TÉCNICO LISBOA

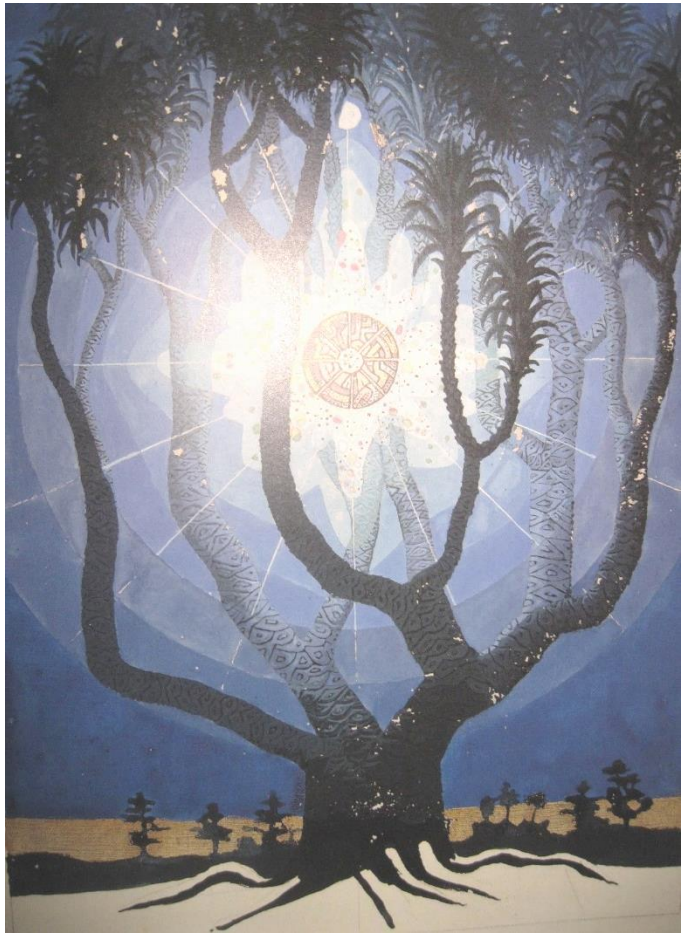
Introduction to Information Retrieval

Essentials on IR and search interfaces

PRI 2023/24

Rui Henriques
rmch@tecnico.ulisboa.pt

Outline



IR definition and origins

Data: structure levels

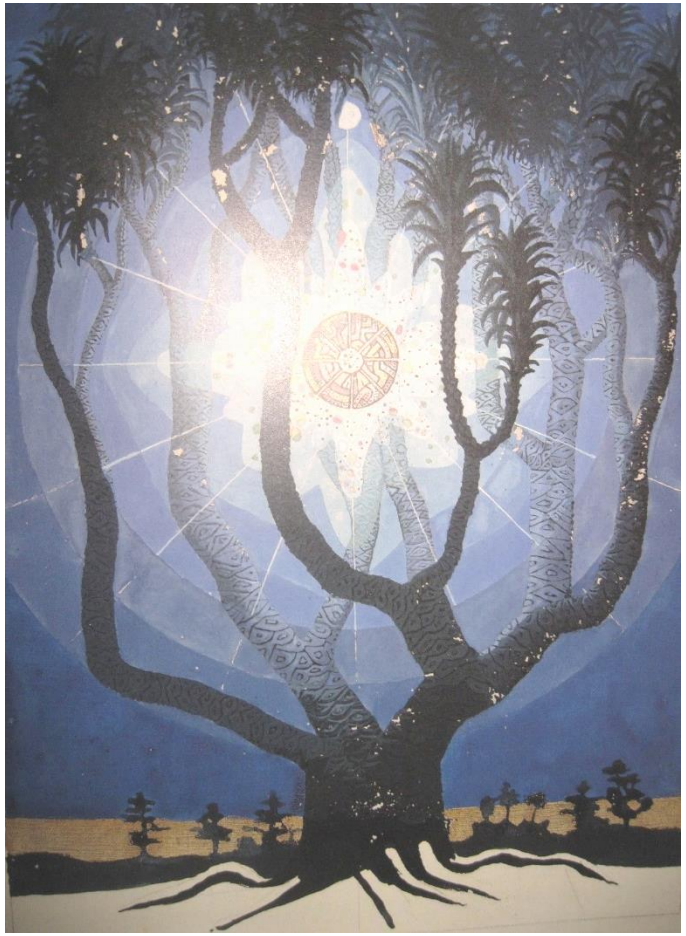
IR systems

Dynamic search model

Search interfaces

- query support
- results display
- evaluation

Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

– query support

– results display

– evaluation

Information retrieval

What is *Information Retrieval* (IR)?

- **processing** of **libraries** to answer an **information need**

The processing of libraries commonly entails:

- **finding** and **organizing** material (usually **documents**) of an **unstructured** nature (usually text)
- from a **large collection** (usually in digital format)

– **document** = book, webpage, media file, article, form
– **library** = **document collection**
  – also known as **corpus** if the collection refers to a well-defined topic
  – *static*: well-defined set of documents
  – *dynamic*: documents can be added and removed along time

Information retrieval: application domains

When we think of IR we think first of ***web search***, ***fact retrieval*** and ***question answering***… yet:

- e-mail and laptop search
- corporate knowledge bases
- academic bases (peer-reviewed articles)
- legal information retrieval

… retrieval at different levels:

- **global** (web)
- **organizational** (institutions, companies, academy)
- **individual** (personal computers, e-mail, social media)

Information retrieval: tasks

Early core task of IR:

– **searching** (also known as querying)

  Given a collection, retrieve relevant documents to the user's information need (helping the user complete a task)

Nowadays, research in IR:

– **crawling** and **indexing** massive collections
– **fact retrieval**
– **question answering** using **large language models**
– **browsing** (also known as navigation)
– document **annotation** (e.g., **fake text detection** and **sentiment analysis**)
– corpus **organization**
– **cross-language** retrieval
– document **summarization** and restyling…

The beginning…

- For over 5000 years, man has organized information for later retrieval and searching
  - compiling, storing, organizing, and indexing **hieroglyphics**, **papyrus** and **books**

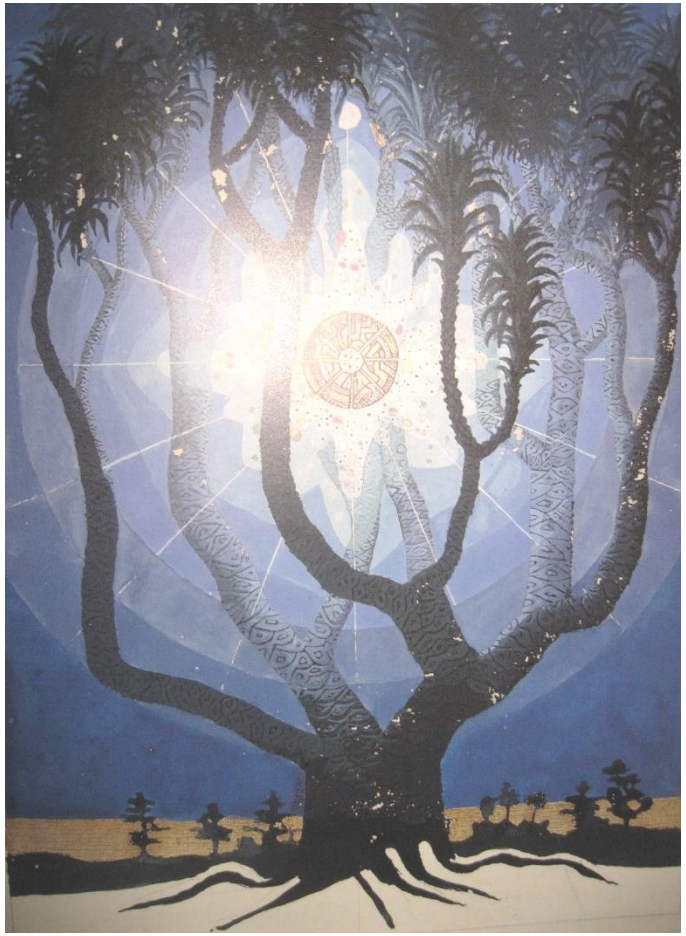
- To hold the various items:
  - ***physical libraries*** were built
    - oldest known library in Elba in the Fertile Crescent (between 3000 and 2500 BC)
    - Great Library at Alexandria (300 BC)
    - since then: libraries everywhere
  - with the advent of informatics: documents are digitalized and maintained within **digital libraries**
    - from librarians to information and companies everywhere
  - nowadays: a global digital library – the **Web**

From manual to automatic IR

- Volume of information in physical, digital and global libraries growing…
  - necessary to build specialized data structures for efficient search —*indexes*
    - for centuries: *manual indexing* (e.g., card catalogs' searching)
    - since 50s: *automatic* ***indexing*** and Boolean querying
    - today: enriched with advanced query support, graphical search interfaces, hypertext features

- Brief **timeline**
  - IR concepts introduced in 50's
  - in late 60's: the *TF-IDF term weighting scheme,* largely used up to date
  - in 70's: first ACM SIGIR International Conference on Information Retrieval held in Rochester

Outline



IR definition and origins

Data: structure levels

IR system

Dynamic search model

Search interfaces

– query support

– results display

– evaluation

Structured vs unstructured data

- **Structured data**
  - … from measuring systems along time
    - geophysical, digital, mechanical, physiological, biological, societal, organizational, hybrid systems
  - typically organized in series (sensorized systems), relational and multi-dimensional formats

- **Unstructured data**
  - … from ad-hoc and knowledge systems
    - communication acts (whether spontaneous, opinions, published material) from individuals and organizations
  - typically in reference to free text (written or transcribed)

- What about **media data**?
  - audio, image and video

Structured data

- searching structured data: ***data engineering*** (databases)
  - well defined structure and semantics
  - a single erroneous object among a thousand of retrieved objects means failure

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

… seeks exact match queries and allows diverse numerical operators

Salary < 60000 AND Manager = Smith

- learning models from structured data: ***machine learning***

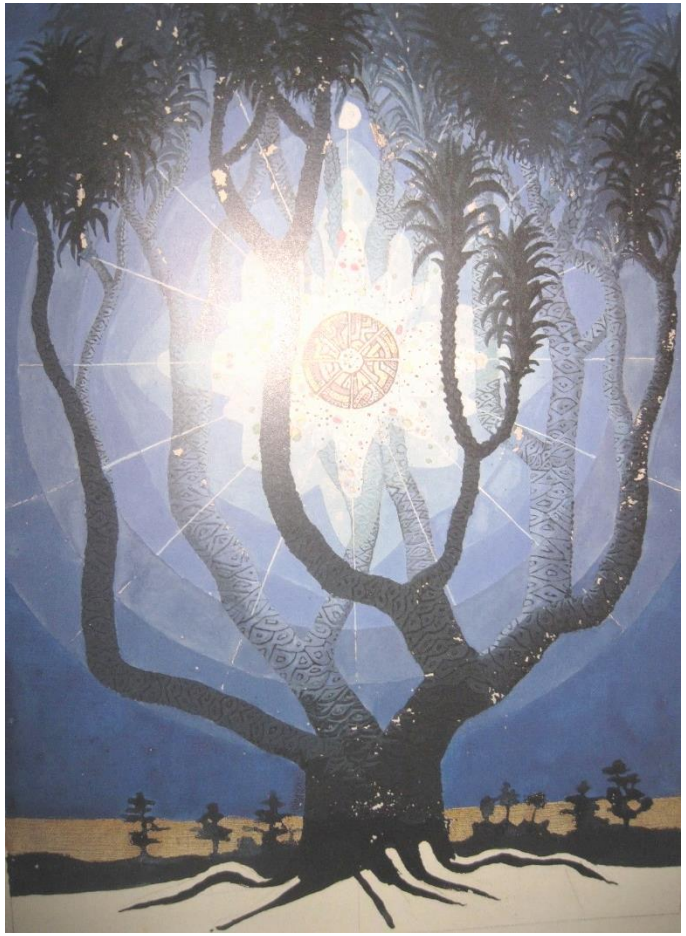
Semi-structured data

- In fact almost no data is *unstructured*
  - *this slide* has distinctly identified zones such as *Title* and *Bullets*
  - as well as **linguistic structure**!

- Semi-structured data
  - research articles: zones?
  - web pages: zones? enriched text?
  - what about XML data?

- Examples on searching semi-structured data:
  - *title* contains data AND *bullets* contain search
  - *title* is about object-oriented programming AND *author* is stro*rup
    where * is a wildcard operator

Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

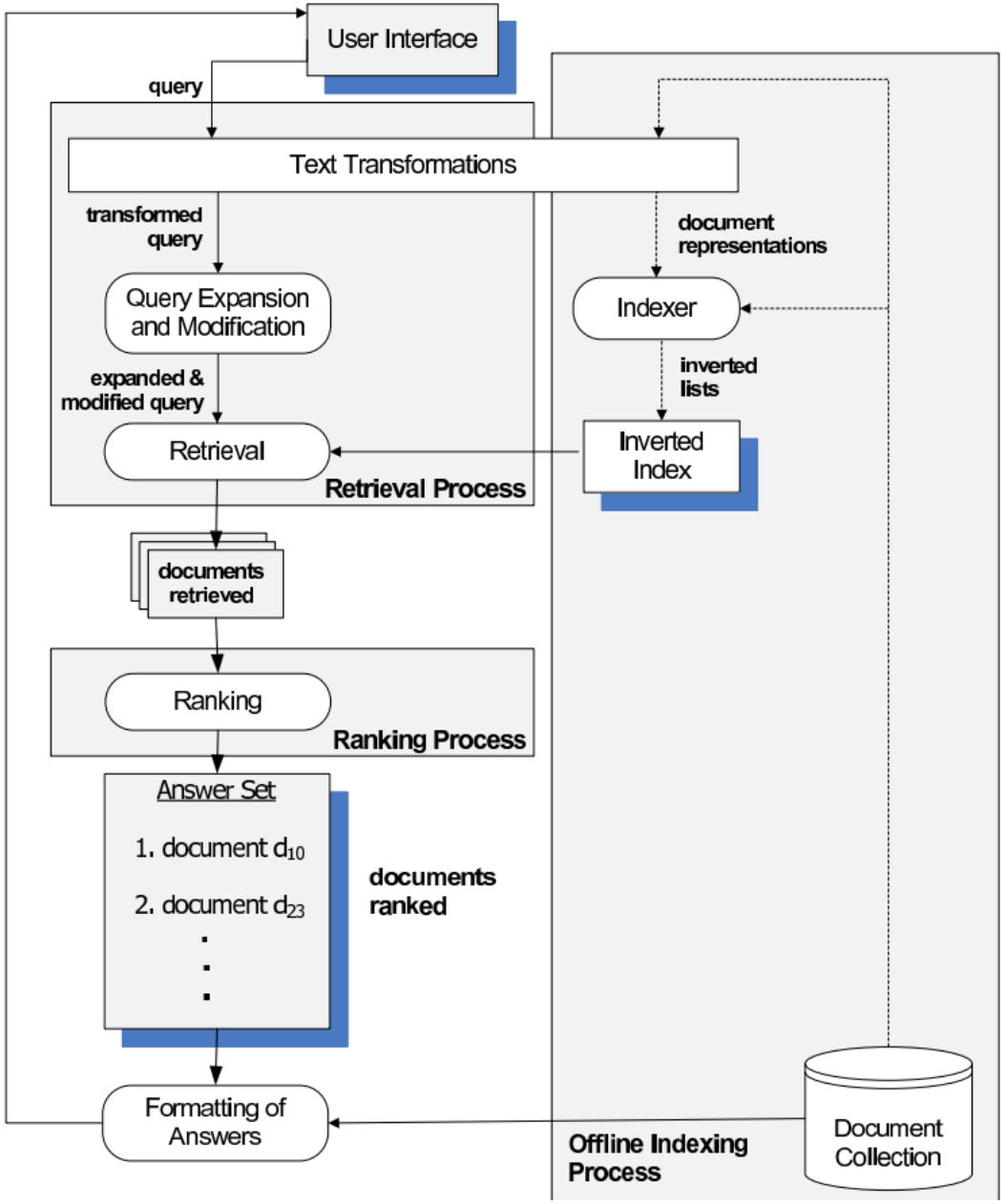
– query support

– results display

– evaluation

IR system

- a **computational system** to support information retrieval
- document **processing, indexing, retrieval, ranking,** and **annotation** as core IR tasks
- let us focus on a specific task: **document retrieval**
  - *architecture* on the left



How good is the retrieved information?

IR systems may correctly answer a **query**, yet not necessarily satisfy the **information need**

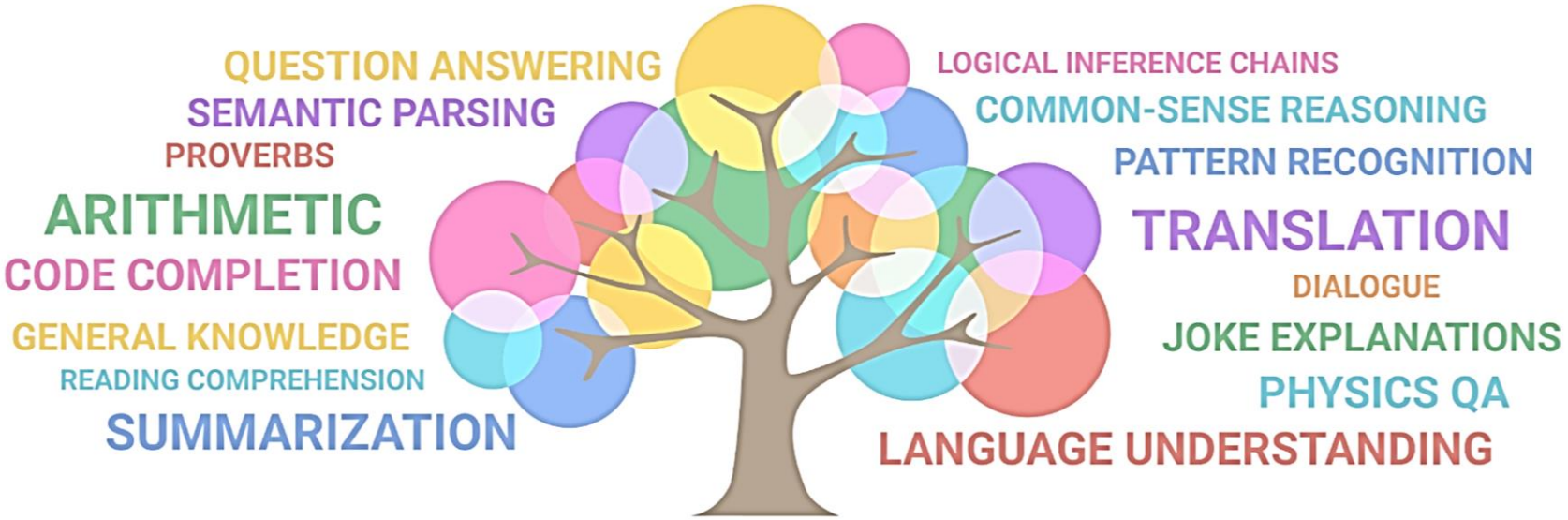
- "*wine death*" (query) to learn about reduced cardiovascular risks (information need)
  may return information on car deaths under alcohol effect (ok for the query yet not the need)

Considering **searching** (document retrieval)…

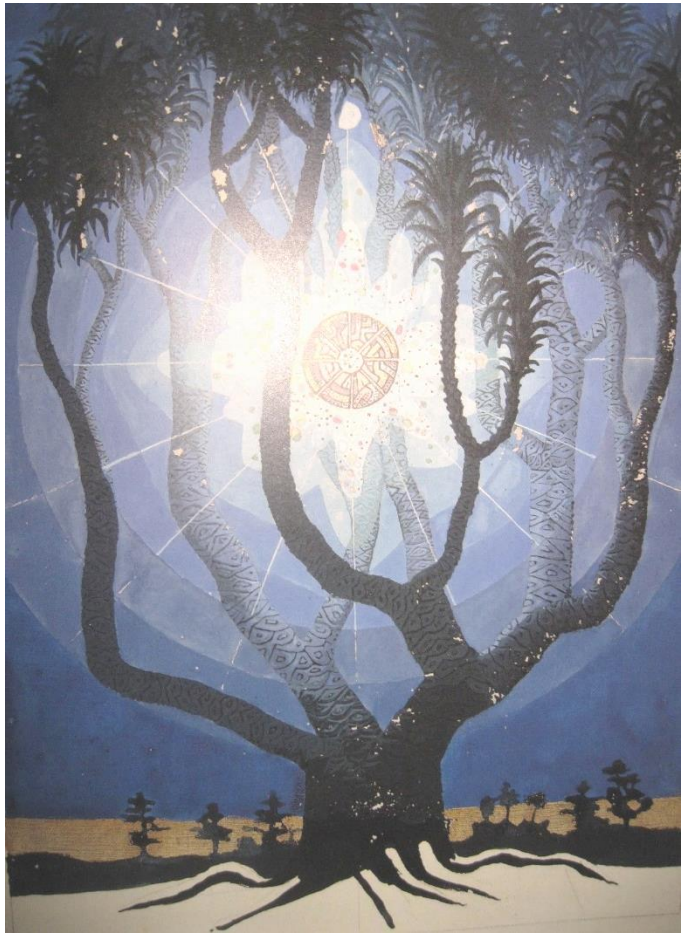
- **goal**: an IR system able to retrieve all relevant items to a user query
  while retrieving as few non-relevant items as possible
  - ***precision***: fraction of retrieved docs that are relevant to the user's information need
  - ***recall***: fraction of relevant docs in collection that are retrieved
- simple retrieval *versus* **ranking**
  — ranking tasks order information items according to a degree of relevance to the user query
  — the notion of relevance is of central importance in IR!
  — evaluation of ranking outputs require additional metrics!

Going beyond document retrieval…

- How do we extend our IR systems to handle additional tasks?
  How do we evaluate the success of IR systems for those tasks?
  - annotating documents (e.g., tags, spam e-mail, fake web content, hateful text)
  - manipulating documents (e.g., translation, restyling, summarization)
  - answering natural language questions



Outline



IR definition and origins

Data: structure levels

IR systems

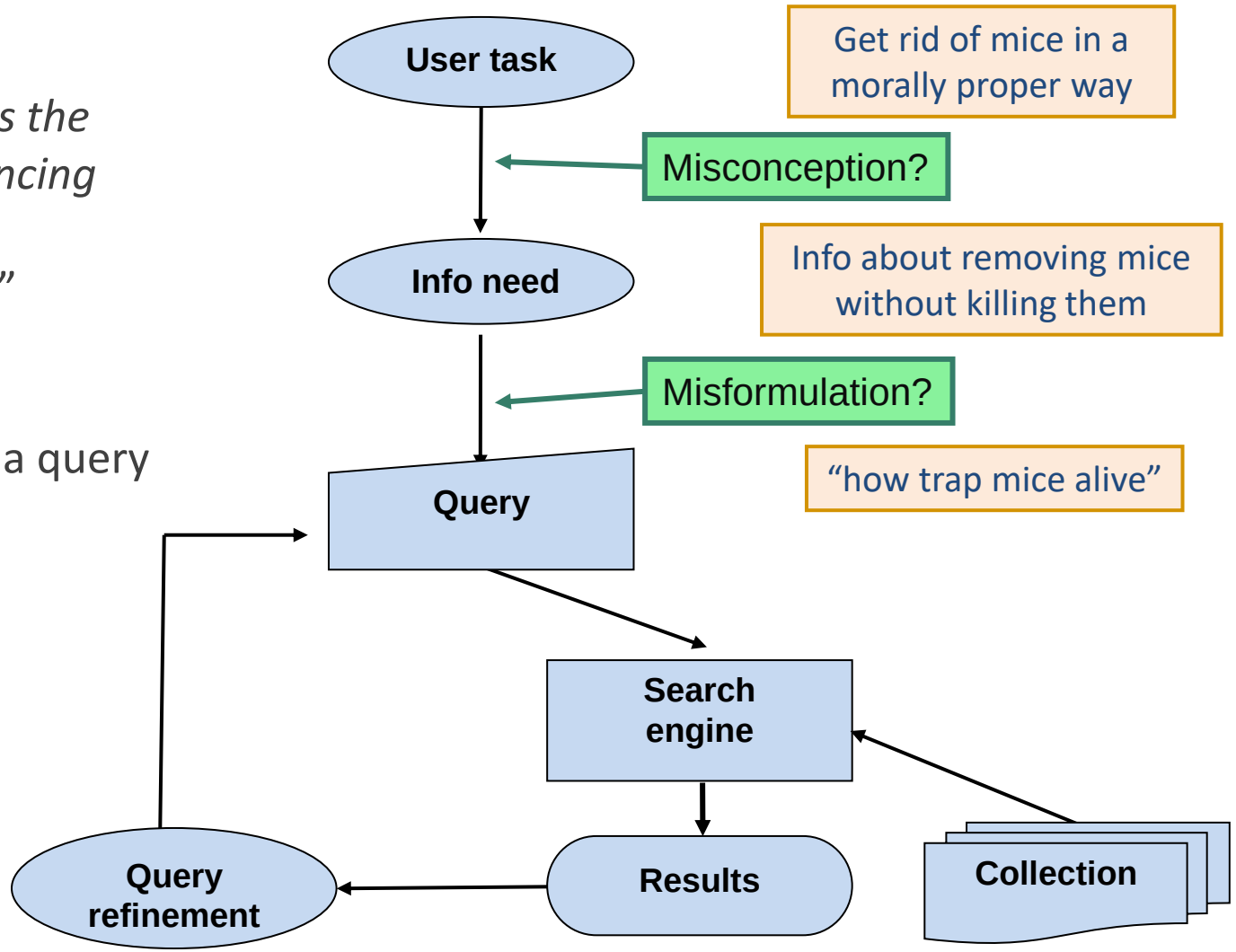
Dynamic search model

Search interfaces

- query support
- results display
- evaluation

Classic search model

- **need**: *"Find all documents that address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)"*
- **querying** process
  - Translate the information need into a query
  - executing query on the IR system



Classic *versus* dynamic search model

- **Classic** information seeking process:
  1. problem identification
  2. articulation of information need(s)
  3. query formulation
  4. results assessment
- More recent models emphasize the **dynamic nature** of the search process:
  - users learn as they search
  - information needs adjust as they see retrieval results and document surrogates (**orienteering**)
- This dynamic process is sometimes referred to as the **berry picking** model of search
  - motivated by the rapid response times of today's web search engines
  - [Jansen *et al.*] search logs reveal that 52% of users commonly modify their queries

Types of searches

Marchionini makes a distinction between:

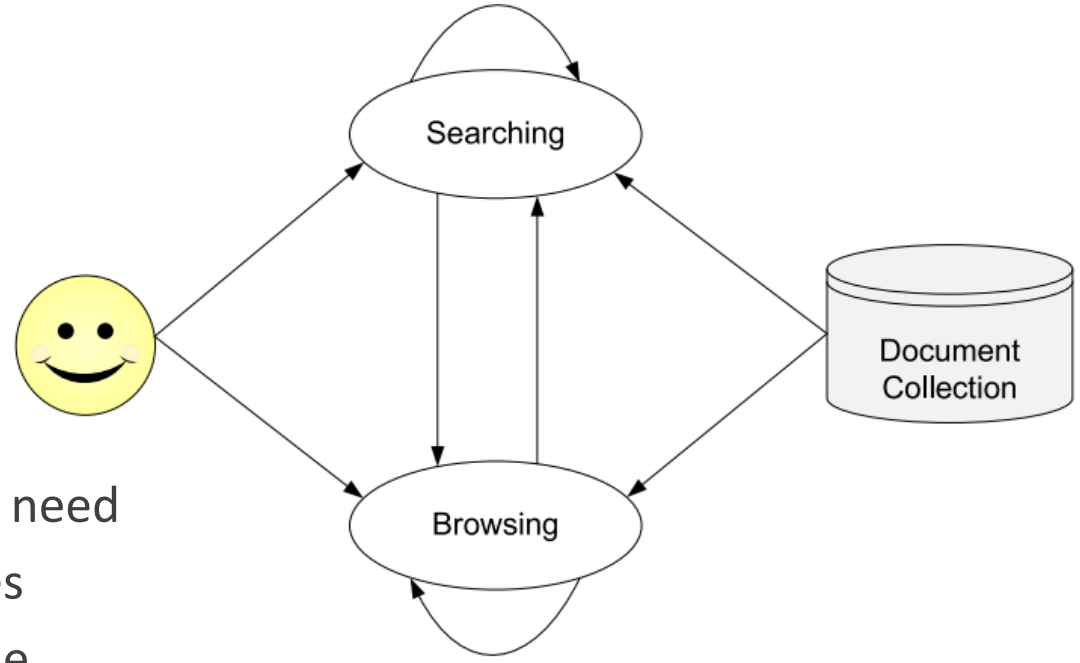
- **information lookup** tasks
  - can be satisfied by discrete pieces of information: numbers, dates, names or websites
  - *searching, fact retrieval* or *question answering*
- **exploratory** tasks
  - **learning**
    - requires more than a single query-response interaction
    - requires time scanning multiple items and synthesizing content to form new understanding
  - **investigation**
    - longer-term process requiring iterations that take place over perhaps long periods of time
    - example: finding a large portion of relevant information available

In the course, we further consider **document annotation** and **manipulation** tasks as part of IR

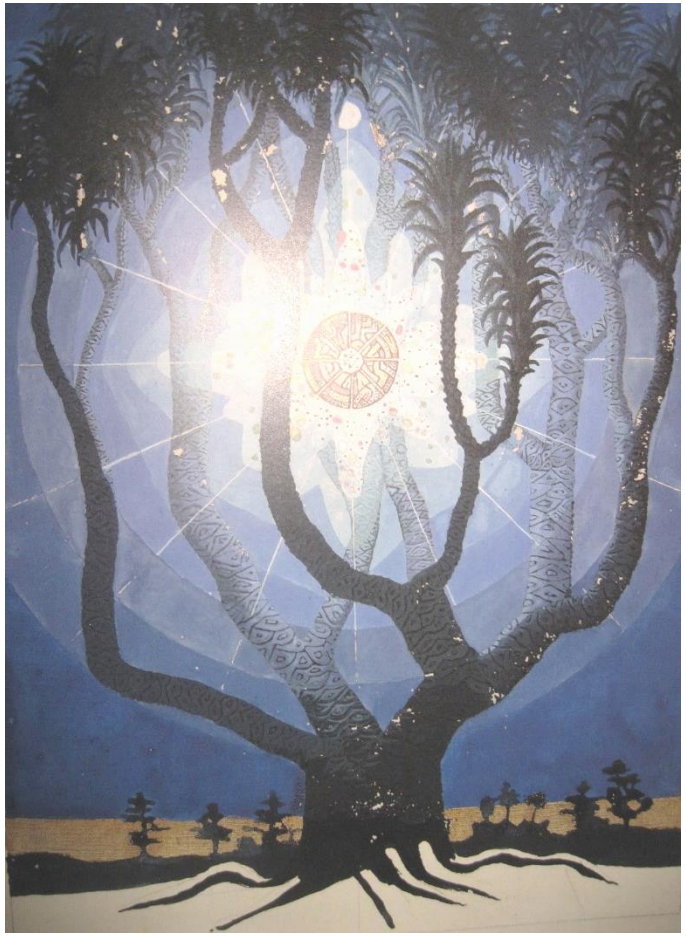
Hybrid searching and browsing

Lookup and exploratory searches can entail **browsing**

- navigating through documents of a given collection
- browsing strategy engaged when:
  - seeking information on a topic that is either poorly defined or inherently broad ⇒ *searching* and *browsing*
    - seeking information on a well-defined topic of interest ⇒ *searching*
  - the information structure well-matches the information need
  - for mentally less taxing recognition of information pieces
  - there are appropriate links among documents, otherwise browsing experience is frustrating



Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

– query support

– results display

– evaluation

Aiding information retrieval

- **Aiding IR systems**
  - supportive **querying interfaces**
  - organization and **visualization** of IR outputs
  - guidance from **relevance feedback**: supervised IR
  - proactive **corpus exploration** (relations among documents, tagging): unsupervised IR
  - advanced **text processing**: natural language processing (e.g., query expansion, correction)

- IR systems are described by three major components
  - **search interface module** (facilities to input queries and output results) => **today**!
  - **indexing module** (facilities to crawl and store document collections) => **next** class!
  - **processing module** (the IR brain to answer queries) => our focus for the period!

Insights from user behavior

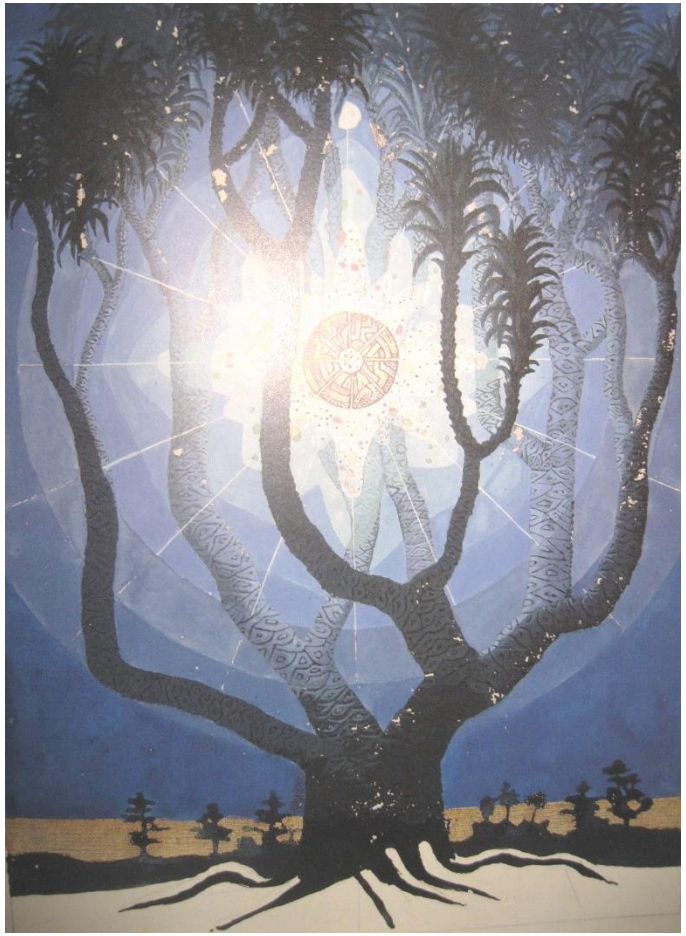
- Searchers typically…
  - **reformulate** their queries with slight modifications
  - look **only** to the **top-ranked** retrieved results or to the **highlights** of a text answer
  - biasedly think the **top documents** (or initial paragraphs) **are better** than those beneath
  - poorly estimate how much of the relevant material was found
  - **not** particularly good at **judging the relevance of information** (especially for unfamiliar topics)
  - search for information previously accessed
    - search strategies differ when searching over previously seen and new materials
    - search interfaces should support **query history** and **revisitation**

Search interfaces

- User interaction with search interfaces differs depending on:
  - the target **task**
  - the **domain expertise** of the info seeker
  - the **amount of time and effort** available to invest in the process

- This interaction – also known as ***information seeking*** – can be seen as being part of a larger process referred to as ***sensemaking***
  - **sensemaking:** iterative process of formulating a conceptual representation
  - examples: legal discovery process, epidemiology (disease tracking), studying customer complaints to improve service

Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

- **query support**
- results display
- evaluation

Search interfaces

- *Classic* methods of information retrieval:
  - entering query into a *search entry form* ⇦
    - generally small queries: *"tests the waters"*
    - if results not relevant: query reformulation (orienteering)
  - selecting links from a directory (website, bookmarks)
- *Contemporary* methods of information retrieval
  - natural language: text or speech form
  - possibility to upload documents and further specify goals (doc retrieval, access facts annotation…)
- **Search entry forms** can be classified according to the supported syntax
    - Boolean operators and auxiliary operators
    - language query in written or audio format (e.g. Alexia)

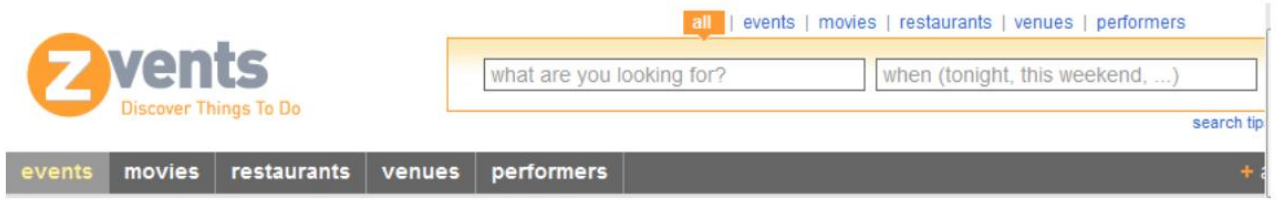
Search interfaces

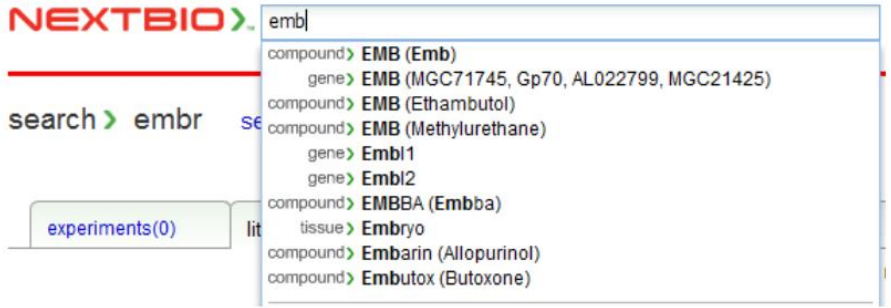
- Can provide multiple forms (for **specialized inputs**)
  - offering hints on the kind of info (e.g., author, minimum #cites, reference document) to enter per form
- Typically support **auto-complete** and **auto-suggest** facilities (dynamic **query suggestions** and **reformulation**)
  - suggestions shown are those whose prefix matches the characters typed so far
  - suggestions may show **synonyms** of typed words
  - **spelling corrections**
  - [Anick *et al.*] users click on dynamic Yahoo suggestions more than one third of times
- **How to** implement query suggestions/reformulations?
  - available metadata and dictionaries (background knowledge)
  - user's own query history
  - querying behavior of similar users
  - querying behavior of all users

Search interfaces

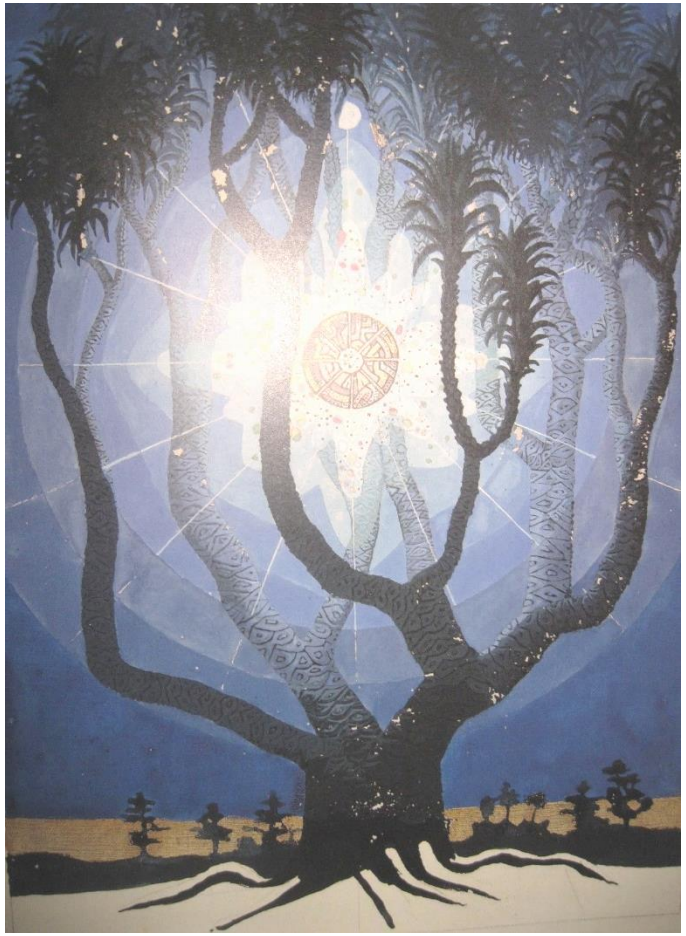
Examples:

- google
- scholar
- yelp
- zevents
- netflix.com
- nextbio





Outline



IR definition and origins

Data structuring

IR system

Dynamic search model

Search interfaces

- query support
- **results display**
- evaluation

Visualizing results

Visualization differs depending on the task

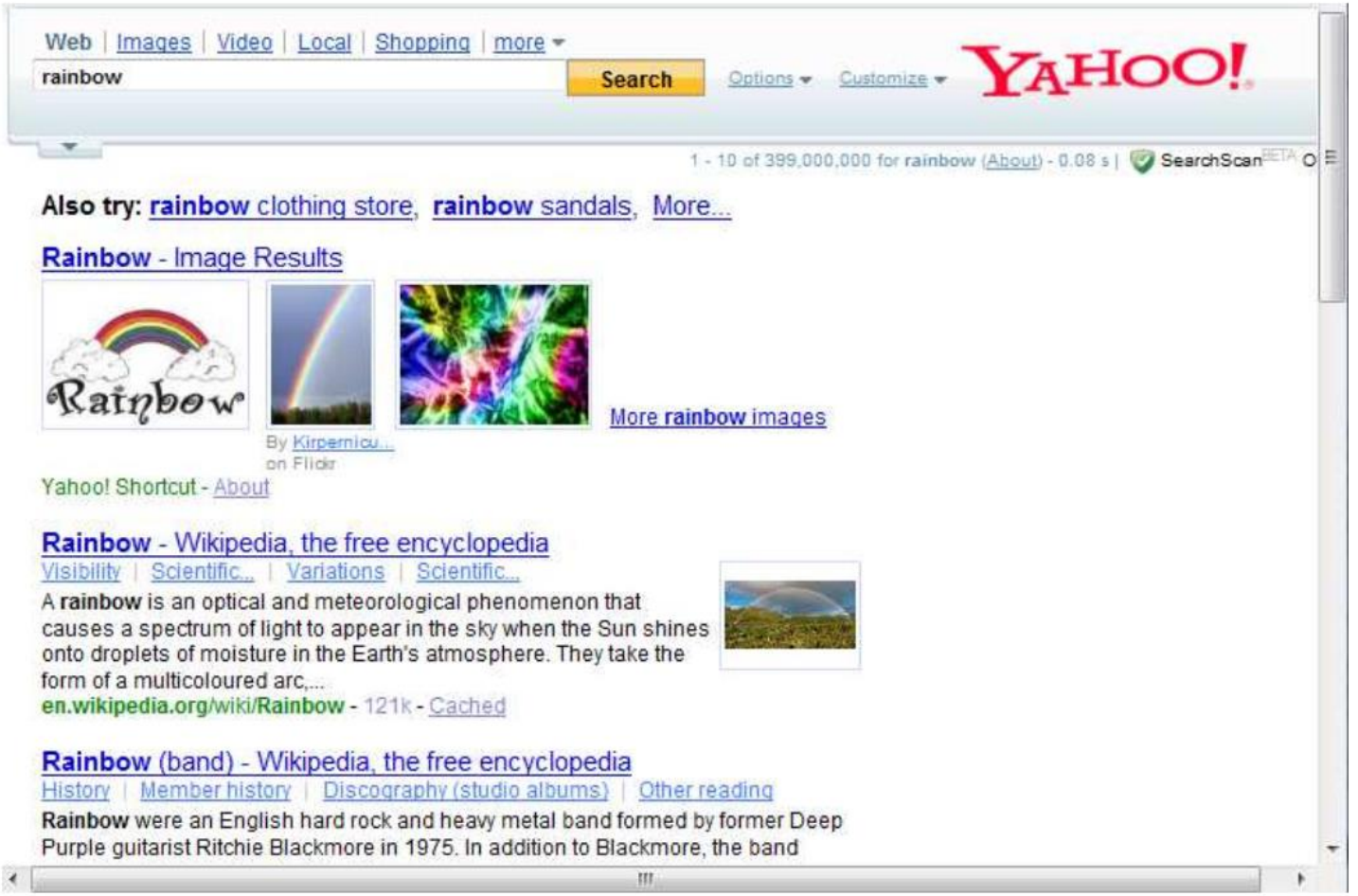
- **fact retrieval**: relevant pieces (single fact *versus* multiple facts with context and/or likelihood)
- **question answering**: simple plain text versus rich formatted text (*css*, *rft*, *markdown*)
- **document retrieval**: let us revisit *google* as a reference IR…

How to display an (ordered) set of retrieved documents?

- *adequate visualization* ***per document***
  - page title, author, URL, date published, summary (snippet)
  - document **surrogate** refers to the information that summarizes the document
    - the quality of the surrogate greatly affects the perceived relevance of the search engine
  - query terms highlighted in the context in which they appear in the document
    - referred as **term highlighting** or keywords in context (**KWIC**)
    - improves user's ability to gauge document relevance
- *organizing the retrieved documents:* flat/list, group, hierarchy, faceted

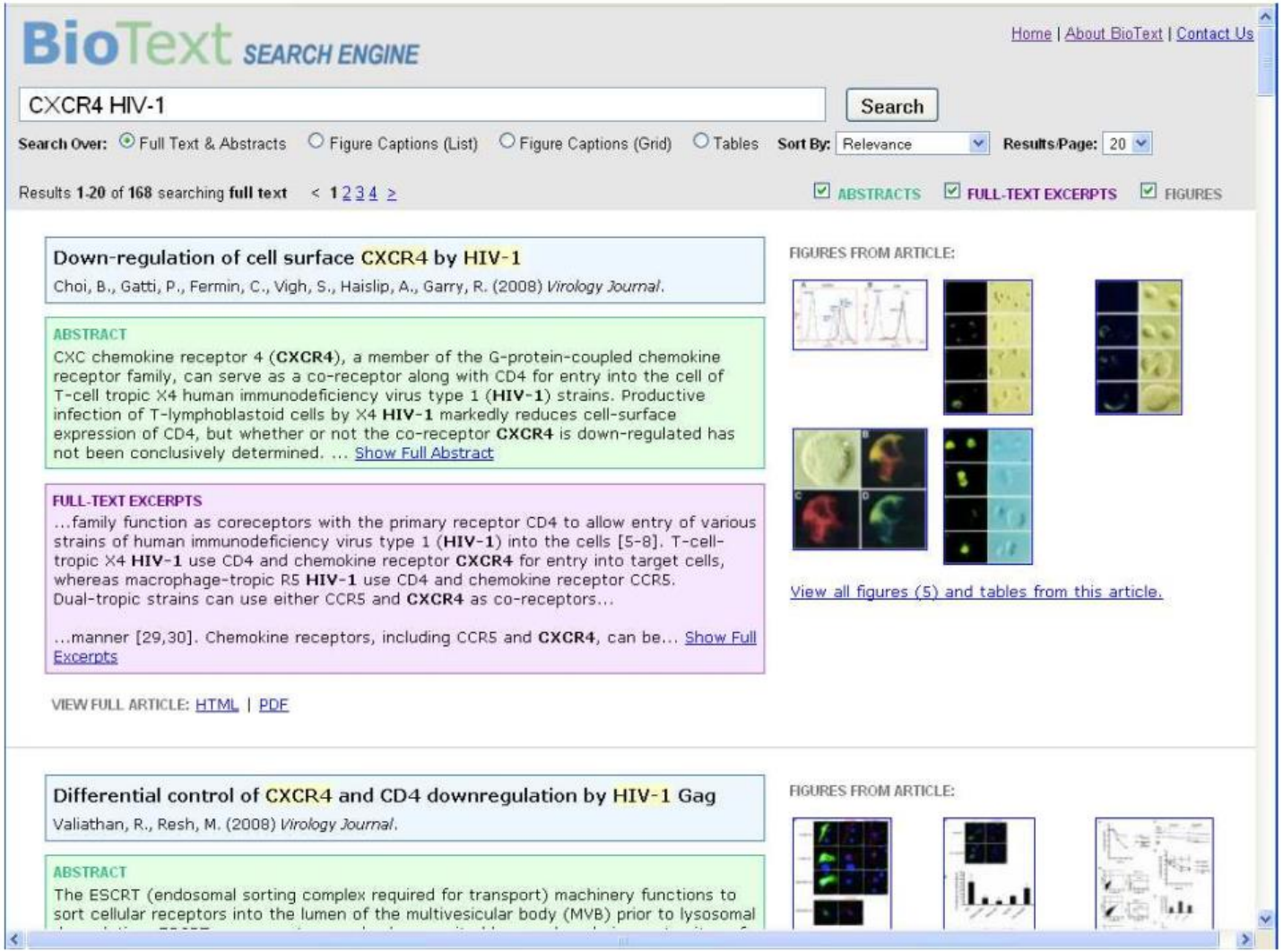
Retrieval results display

- **blended surrogate** can combine:
  - text summaries
  - metadata (annotations)
  - media
- text **summaries**
  - can either follow **extractive** (copy of central sentences) or **generative** approaches (synthesized text)
  - either **query-independent** or **query-dependent** (also termed query-biased, query-oriented, or user-directed summaries)



Retrieval results display

- other *blended surrogates*:
  - figures from journal articles alongside the search results
  - speech assistance
  - audio and video snippets (e.g. pre-loading on hover)
  - going beyond doc retrieval:
    - multimodal answers, e.g. (generated) images and captioning

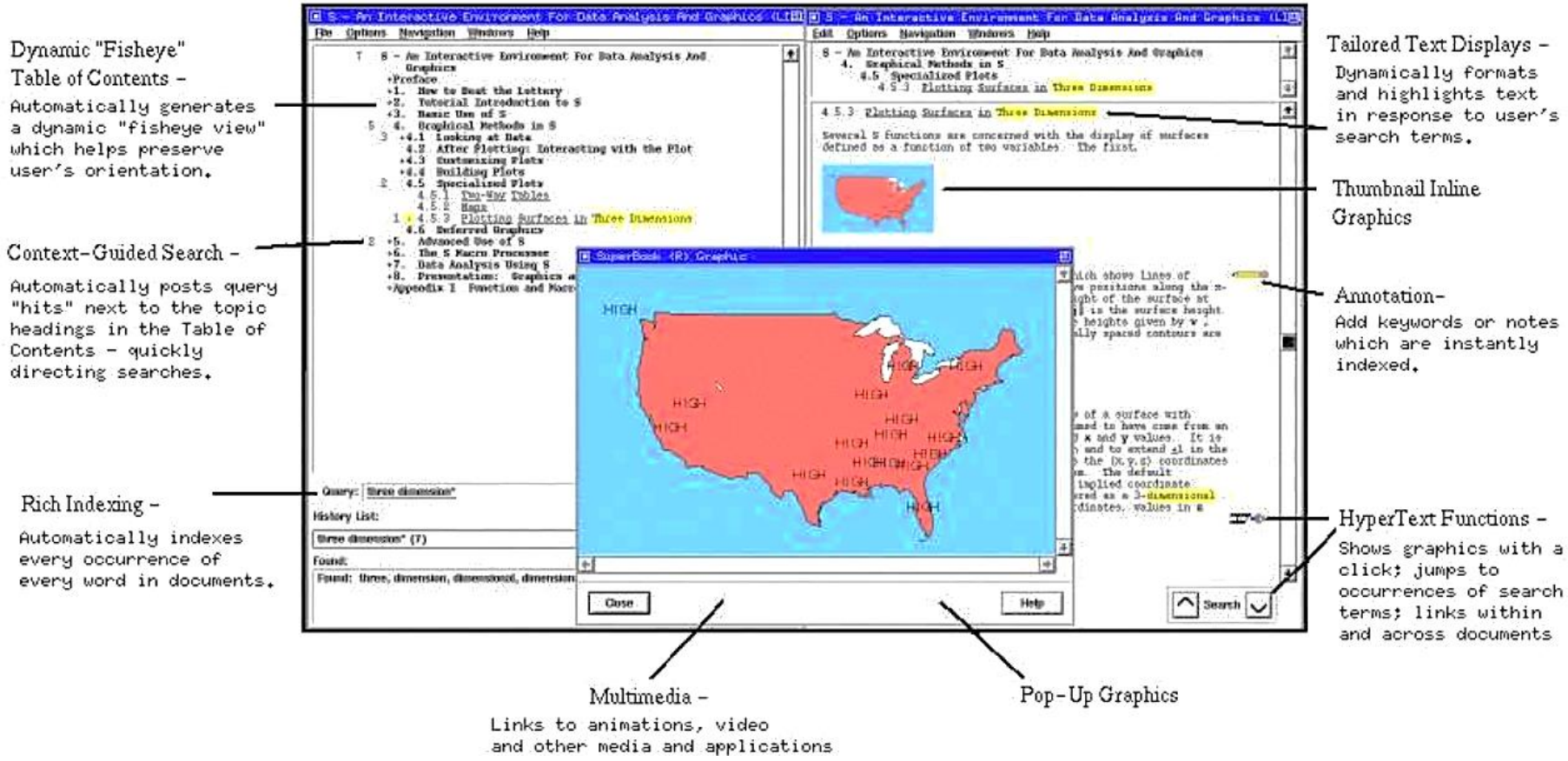


Retrieval results display

Organizing the retrieved documents:

- **category system**: meaningful labels reflecting the concepts relevant to a domain
  - good categories: coherent, complete per document, predictable across different searches
  - **type of categories**
    - **flat:** simple list of topics/subjects
    - **hierarchical:** in early web days, hierarchical systems such as Yahoo's were popular
    - **faceted** (tagging)**:** assignment of multiple categories to a single item
      - can be used to group, filter (narrow) and sort documents
      - each category corresponds to a different facet
- **clustering system**: unsupervised grouping of documents

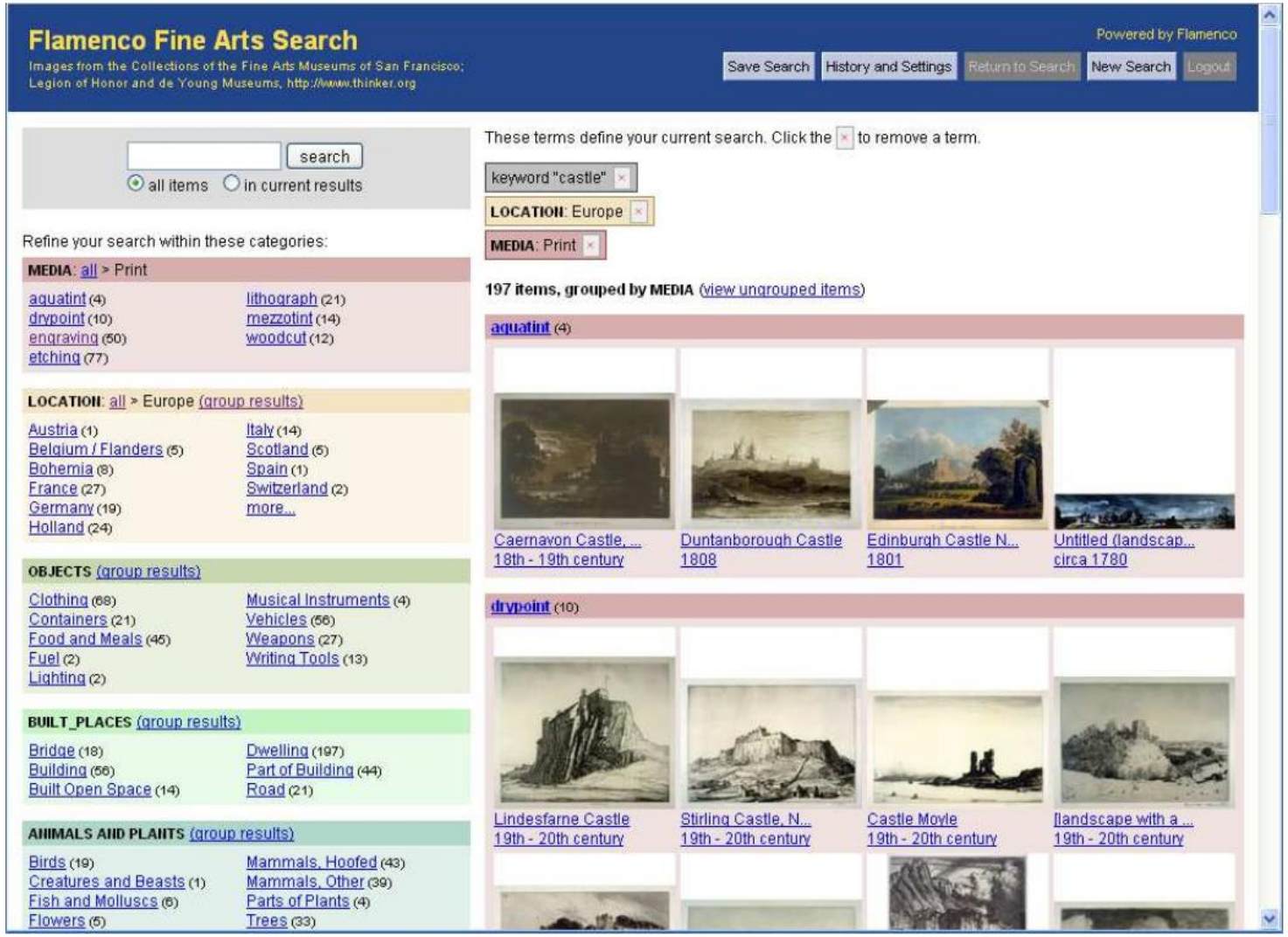
Retrieval results display



SuperBook IR system: **hierarchical** organization and blended surrogates (after document *zoom-in*)

Retrieval results display

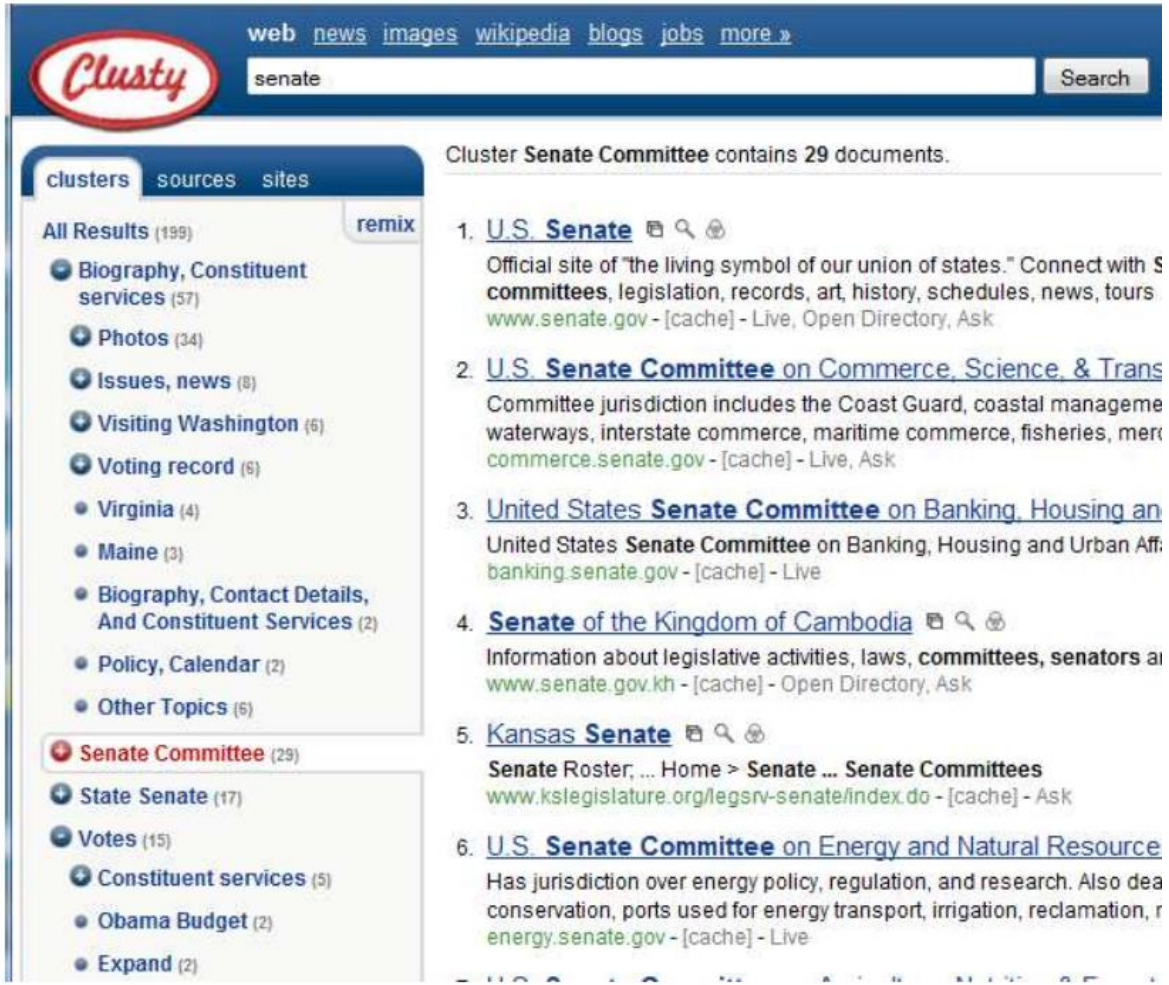
- *faceted navigation*
  How to implement?
  - extracting **topics** (topic modeling)
  - discovering **content patterns** (concept analysis)



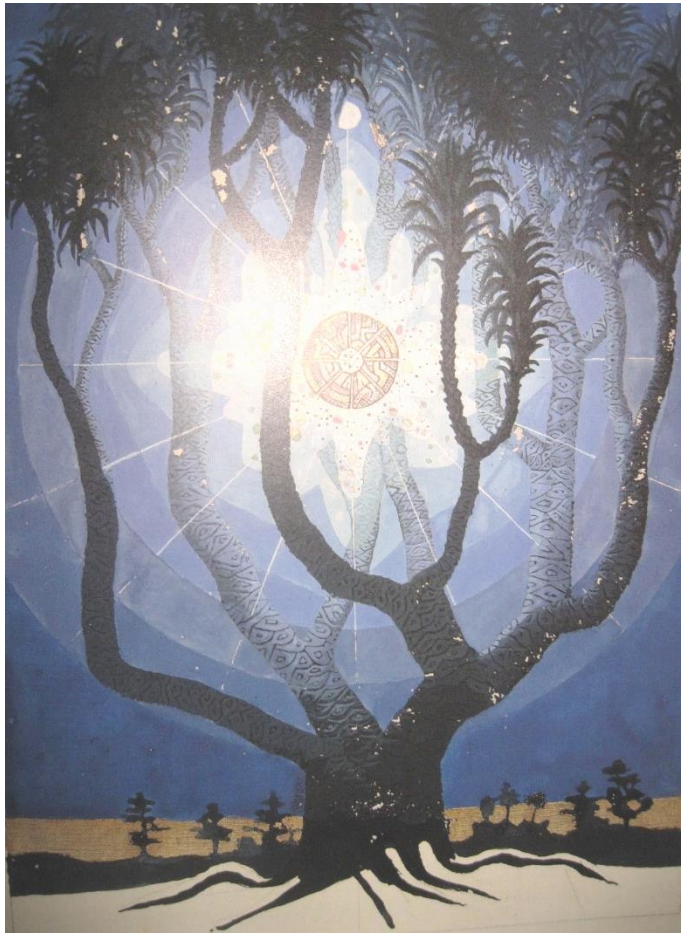
Retrieval results display

clustering: grouping documents according to some measure of similarity

- fully automatic yet not always predictable
- grouping can be led by specific doc properties (term similarity, temporal and style proximity)



Outline



IR definition and origins

Data structuring

IR system

Dynamic search model

Search interfaces

- query support
- results display
- **evaluation**

Evaluating search interfaces

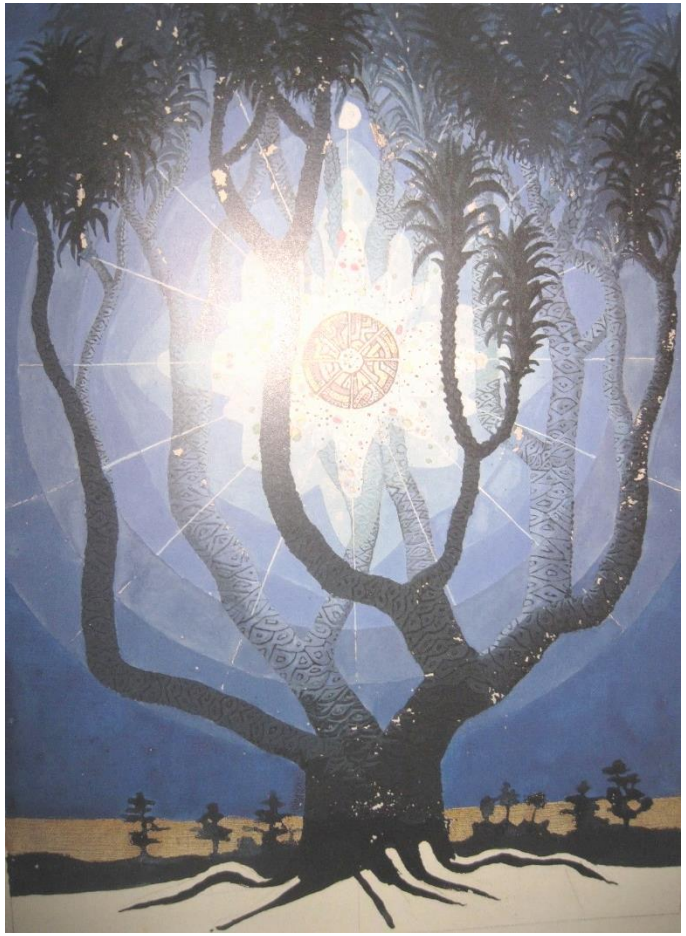
- Evaluation of IR systems entails different aspects:
  - **efficiency** and **efficacy** of their different components
    - the IR brain (**IR processing module**) to ensure proper answers in useful time
    - the **IR indexing module** to ensure updated and well-organized collections
    - what about **search interfaces**? Efficient and proper query support and result display also essential!
- User interface design: a field of **human-computer interaction** (HCI)
  - **user-centered design**: how people think about, respond to, and use technology
- Evaluation of user interface radically different from evaluating IR algorithms
  - a ranking algorithm can be evaluated by precision, recall and efficiency
  - user interface evaluation: subjective responses are as, if not more, important than quantitative measures
    - criteria: **speed, familiarity, aesthetics, preferred features, *perceived* ranking accuracy**
    - "*if a person has a choice between two systems, they will use the one they prefer*" (often the familiar one)

Evaluating user interfaces

How best to evaluate a user interface depends on the **IR system maturity**/stage:

- when starting with design and idea
  - **discount** usability methods: showing a few users different designs and asking pros and cons
  - **heuristic evaluation:** usability experts “walk through” a design and evaluate the functionality
  - *difficulties*: difficulty to mimic long-standing interactive search sessions
- advanced development stage
  - **longitudinal studies**
    - participants can test a new interface for an extended period of time
    - evaluation is based both on log analysis and questionnaires
- well-established and heavily-used user interfaces
  - **bucket testing** (**A/B testing**)
    - a randomly selected subset of the users is shown a new design their actions are logged and compared to another randomly control group that continues to use the existing interface

Outline



IR definition and origins

Data: structure levels

IR systems

Dynamic search model

Search interfaces

- query support
- results display
- evaluation

Thank You

rmch@tecnico.ulisboa.pt