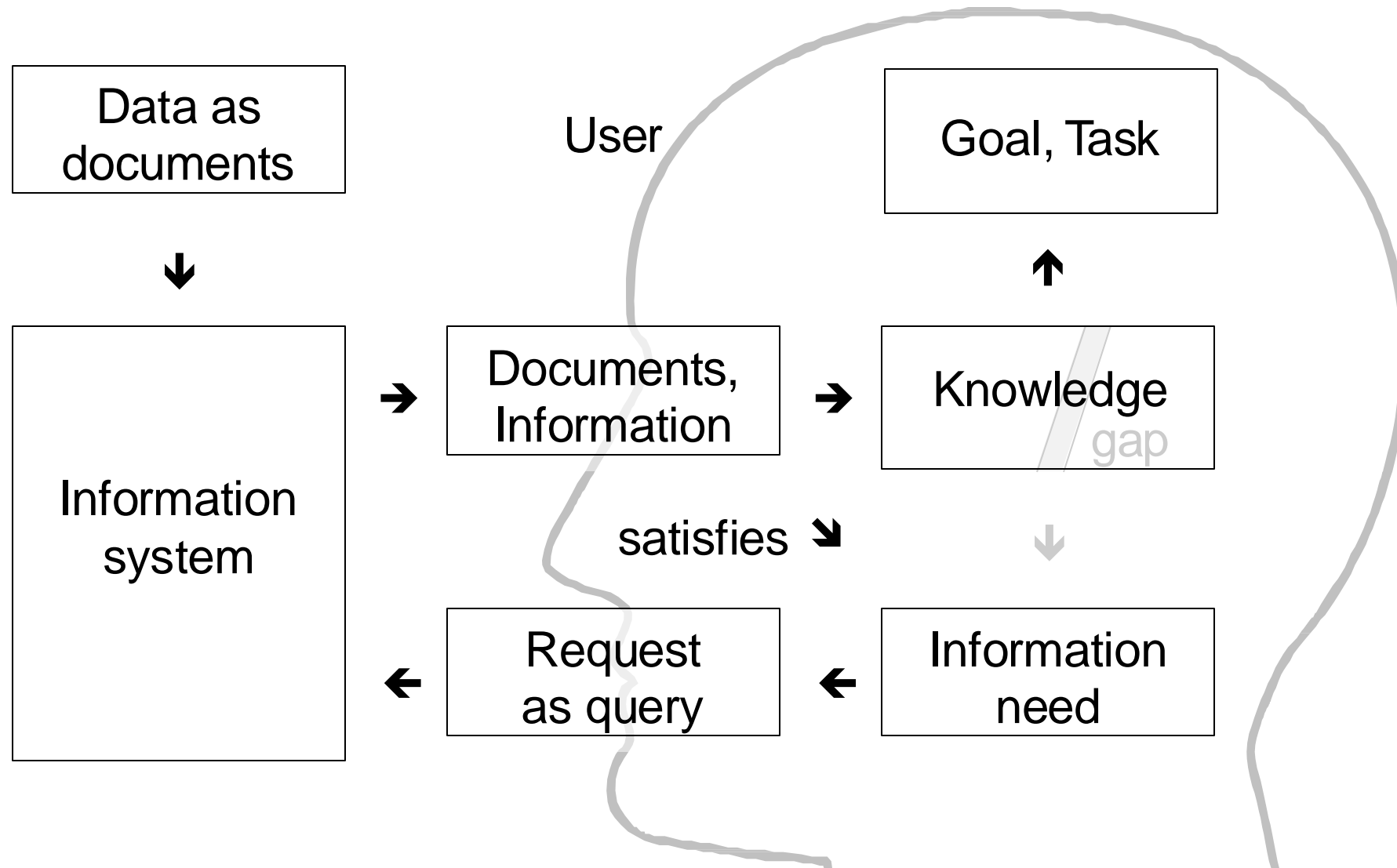


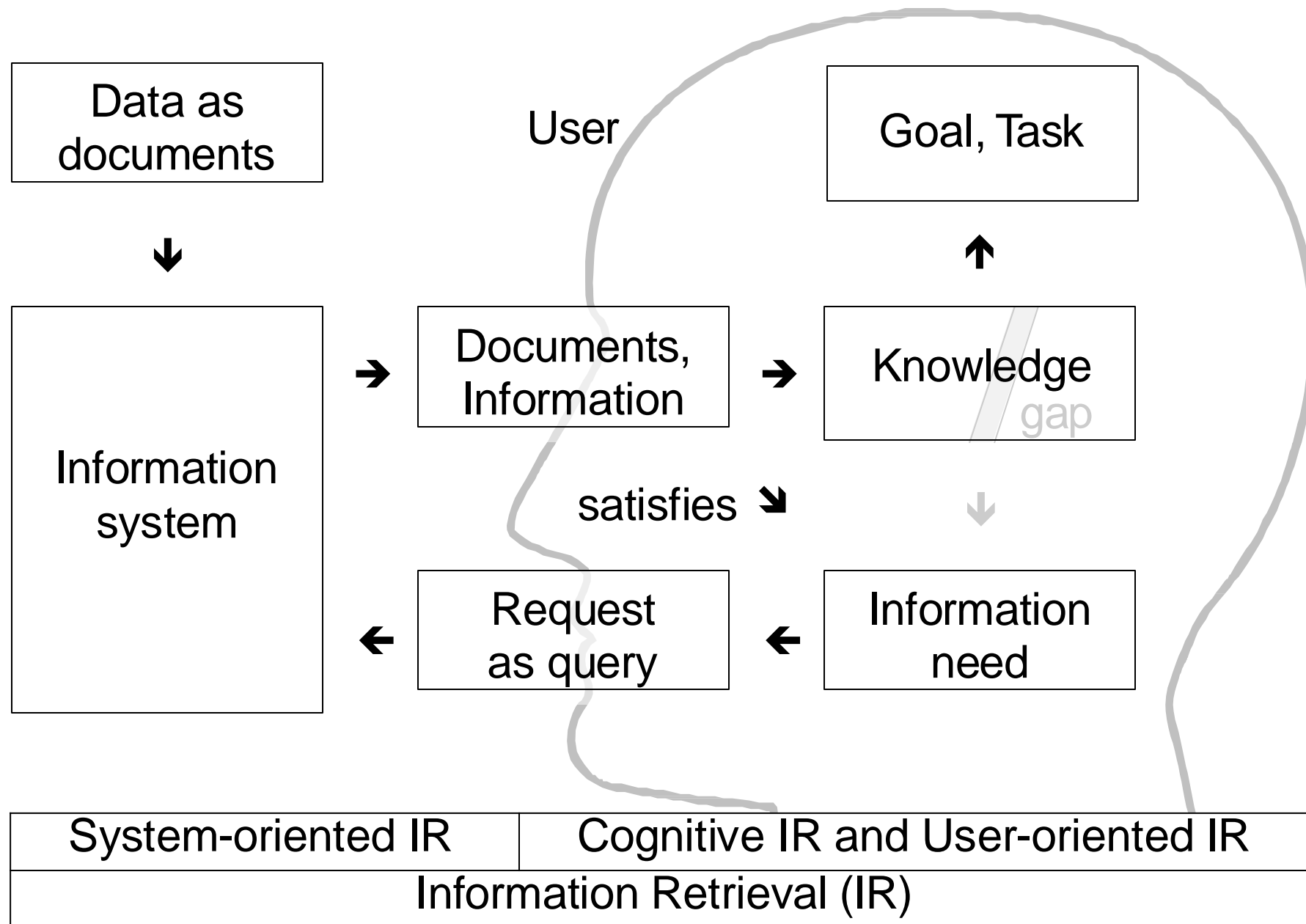
Information need











Was ist also nun Information Retrieval?

Definition 7 (Information Retrieval, IR)

- Beschreibt den Akt der Beschaffung von Informationen aus Daten, die für einen Informationsbedarf relevant sind.
- Als Forschungsgebiet untersucht Information Retrieval die Rolle und Funktion von Informationssystemen bei der Übertragung von Wissen über Daten sowie die Konzeption, Implementierung, Bewertung und Analyse solcher Systeme.

Was ist also nun Information Retrieval?

Definition 7 (Information Retrieval, IR)

- Beschreibt den Akt der Beschaffung von Informationen aus Daten, die für einen Informationsbedarf relevant sind.
- Als Forschungsgebiet untersucht Information Retrieval die Rolle und Funktion von Informationssystemen bei der Übertragung von Wissen über Daten sowie die Konzeption, Implementierung, Bewertung und Analyse solcher Systeme.

Die Probleme dabei sind (u.a.):

- **Vage Anfrage** (ohne klares Ziel, uneinheitliche Sprache, usw.)
- **Unvollständiges und unsicheres Wissen** (z.B. Bias durch politische Einstellungen)
- **Genauigkeit der Ergebnisse und Effektivität**
- **Effizienz**

Weitere Definitionen von IR

Drei weitere Definitionen (von vielen)

- „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).“ (Manning et al., 2008)
- „Information retrieval (IR) is a field concerned with the design, development, and evaluation of interactive systems that help users find information.“ (Arguello, 2017)
- “Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.“ (Salton, 1968)

Weitere Definitionen von IR

Drei Definitionen (von vielen)

Besser!



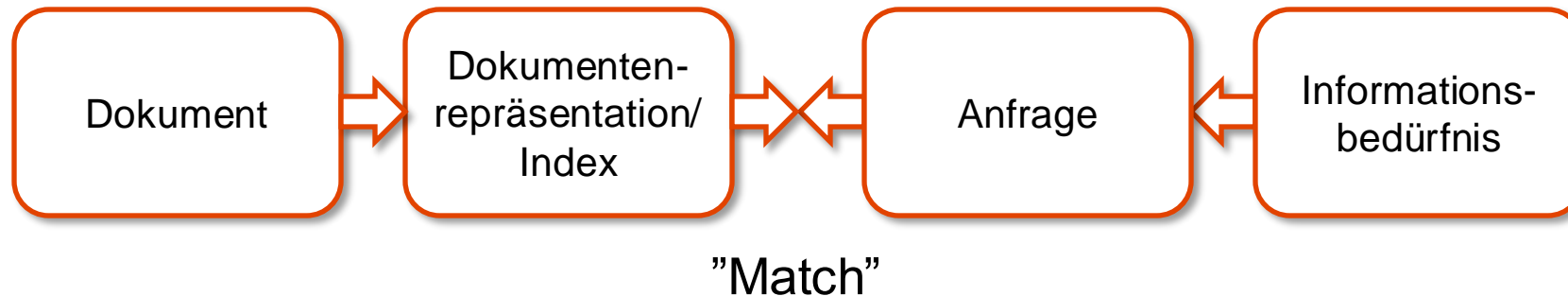
- „Information retrieval (IR) is **finding material** (usually documents) of an **unstructured nature** (usually text) that **satisfies an information need** from within **large collections** (usually stored on computers).“ (Manning et al., 2008)
- „Information retrieval (IR) is a field concerned with the design, development, and evaluation of interactive systems that help users find information.“ (Arguello, 2017)
- “Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.“ (Salton, 1968)

Klassisches Information Retrieval-Modell

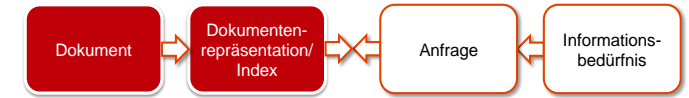
Das klassische **Ad-Hoc-Retrieval** basiert auf Abgleich von

- Dokumenttermen (Document Representation) und
- Anfragetermen (Query).

Im klassischen Information Retrieval-Modell sind das Informationsbedürfnis als auch die Anfrage starr und verändern sich nicht.



Strukturierte Daten

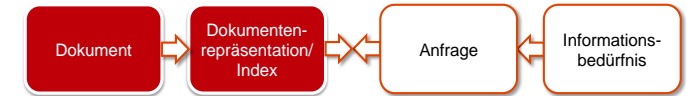


Strukturierte Daten sind z.B. Tabellendaten

Angestellter	Boss	Gehalt
Berthold Heisterkamp	Bernd Stromberg	50000
Ulf Steinke	Bernd Stromberg	60000
Sinan Turçulu	Timo Becker	50000

- Numerische Anfragen und Exact Match sind möglich (wie in SQL), bspw.: ***Gehalt < 60000 AND Boss = Timo Becker***
- Toll, aber meistens nicht das was wir im Information Retrieval vorfinden
→ **Wir suchen in unstrukturierte Daten!**

Unstrukturierte Daten...



Log in / create account

Article Discussion Read Edit View history

Gerard Salton

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of [information retrieval](#) during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in [Applied Mathematics](#) in 1958, the last of [Howard Aiken](#)'s doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval^[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced [TF-IDF](#), or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)^[2].) Later in life, he became interested in automatic text summarization and analysis^[3], as well as automatic hypertext generation^[4]. He published over 150 research articles and 5 books during his life.

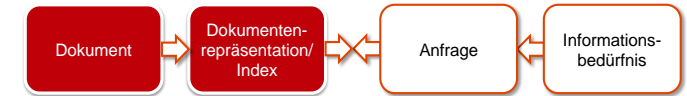
Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired [SIGIR](#). He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an Award of Merit from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

References

[edit]

- [↑] G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing , Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975
- [↑] Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval" , *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526 

Das Dokument ist teilstrukturiert...



Log in / create account

Article Discussion Read Edit View history

Gerard Salton

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of [information retrieval](#) during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in [Applied Mathematics](#) in 1958, the last of [Howard Aiken](#)'s doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval^[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced [TF-IDF](#), or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)^[2].) Later in life, he became interested in automatic text summarization and analysis^[3], as well as automatic hypertext generation^[4]. He published over 150 research articles and 5 books during his life.

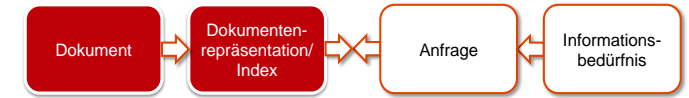
Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired [SIGIR](#). He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an Award of Merit from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award for outstanding contributions to study of information retrieval](#) (1983) -- now called the [Gerard Salton Award](#).

References

[edit]

- [↑] G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing [↗](#), Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975
- [↑] Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval" [↗](#), *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526 [↗](#)

Dokumentenstrukturen

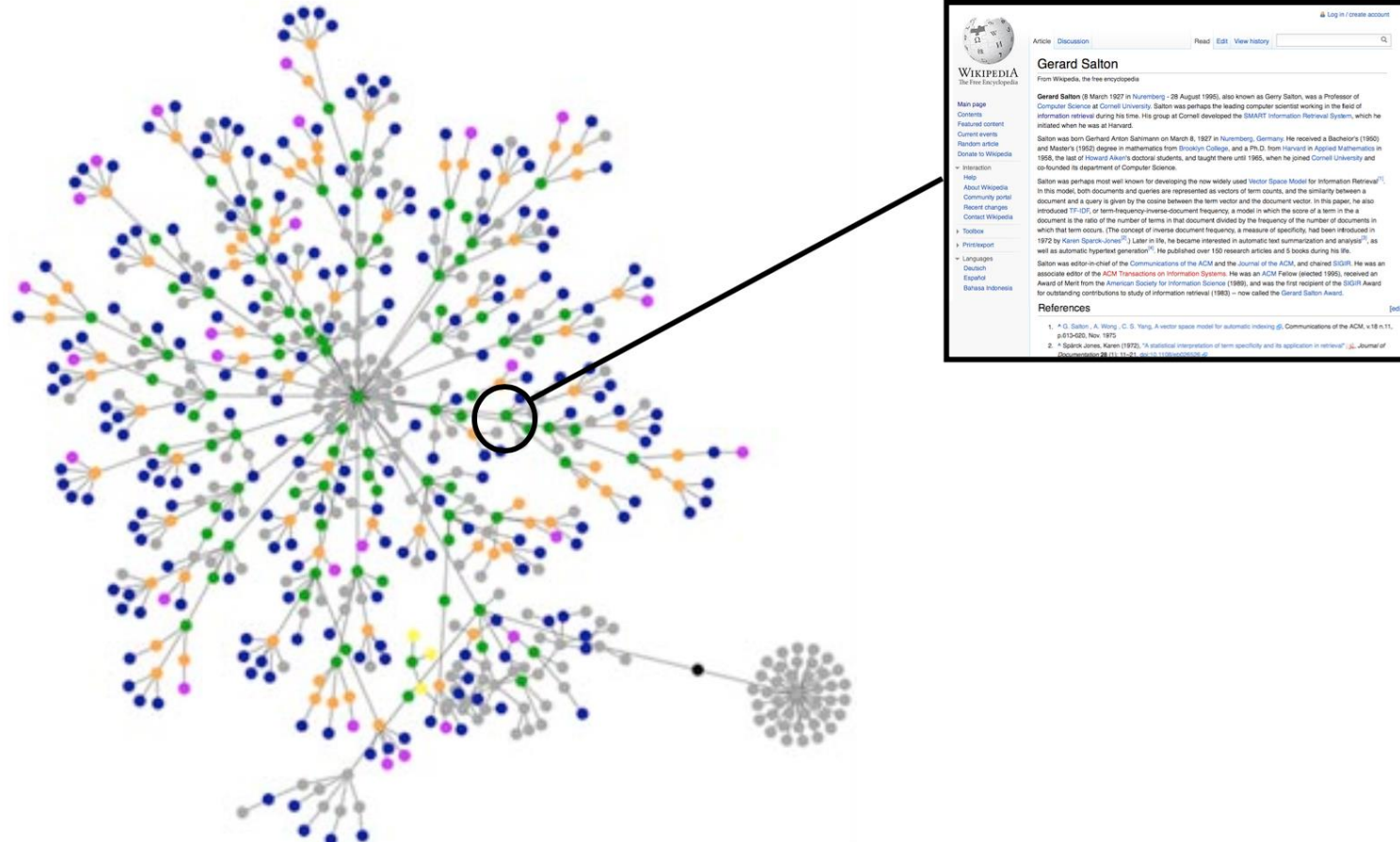
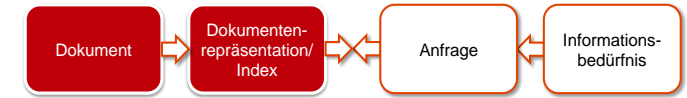


Es gibt zwar eine Struktur, aber der für uns wichtige Teil ist der Text des Artikels. Der Text hat nur **wenig bzw. gar keine Struktur**, die der Computer verstehen könnte.

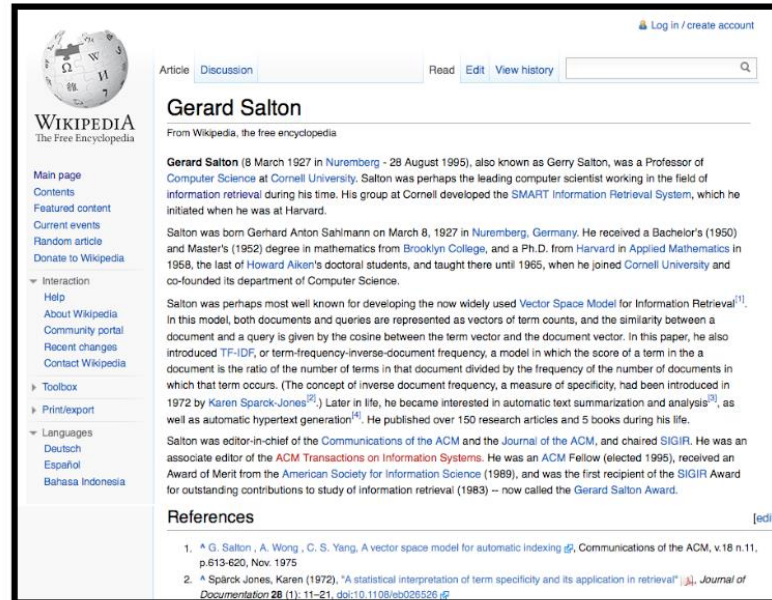
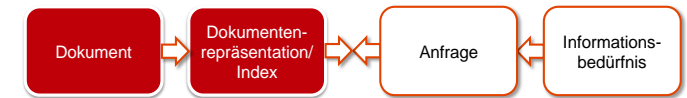
Allerdings wird sich herausstellen, dass es für eine Suchmaschine gar nicht notwendig ist, den Text „zu verstehen“ um eine Anfrage z.B. nach „Gerald Salton“ mit relevanten Ergebnissen zu beantworten. Es reichen uns **unstrukturierte/semi-strukturierte Daten!**

2. ¹ Sparck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation* 28 (1): 11–21, doi:10.1108/jeb028526

Kollektionsstrukturen

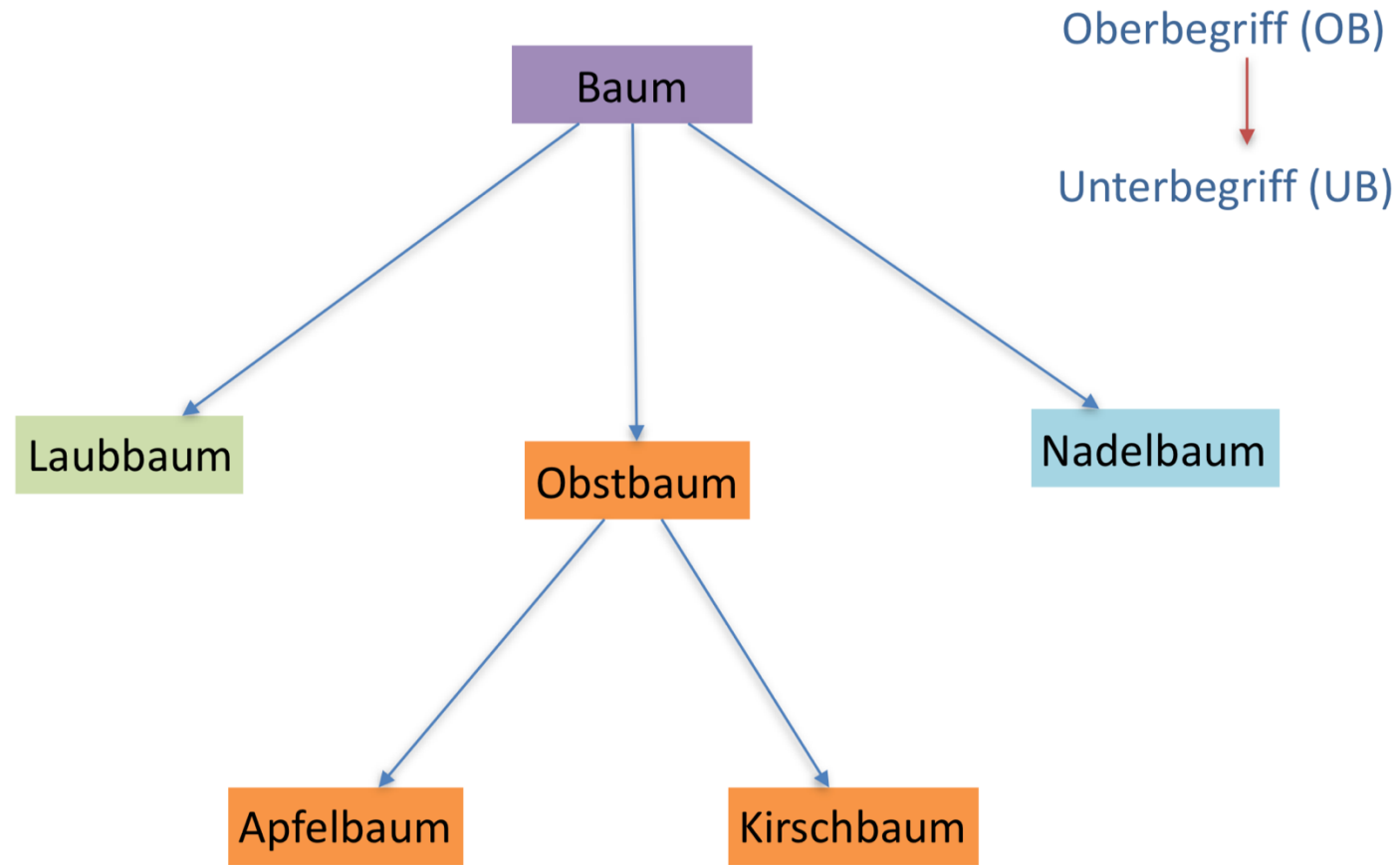
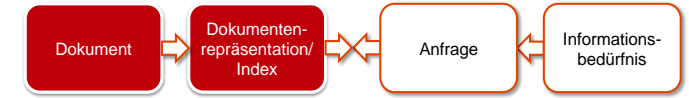


Strukturen durch Klassifikationen

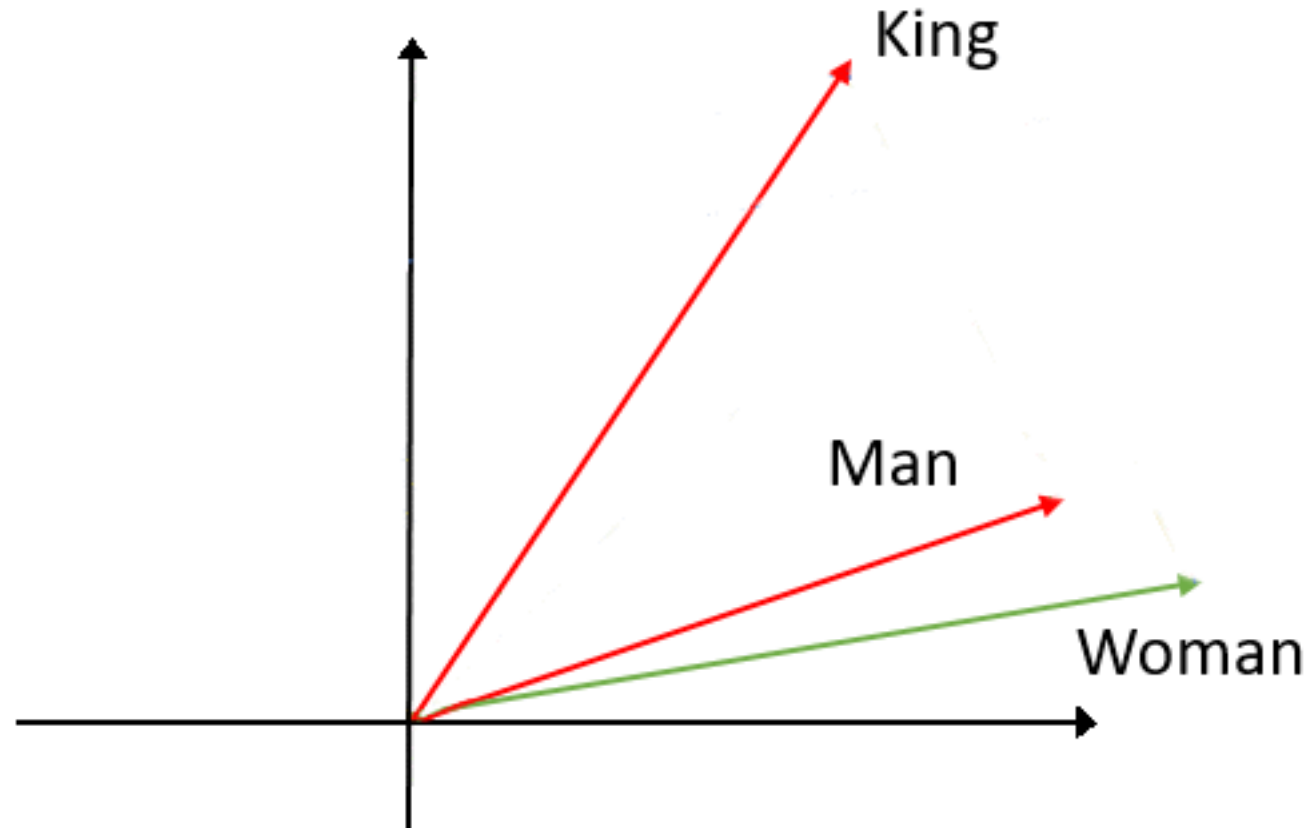
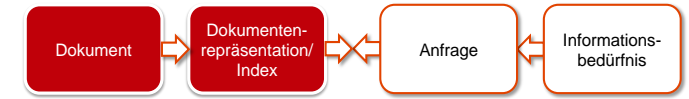


Categories: 1927 births | 1995 deaths | American computer scientists | Computer pioneers | Harvard University alumni | Harvard University faculty | Cornell University faculty | Fellows of the Association for Computing Machinery | Guggenheim Fellows

Semantische Strukturen durch Konvention

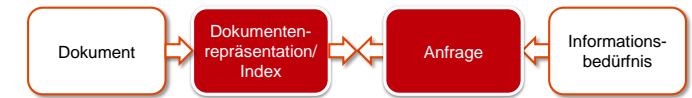


Semantik durch statistische Analyse



$$X = woman + king - man \approx queen$$

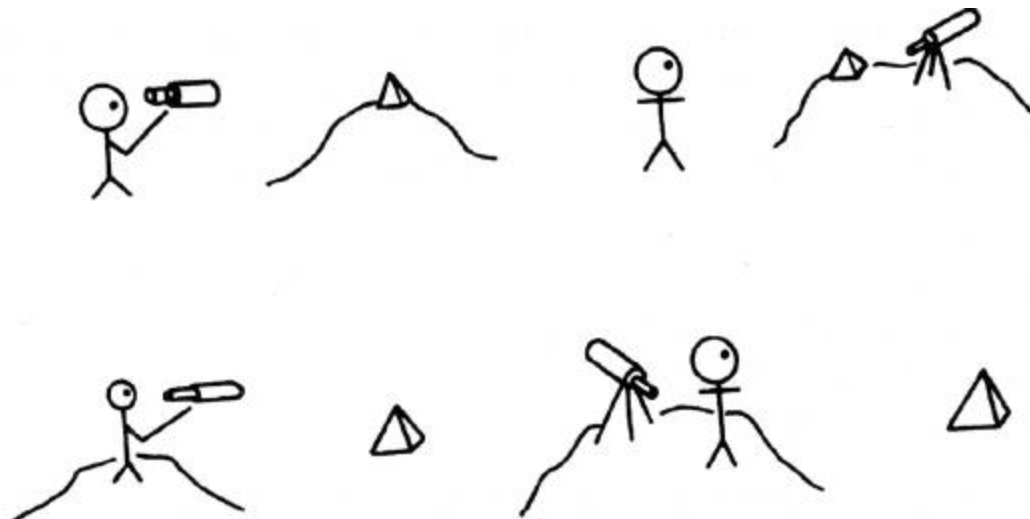
Was ist eigentlich Relevanz?



Schwierig, da Anfrage gerne sehr vage sind.

„The man saw the pyramid on the hill with the telescope.“

- Viele Interpretationen dieses Satzes sind denkbar...



Welches der Bilder ist relevant?

Anfragen sind nur eine Krücke



Eine Suchanfrage ist nur eine verkürzte Beschreibung des **Informationsbedürfnisses** des Nutzers.

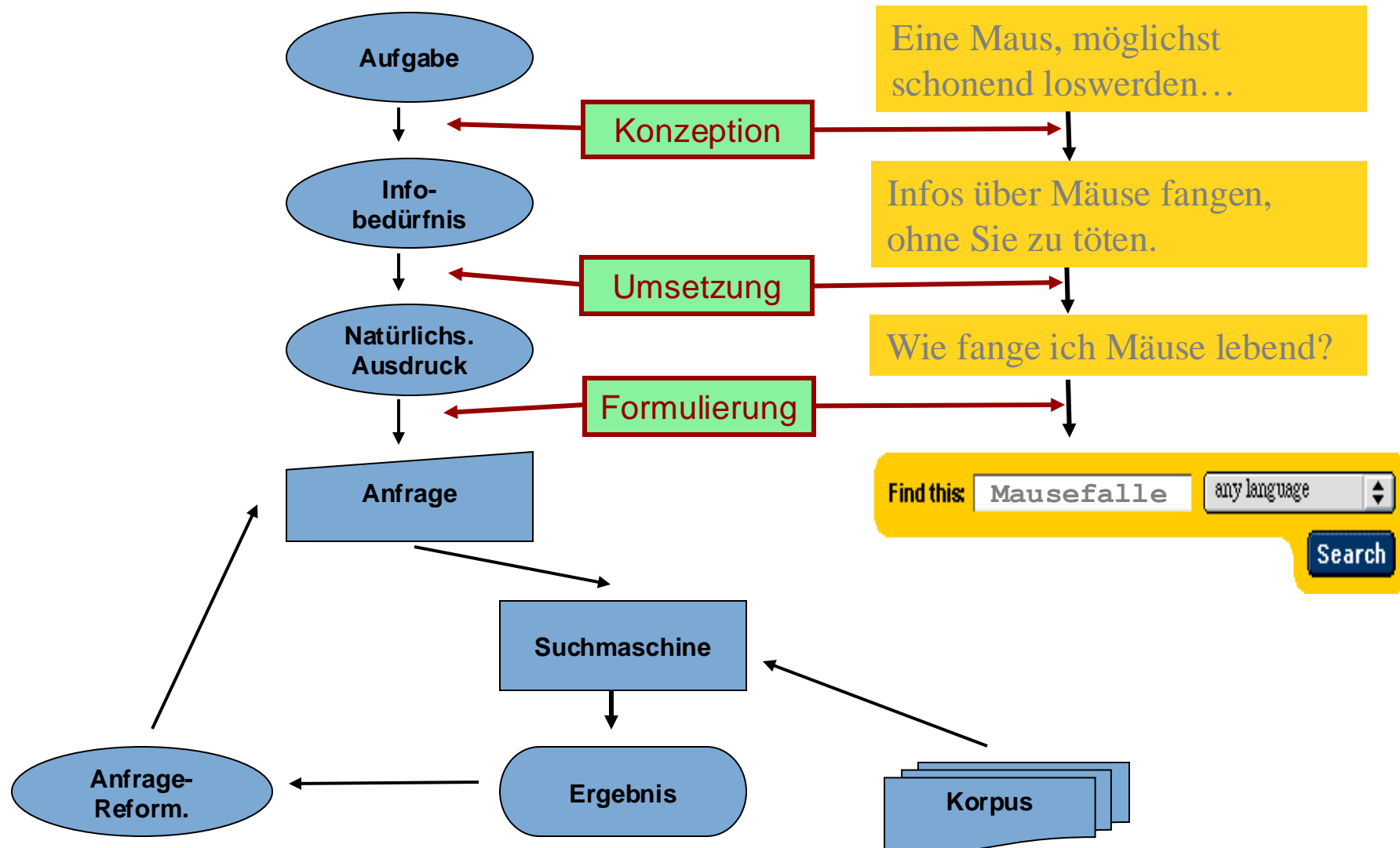
- Mehrdeutig und eigentlich nur für den Nutzer selbst verständlich
- Eine Anfrage und ein Informationsbedürfnis setzt eine konkrete Aufgabe voraus
- Der Nutzer befindet sich in einem konkreten Kontext, den wir nicht kennen

AOL-Logdateien – User 1515830

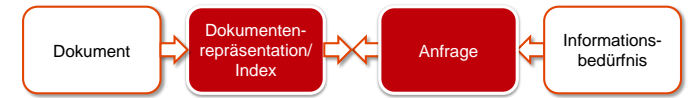


chai tea calories
calories in bananas
aftermath of incest
how to tell your family you're a victim of incest
pottery barn
curtains
surgical help for depression
oakland raiders comforter set
can you adopt after a suicide attempt
who is not allowed to adopt
i hate men
medication to enhance female desire
jobs in denver colorado
teaching positions in denver colorado
how long will the swelling last after my tummy tuck
divorce laws in ohio
free remote keyloggers
baked macaroni and cheese with sour cream
how to deal with anger
teaching jobs with the denver school system
marriage counseling tips
anti psychotic drugs

Anfragen sind nur eine Krücke



Wie gut ist unsere Suchergebnis?

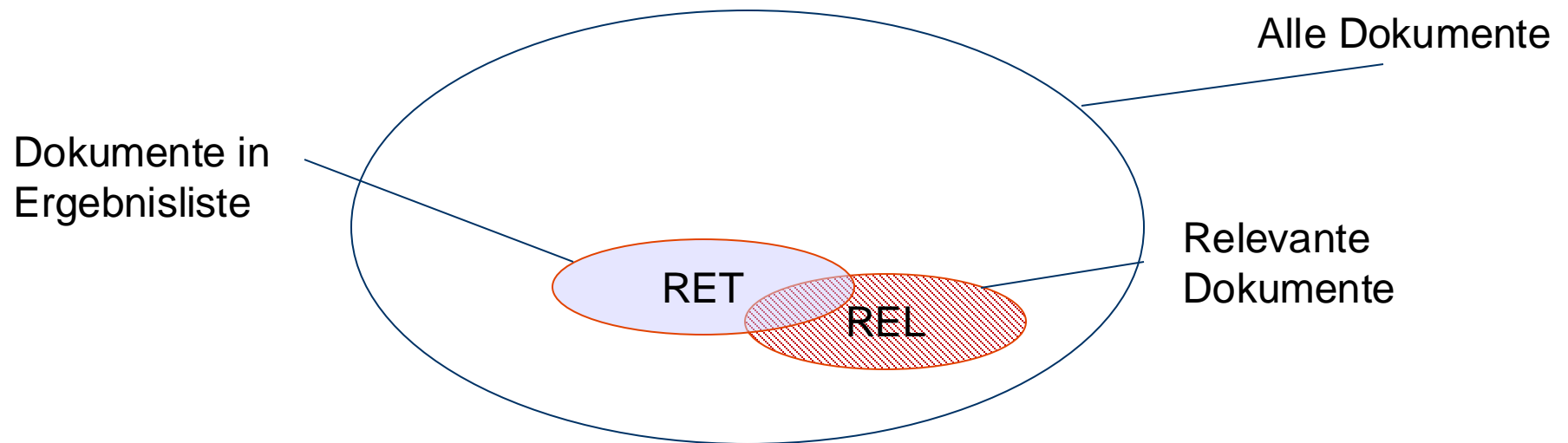


- **Precision** (Treffergenauigkeit)

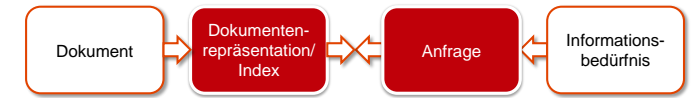
$$\mathcal{P} = \frac{|\text{RET} \cap \text{REL}|}{|\text{RET}|}$$

- **Recall** (Treffervollständigkeit)

$$\mathcal{R} = \frac{|\text{RET} \cap \text{REL}|}{|\text{REL}|}$$



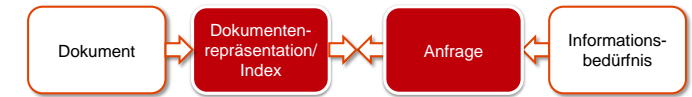
Precision und Recall: Ein Beispiel



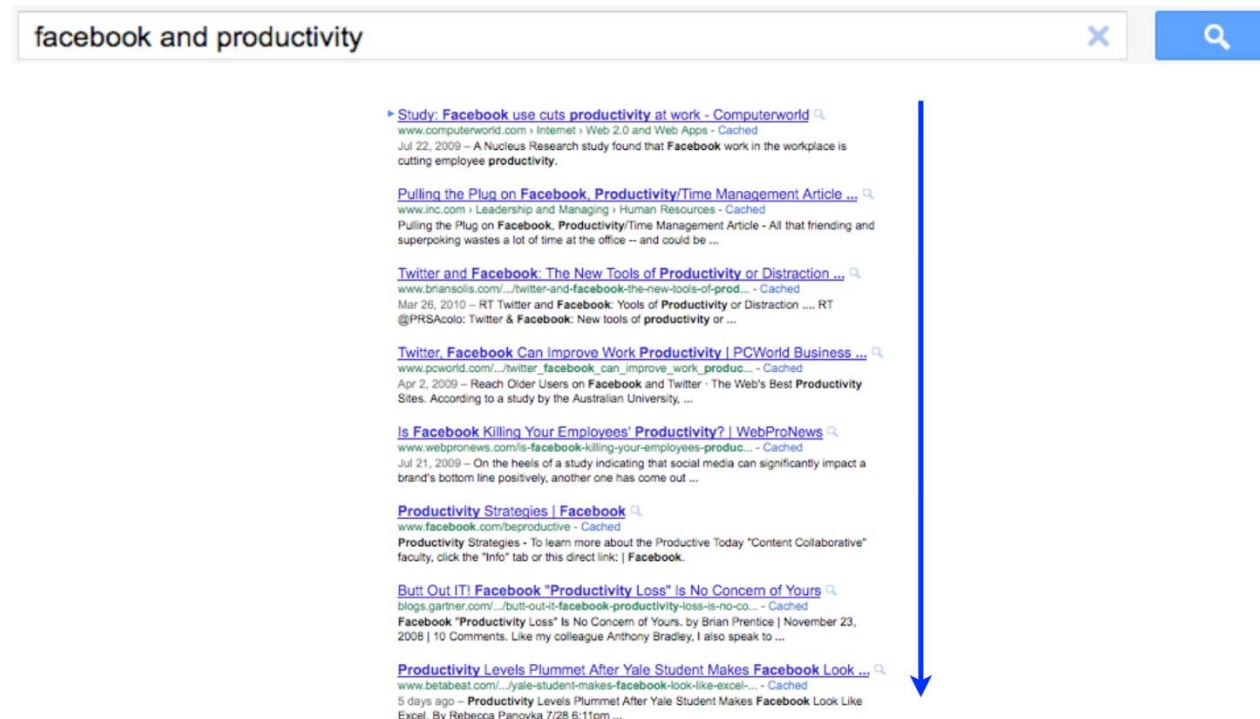
	Relevant	Nicht relevant
Gefunden	30	12
Nicht gefunden	14	44

- Precision $P = 30 / (30 + 12) \approx 0,714$
- Recall $R = 30 / (30 + 14) \approx 0,681$
- Precision und Recall sind Mengen-basierte Maße!
 - Die Reihenfolge der Ergebnisse interessiert also nicht.

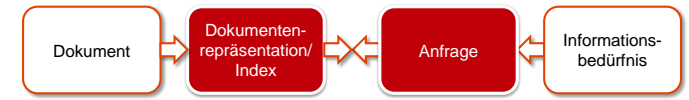
Retrieval als Suchaufgabe



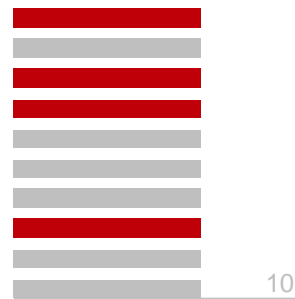
- **Ausgabe:** Ein **Ranking** von Dokumenten, in absteigender Reihenfolge Ihrer geschätzten Relevanz (macht es einfacher!).
- **Annahme:** Der Benutzer schaut sich die **ersten paar Dokument** an und ist zufrieden, wenn er etwas Passendes gefunden hat.



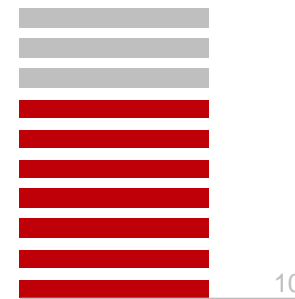
Ranking – Welches ist besser?



System A

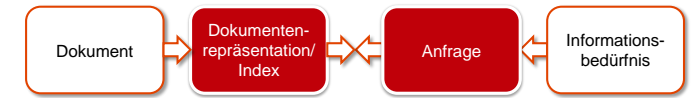


System B

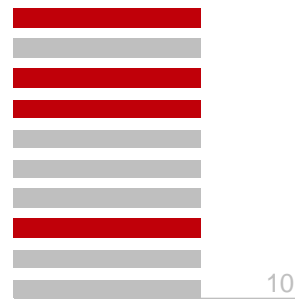


- Relevante Treffer sind rot markiert.

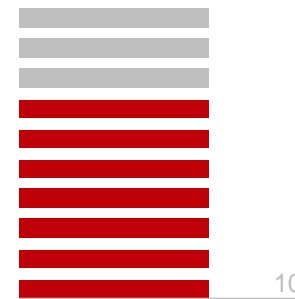
Ranking – Welches ist besser?



System A



System B



- Precision kann z.B. auch auf Listen angewendet werden.
- $P@n \rightarrow$ Precision an der Position n

$P@3$ für System A: 2/3

$P@3$ für System B: 0/3

Warum ist IR also eine schwierige Aufgabe?

Information Retrieval ist ein Prozess mit **Unsicherheiten**...

- Benutzer **wissen nicht was sie eigentlich wollen**
- Benutzer wissen nicht, wie sie das was sie suchen **ausdrücken sollen**
- Computer können Nutzer keine **Kontextinformationen** entlocken, wie es z.B. ein menschlicher Bibliothekar könnte
- Computer verstehen **keine natürliche Sprache**
- Suchmaschinen müssen **erraten, was relevant ist**
- Suchmaschinen müssen erraten, wann ein Benutzer **zufrieden** ist
- ...

Suchanfragen, Relevanz, Ranking...

Dies alles macht Information Retrieval so **schwer** – und **faszinierend**!

- Eigentlich müssten wir viel mehr „verstehen“, was der Nutzer will.

Im IR geht es aber nicht darum Text zu verstehen, sondern darum „zu bestimmen“/„zu berechnen“/„vorherzusagen“ ob ein Dokument zu einer Anfrage relevant ist – oder nicht.

- Das ist in vielen Fälle **viel einfacher**, als „verstehen“.
- Es ermöglicht uns, das Problem an **Computer abzugeben**, z.B. in Form von Suchmaschinen!
- Computer arbeiten mit **Modellen**, um Informationen zu finden, z.B. dem **booleschen Modell**...
- **Reicht aber oft genug aus!**

Zusammenfassung

Das Ziel des Information Retrieval ist es Suchenden die **passenden Informationen** zu liefern, die sie **benötigen!**

- IR beinhaltet die **Analyse, Speicherung** und die eigentliche **Suche**
- **Verschieden Arten** von Suchmaschinen
- **Unwägbarkeiten** und **Unsicherheiten**
- Abhängig von **Aufgabe** und **Kontext** erwarten Nutzende ganz Unterschiedliches
- **IR-Evaluation** ist komplex und setzt ein Verständnis des Nutzerbedürfnisses voraus

