

› Web Information Retrieval Introduction

Frank Hopfgartner
Institute for Web Science and Technologies

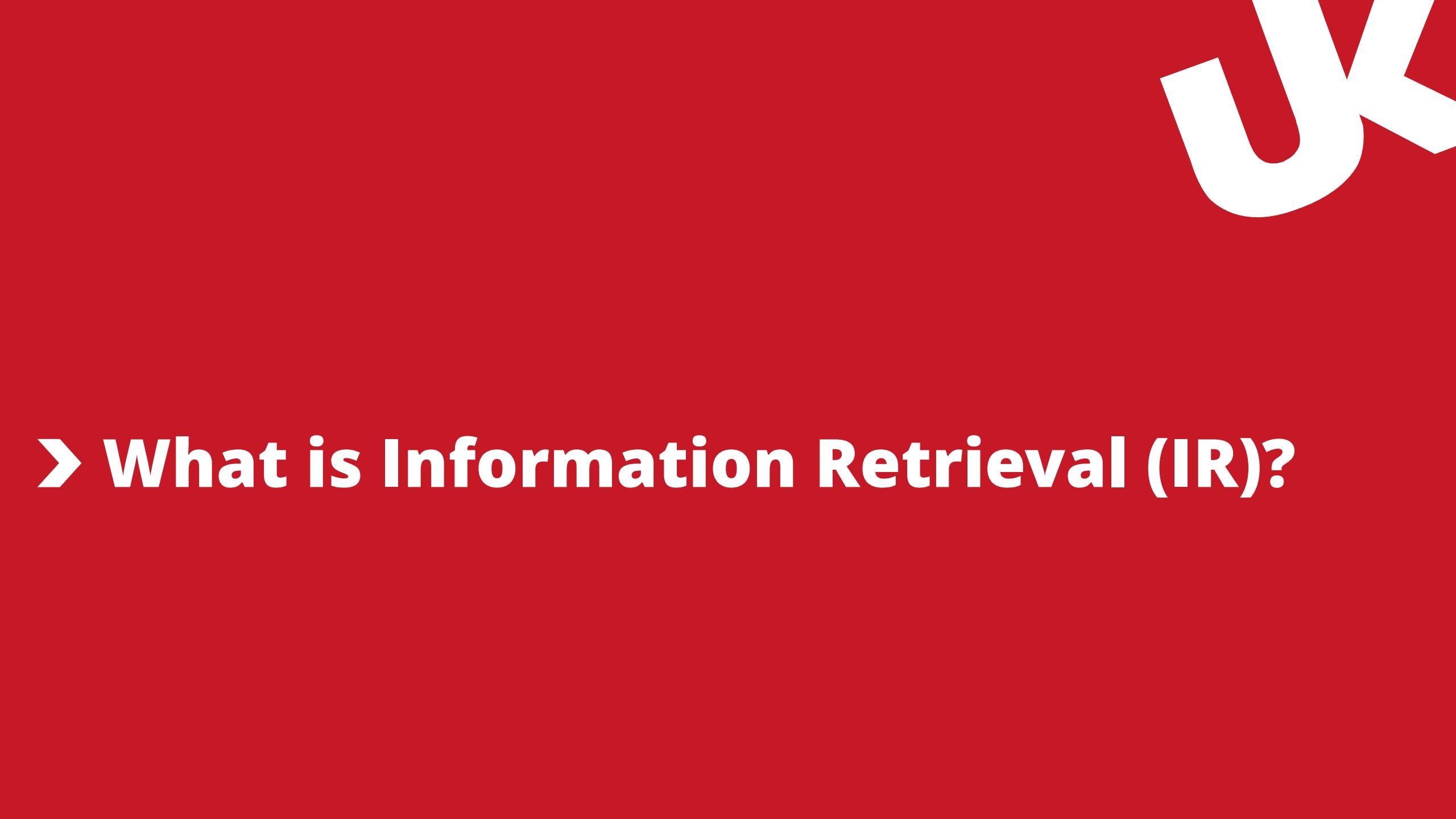
Intended Learning Outcomes

By the end of the lecture you should be able to

- Describe what is meant by IR
- Recognise that IR systems are used to support the broader information seeking activities of users
- Describe the main components of an IR system

Outline

- What is Web Information Retrieval
- Information Retrieval – A definition
- Information Retrieval in Context
- How search systems work (briefly)
- History of IR



➤ What is Information Retrieval (IR)?

Information Retrieval

IR systems support users' searching and seeking activities



Photo credit:

www.flickr.com/people/pfala/

An IR example



- One solution is to go through all documents and read through all the text searching for the query string
 - the computer (e.g. *grep* command in Linux) can perform this process
 - *grep* stands for *global regular expression print*
 - this process is very effective

An IR example

- But
 - the volume of online data compared to the speed of today's computers does not allow all data to be processed in a reasonable time
 - this process retrieves only exact matches and does not allow flexible queries
 - this process can retrieve a lot of irrelevant results that do contain the query string
 - this process cannot rank the retrieval results

How people find information

IR systems are just one way people find information

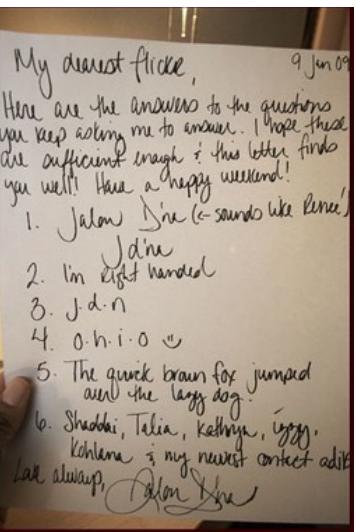
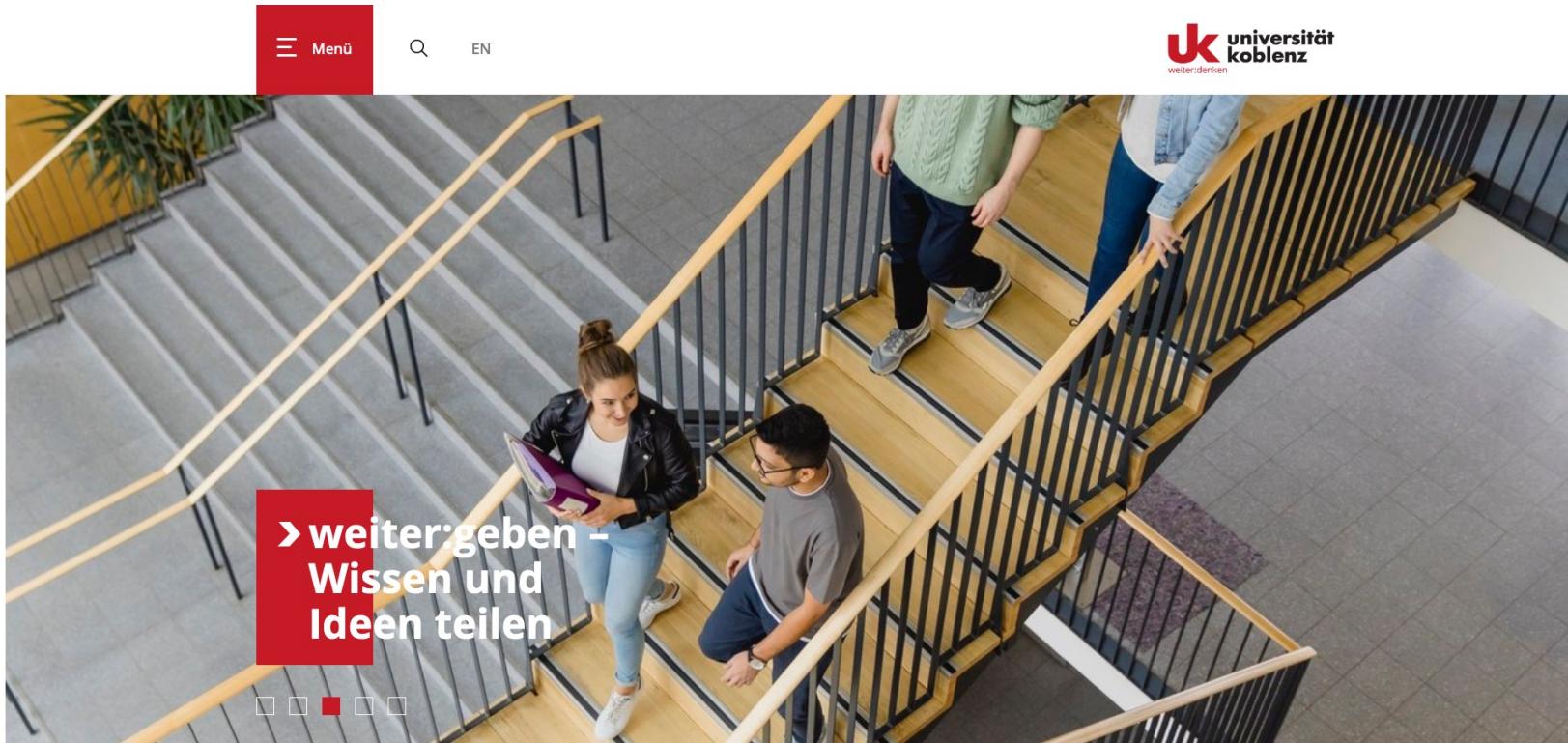


Photo credits:
www.flickr.com/photos/robfenwick/
www.flickr.com/photos/jalon_dna
www.flickr.com/people/dinnerseries/

How people find information

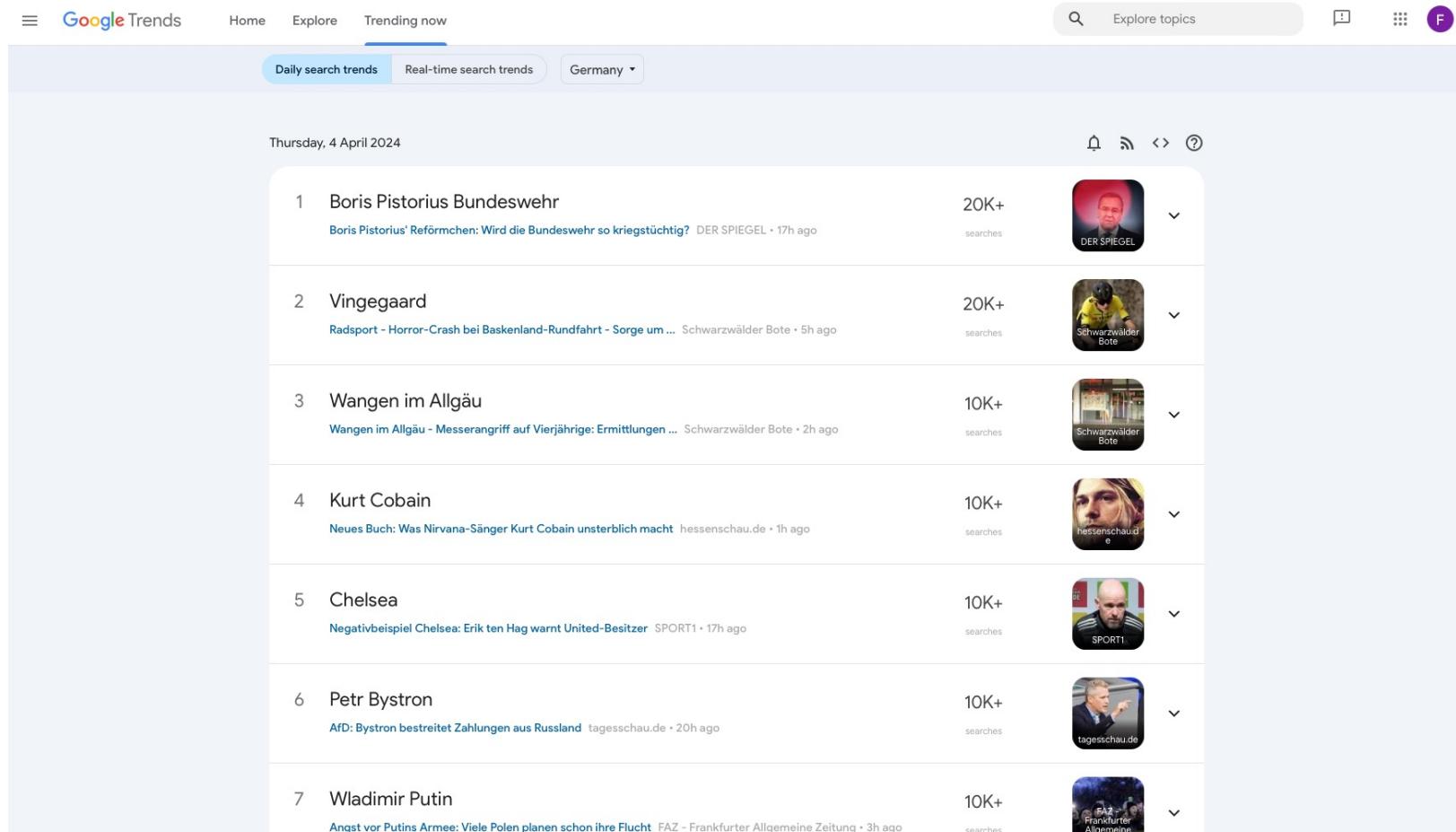
People can browse, navigate or query



Newsroom



What is Germany searching for (on Google)



What IR systems are used for

People use IR systems for a variety of information needs

- Question 1: In the past week, what information retrieval (search) systems have you used?
- Question 2: In the past week, what have you searched for?



› Information Retrieval - A Definition

A definition

“Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections (usually stored on computers)”
Manning, Raghavan & Schütze (2009)

A definition

“Information retrieval (IR) is **finding material** ... of an unstructured nature ... that satisfies an information need from within **large collections** (usually stored on computers)”
Manning, Raghavan & Schütze (2009)

"finding materials ... in large collections"

Google Scholar search results for "information retrieval". The interface shows a sidebar with filters like "Articles", "Books", and "Conference papers". A detailed search bar includes dropdowns for "Any time", "Since 2019", "Review", "Since 2015", and "Creative Commons". The main area displays a grid of search results with titles, authors, and publication details.

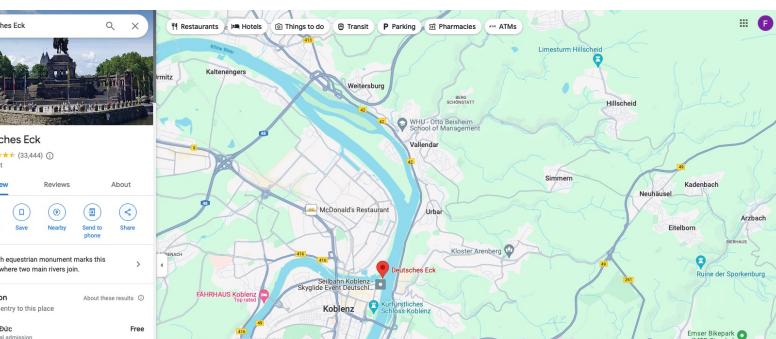
Amazon homepage featuring a banner for the new movie "Road House". Below it, there are several promotional boxes for "Top Deal", "UEFA Champions League", "New to Amazon?", and "Sign in for your best experience". Other sections include "Electronics", "Fashion", "Sports & Outdoors", and "Find gifts for everyone".

YouTube search results for "asmr". The page shows a grid of video thumbnails with titles like "NEW ASMR SOUNDS + mini trigger test with Poki ❤ Sponges, Airpods, Tapping, Whispering, etc!", "ASMR | Layered Sounds // Whispering", and "Air Conditioner - 10 hours of relaxing ambient sounds asmr". Each video has its title, view count, and upload date.

iStock photo search results for "Deutsche Eck". The page features a grid of landscape photographs of the Deutsches Eck monument in Koblenz, Germany, along with other travel-related images.

allrecipes.com search results for "Victoria sponge". The page includes a navigation bar with categories like "RECIPES", "HOW TO", "Dish type", "Ingredients", "Method", "Cuisines", "Diet & lifestyle", and "Popular". The search results show a list of recipes with titles like "Traditional Victoria sponge" and "Victoria sponge", each accompanied by a small image, reviews, and cooking time.

Spotify profile for Beyoncé. The page displays her top tracks: "Irreplaceable", "Halo", "Love On Top", "If I Were a Boy", and "Crazy in Love (feat. Jay Z)". Each track is shown with its name, artist, album, and duration.



A definition

“Information retrieval (IR) is finding material ... of an **unstructured nature** ... that satisfies an information need from within large collections (usually stored on computers)”
Manning, Raghavan & Schütze (2009)

“unstructured nature”

- IR systems can deal with unstructured information where the meaning of the information is not made clear (semantics not explicitly defined)
- IR systems can deal with incomplete (and vague) specifications of the user's query

IR system

The screenshot shows a search bar with the query "university of koblenz professors". Below the search bar are navigation links: ALL, COPILOT, MAPS, IMAGES, VIDEOS, NEWS, SHOPPING, MORE, and TOOLS. There are also microphone, camera, and search icons. Below these are three blue-highlighted suggestions: "university of koblenz ranking", "university of koblenz & landau", and "university of koblenz acce...". A message indicates "About 2.700.000 results". At the bottom is a map titled "University Of Koblenz Professors" showing Coblenz and surrounding areas with various roads labeled (E31, A61, E44, L52, B9, L127, B261, etc.).

Database

EMP-NO	NAME	GRADE	HOURS	OTIME
.....				
0023	Jürjens	IV	36	0
0024	Hopfgartner	IV	20	0
0025	Kschischo	III	36	12
0026	Müller	II	20	4
0027	Lämmel	IV	36	2
0028	Frey	II	20	0
....				

“unstructured nature”

What is unstructured data?

- “it refers to data which does not have clear, semantically overt, easy-for-a-computer structure.” (Schütze, Manning, Raghavan, 2008)



Structured vs. Unstructured Data

	Structured data	Unstructured data
<i>Retrieval system</i>	Data retrieval	Information retrieval
<i>Canonical example</i>	Relational database	Collection of documents
<i>Result</i>	Very precise and always correct.	Relevance varies
<i>Interaction</i>	One shot query	Interaction is important
<i>Type</i>	Text only	Not limited to one type
<i>Amount [6]</i>	20% of enterprise data (2017)	80% of enterprise data (2017)
<i>Volume</i>	Less storage is required	More storage is required
<i>Retrieval velocity</i>	Relatively high	Relatively low
<i>Scalability</i>	Difficult	Highly scalable
<i>Accessibility</i>	Easy	Hard
<i>Schema</i>	Dependent	Free of

A definition

“Information retrieval (IR) is finding material ... of an unstructured nature ... that **satisfies an information need** from within large collections (usually stored on computers)”
Manning, Raghavan & Schütze (2009)

“satisfies an information need”

- IR systems are purposeful devices
- Information need usually expressed as a query
- Satisfied if IR system returns documents relevant to the information need

Queries (expression of information need)

- Queries to an IR system are expressed in natural language (or 'text-free')
- People often use short and ambiguous queries

Returns relevant documents

- *Relevance* is of central importance in IR
- Relevance refers to whether the retrieved information meet the expected user need
- IR systems aim to *rank* information objects according to the degree of relevance to the user query

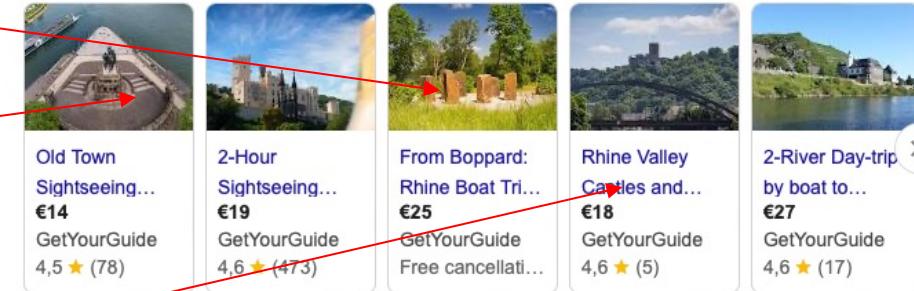
When is a result relevant?

- Topical relevance is most often considered, but there are many more criteria
 - I'll search for 'Koblenz tourist destinations"

About 10.300.000 results (0,41 seconds)

Results for Koblenz · Choose area :

Sponsored · Tickets & tours :



Actually this is
another town, not
Koblenz

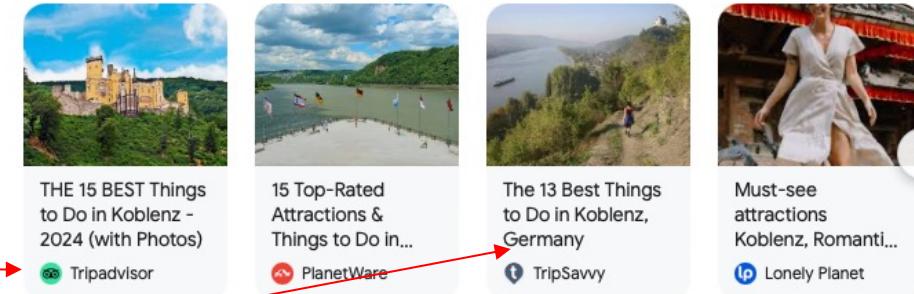
I've already been
there

Need a car - I
have one so
that's ok

I've used this
site before - I
trust it

Sounds a bit
whacky - I'll skip
this

Places sites :



See more →

Top sights in Koblenz :



How do you decide relevance?

Do a Google search (or think about the last search you did). Try different searches – do your criteria change?

- Question 3: How did you decide which result to select, i.e., why was the result relevant?
- Question 4: Why did you not select some results, i.e., why was the result not relevant?

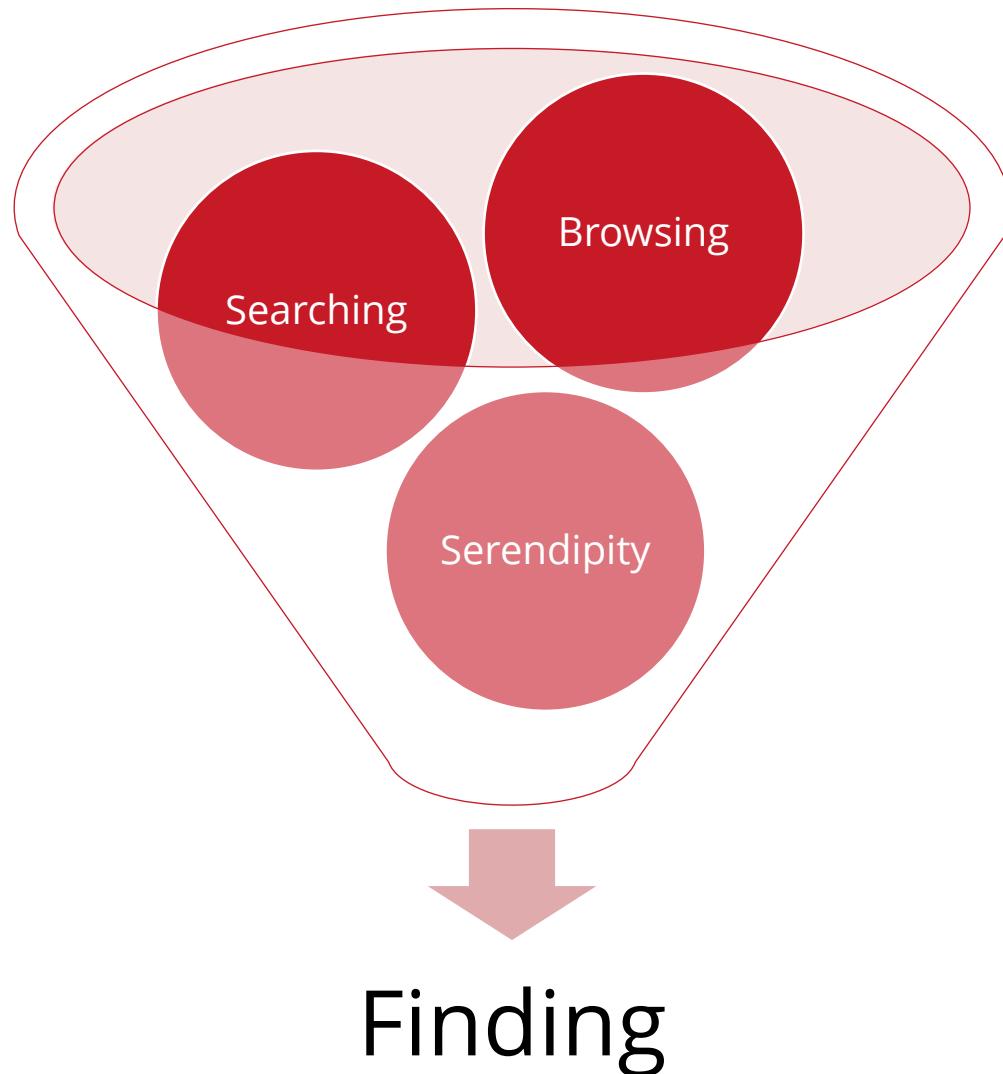
Challenges with “satisfies an information need”

- Ambiguous underspecified queries
- Relevance is subjective, situational, dynamic and multi-dimensional

Definition revisited

“Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections (usually stored on computers)”
Manning, Raghavan & Schütze (2009)

Three ways of finding information

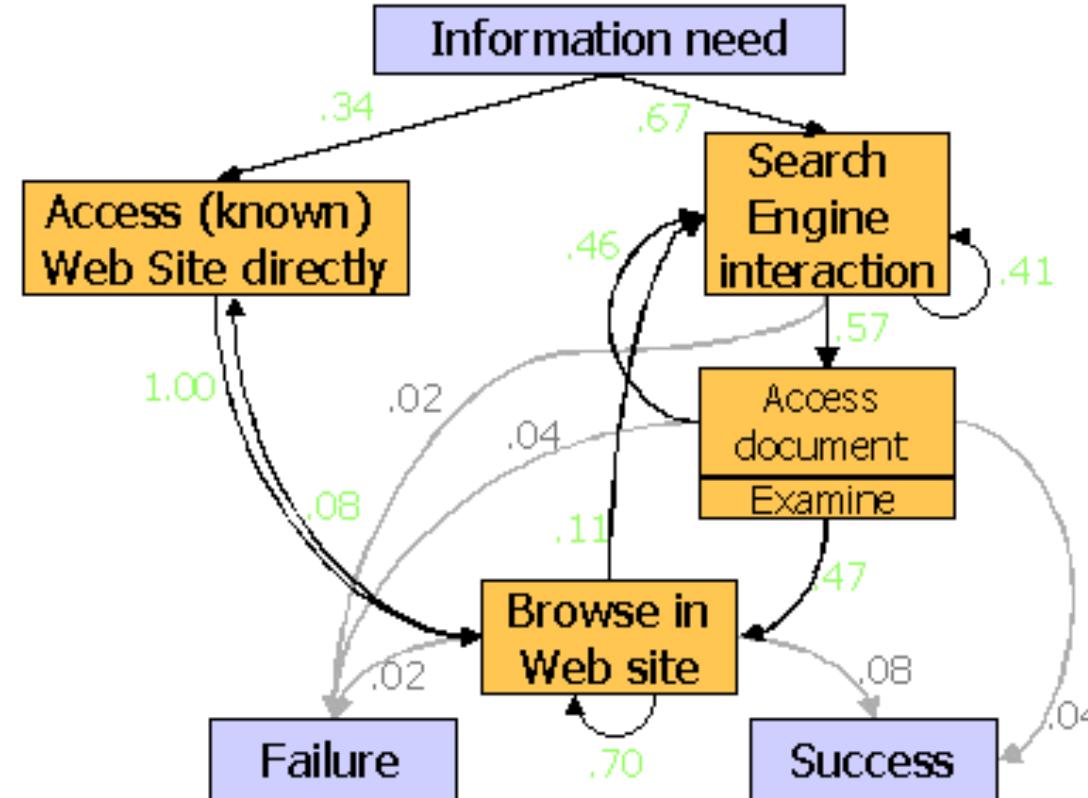


- *Searching* is a structured activity and has long been in use (e.g., querying databases)
 - Different *types* of searches, e.g. known-item vs. subject/topical
 - Searching can be effective ...
 - If the user knows what s/he is looking for (i.e. has specific goal)
 - If the query is specific (e.g. a known-item)
 - ... but searching can also be difficult (e.g. finding the right query terms)
- Typically searching involves *formulating queries*
 - Recall potential words or select suggested categories or terms
 - Can be expressed in natural language (e.g. keywords or concepts) or Boolean logic

- *Browsing* allows users to look for information in a more random and unstructured way than search
 - Suitable when people don't have specific search goals
 - Provides a way of exploring collections
- There are many types of browsing
 - Systematic, exploratory, casual
 - Directed, semi-directed and undirected
- Browsing online content comes in various forms
 - Viewing groups of items by category (e.g. Amazon, Yahoo!)
 - Following hyperlinks
- Many search systems support search and browse behaviours

- *Serendipity* occurs when people find information that they were not actively looking for, but they recognise it might be useful
- Can result from other information seeking activities such as searching and browsing on unrelated tasks
- Can be 'encountered' in the course of more general work and leisure activities
- It is difficult to design for serendipity, but some features such as recommendations may go some way towards it

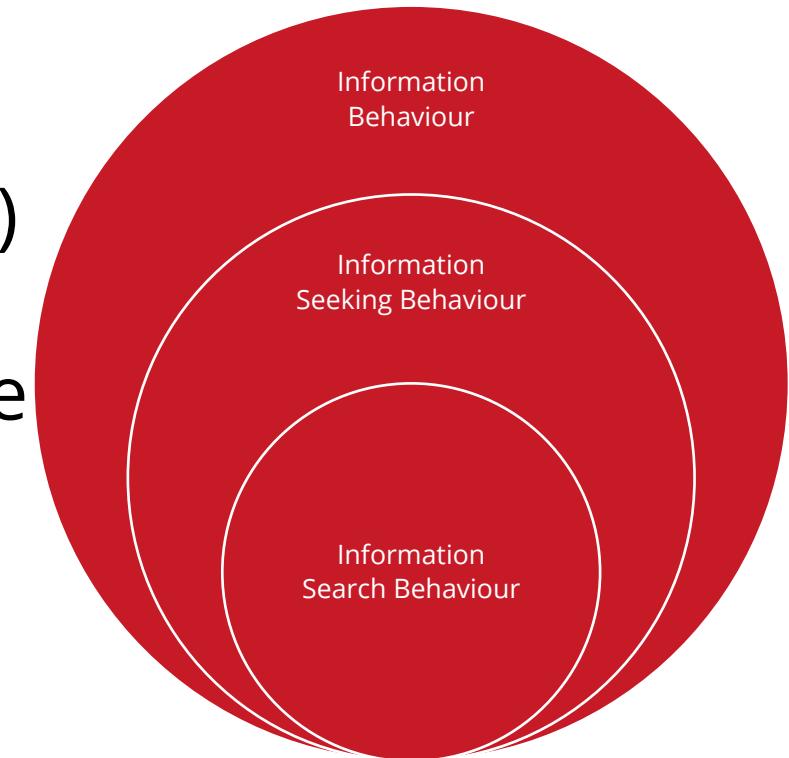
People often use multiple behaviours



Christoph Hölscher and Gerhard Strube. 2000. Web search behavior of Internet experts and newbies. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking*. North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands, 337-346. <http://www.www9.org/w9cdrom/81/81.html>

Information seeking

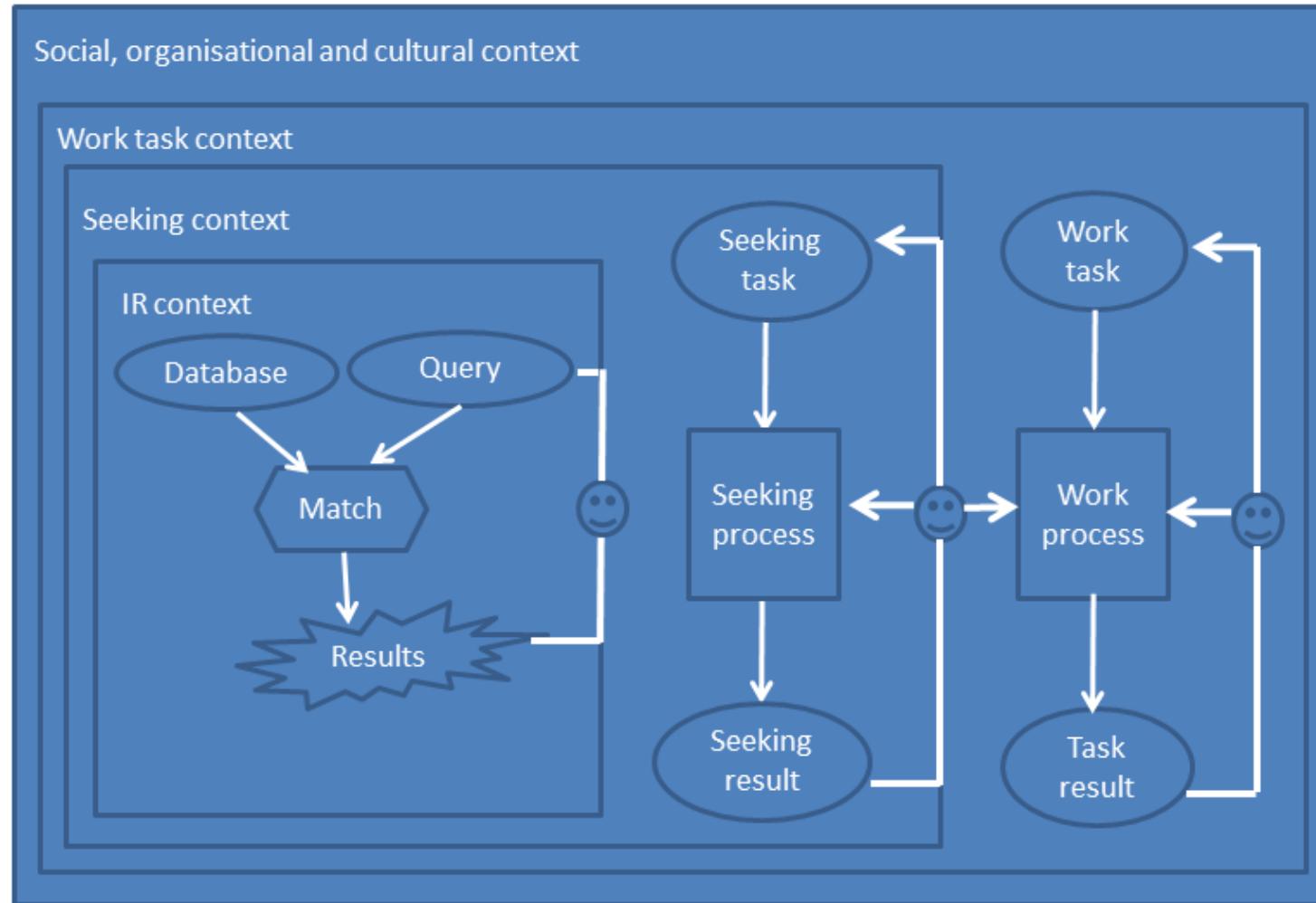
- Information seeking is often defined as human information behaviour that deals with searching or seeking information, by means of information sources and (possibly) interactive information retrieval systems
- Information seeking and search activities are the result of some *information need*
 - Consciously identified gap in knowledge
 - Commonly driven by context, e.g. work task
 - Types include: exact information, a few good things, everything, don't know yet and will revisit



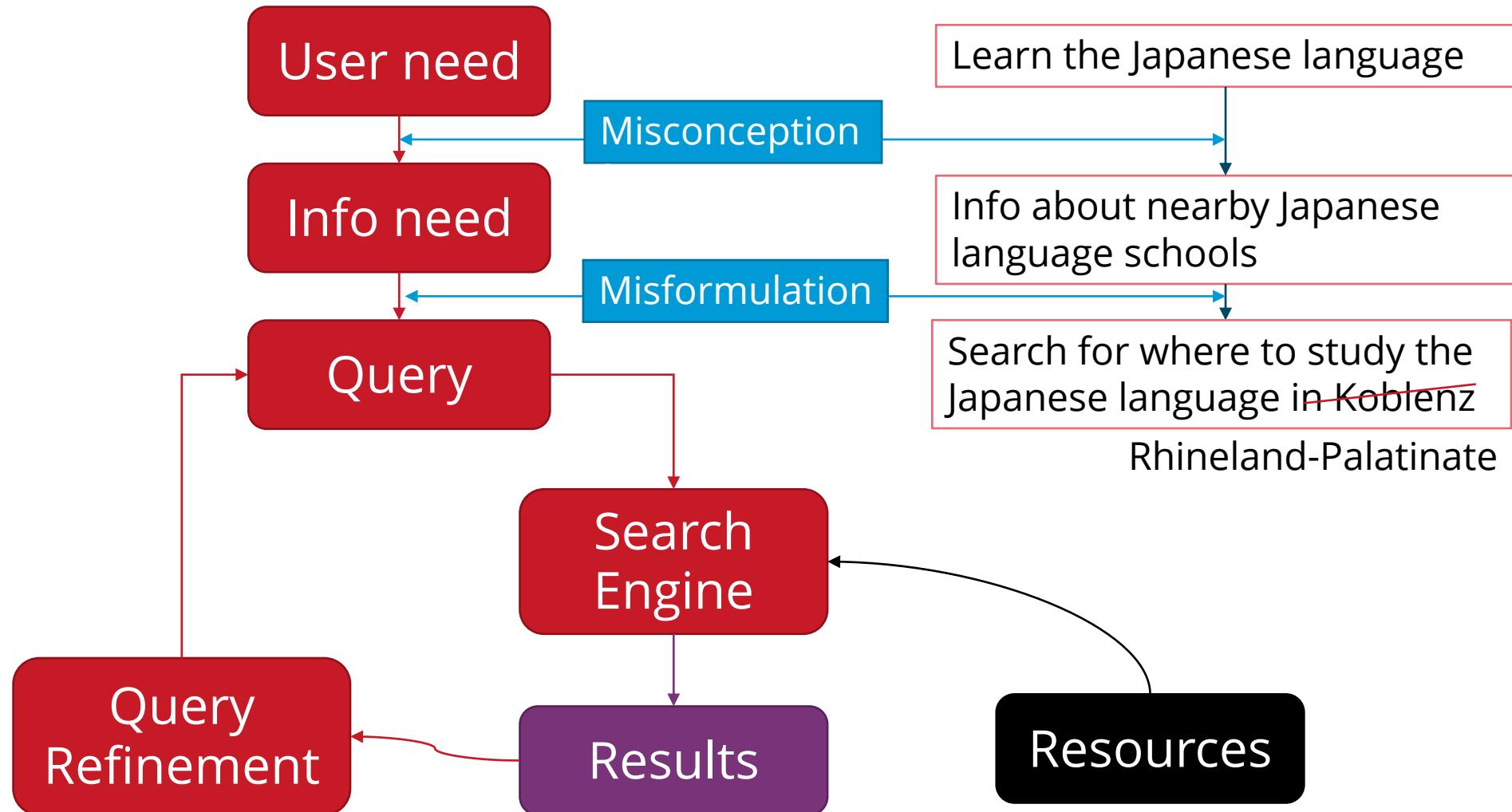


› Information retrieval in context

IR takes place in context(s)



Järvelin, K. & Ingwersen, P. (2004). "Information seeking research needs extension towards tasks and technology" *Information Research*, 10(1) paper 212 [Available at <http://InformationR.net/ir/10-1/paper212.html>]



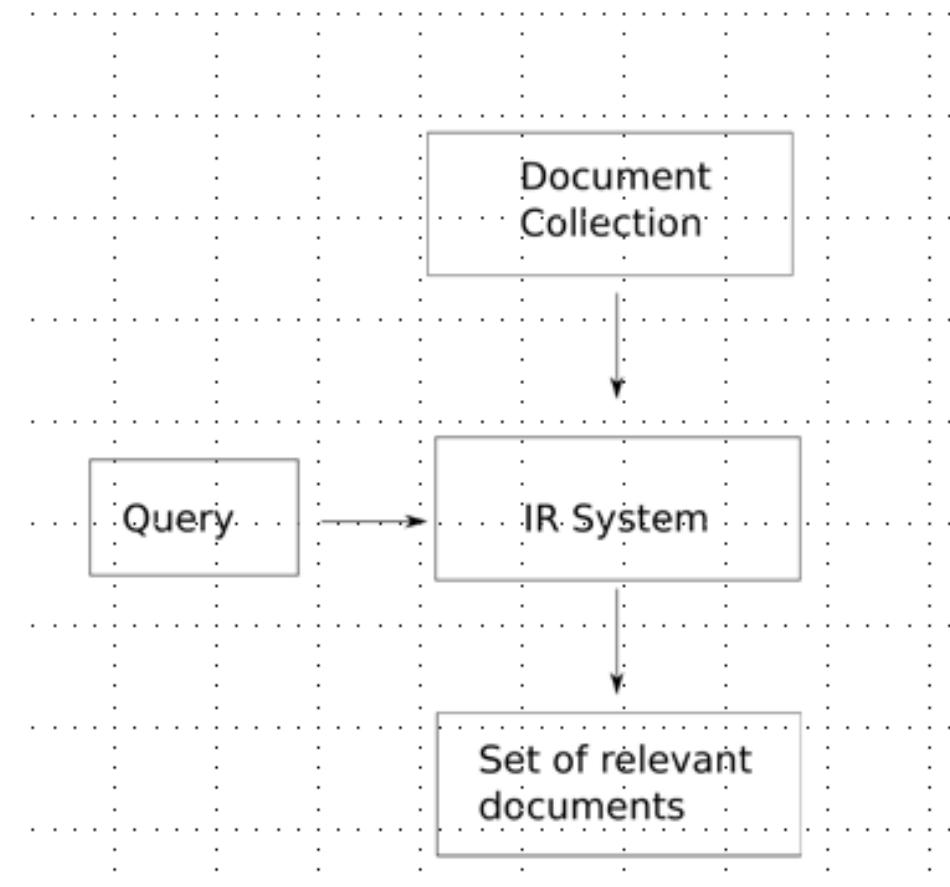
- Question 5: In the past week, on how many different devices have you used an IR system?
- Question 6: In the past week, on what kinds of devices have you used an IR system?
- Question 7: In the past week, in which locations have you used an IR system?



› How IR systems work (briefly)

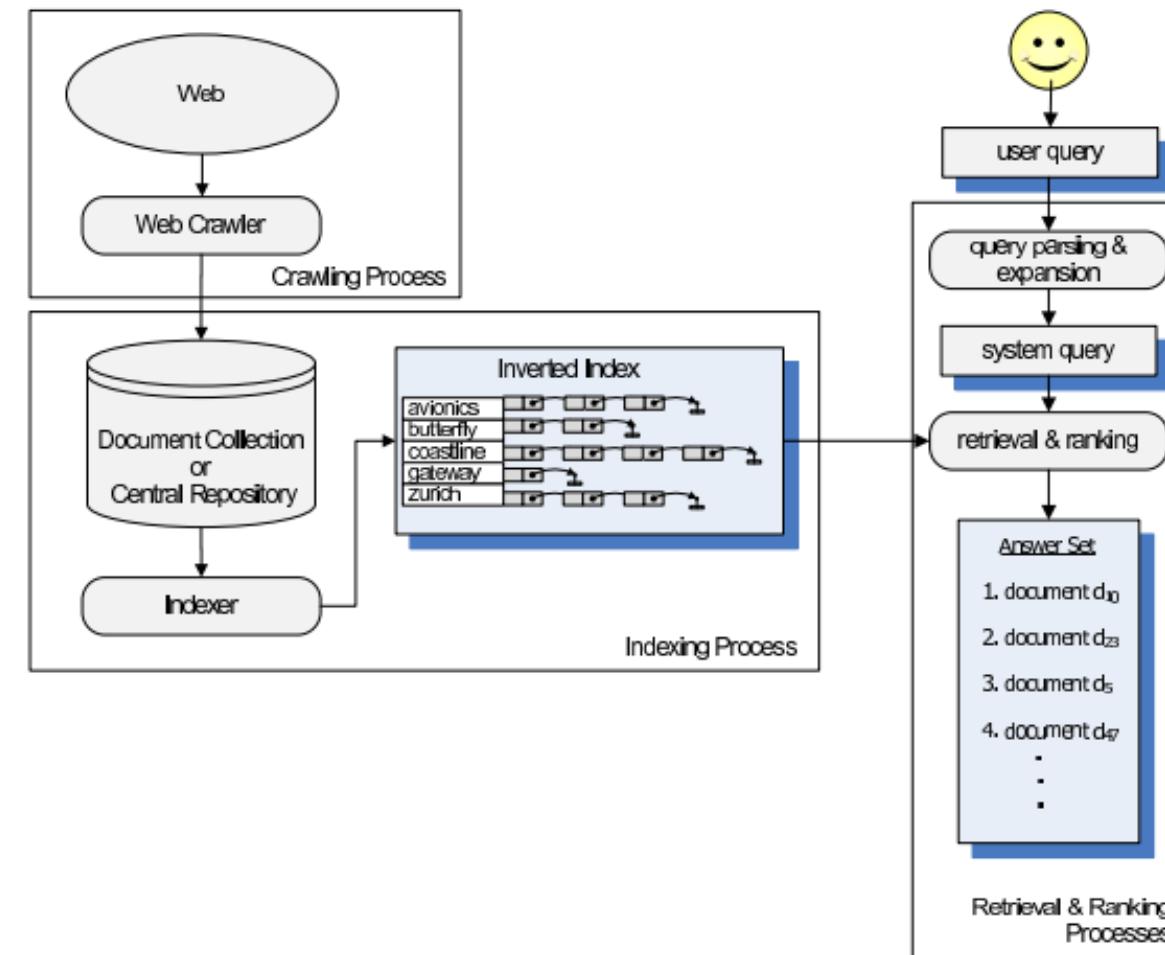
Aim of IR system

- The IR system aims to retrieve the information items which are *relevant* to the user's query
- The output of the IR system is a ranked list of items



Architecture of an IR system

High-level software architecture of an IR system (Baeza-Yates & Ribeiro-Neto, 1999)



- Interface acts as intermediary between users and IR systems
- Well designed interface will help users to
 - Clarify their information needs
 - Formulate suitable queries
 - Understand the results
 - Carry out a range of search tasks effectively



➤ History of IR

A long, long time ago



206BC – Han
imperial library

2500 BC – Ebla tablets, Syria.
World's oldest library (?)



- Increasing interest in automatic indexing
- Vannevar Bush in Atlantic Monthly
 - “Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.”

1950s onwards

- Term “IR” coined by Calvin Mooers in 1950s
- First IR research group Gerald Salton in 1960s
- Development of ideas still in use today
 - Relevance feedback – 1960s
 - Document clustering – 1960s
 - Evaluation techniques – 1960s
 - Term / document frequency – 1970s + 1980s

- The web – IR takes centre stage
- Wayback Machine (<https://archive.org/web/>)
 - Excite, 1996 and Google, 1998



Excite Reviews: site reviews by the web's [best editorial team](#).

- | | | | |
|---------------------------------|------------------------------------|--------------------------------|----------------------------|
| ◦ Arts | ◦ Health | ◦ Magazines | ◦ Politics |
| ◦ Business | ◦ Hobbies | ◦ Movies | ◦ Regional |
| ◦ Computing | ◦ Investing | ◦ Music | ◦ Science |
| ◦ Education | ◦ Law | ◦ News | ◦ Shopping |
| ◦ Entertainment | ◦ Life & Style | ◦ People Pages | ◦ Sports |

Excite City.Net

Plan your weekend, your travels.

Find-A-Destination

[Take me there!](#)

[Maps](#) ◦ [Top Cities](#) ◦ [Concierge](#)

Excite Live!

Your news, your way.

◦ [Latest news](#) ◦ [Stock quotes](#)
 ◦ [Sports scores](#) ◦ [TV listings](#)

Excite Reference

Just the facts, ma'am.

◦ [Yellow Pages](#) ◦ [Maps](#)
 ◦ [People Finder](#) ◦ [Shareware](#)

- Link analysis for Web Search
 - Google
- Automated Information Extraction
 - Fetch
 - Burning Glass
- Question Answering
 - TREC Q/A track

2000s onwards

- Multimedia IR
 - Image
 - Video
 - Audio and music
- Cross-Language IR
 - DARPA Tides
- Document Summarization
- Learning to Rank

■ Generative IR

The screenshot shows the Microsoft Bing search interface with the 'COPilot' tab selected. The main query is "What are the main sites of Koblenz?". The Copilot section provides a list of nine sites:

1. **Deutsches Eck (German Corner)**: This iconic headland extends into the waters where the Rhine and Mosel rivers meet. It has been a vital part of Koblenz's natural defenses since 1216. Standing here, you'll feel like you're on the bow of a ship, with the rivers flowing around you.¹ ².
2. **Fortress of Ehrenbreitstein**: Dominating the right bank of the Rhine, this historic fortress offers commanding views of both rivers. It's a must-visit for history enthusiasts and provides a glimpse into the city's strategic past.¹ ².
3. **Koblenz Cable Car**: Hover over the Rhine in the cable car and discover the Ehrenbreitstein Fortress. The ride offers spectacular vistas of the city and its surroundings.¹.
4. **Forum Confluentes**: An artsy hub where you can explore exhibitions, concerts, and cultural events. It's a great place to immerse yourself in creativity.¹.
5. **Historic Old Castle (Alte Burg)**: Wander through this well-preserved castle, which has witnessed centuries of history. Its architecture and ambiance are captivating.¹.
6. **Stolzenfels Castle**: Take a short drive to this fairytale-like castle overlooking the Rhine. Its romantic setting and lush gardens make it a delightful visit.¹.
7. **Basilica of St. Castor**: Explore this ancient church, one of the oldest in Germany. Its Romanesque architecture and serene atmosphere are worth experiencing.¹.
8. **Kaiserin-Augusta-Anlagen (Rhine Gardens)**: Stroll through these beautiful gardens along the Rhine. They offer a peaceful escape and lovely views of the river.¹.
9. **Old Town Koblenz**: Wander the charming streets of the old town, where historic buildings, cozy cafes, and lively squares.¹

Below the list, there are links to "Learn more" with icons for planetware.com and tripadvisor.com, and a "Stop Responding" button at the bottom.

- System-centred vs. user-centred approaches
 - System-centred approach – developing algorithms
 - User-centred approach – better understanding and representation of *users* and their *interactions* with IR systems
- “IR is ultimately a human activity. Humans and machines can bring complementary strengths to the interactive search process; properly coupling these capabilities can result in a synergy that exceeds the capabilities of either human or machine alone” (Oard et al., 2008)

Related areas

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

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

Library and Information Science

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

Artificial Intelligence

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

Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords
- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*)
- Methods for identifying specific pieces of information in a document (*information extraction*)
- Methods for answering specific NL questions from document corpora or structured data

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

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

› Summary



Summary

- At the end of this lecture, you are expected to
 - have obtained a wide overview of WIR
 - know the elements of Web Information Retrieval
 - understand the difference between structured and unstructured data
 - understand the layout of IR system