Information Retrieval

Part 1
Organisation and Introduction

Week 1

Qianni Zhang

Course Aims

- **Indexing**: Representing the information content of documents through the use of e.g. stopword removal, stemming, and term weight calculation.
- **Retrieval**: Building models that select which information objects are relevant to a user's need. Models will include Boolean model, vector space model, probabilistic model, language model, inference network model, and relevance feedback model.
- **Evaluation**: Implementing and evaluating IR models, mainly with respect to effectiveness aspects.
- Tasks other than ad-hoc retrieval: e.g.: Classification, Summarisation

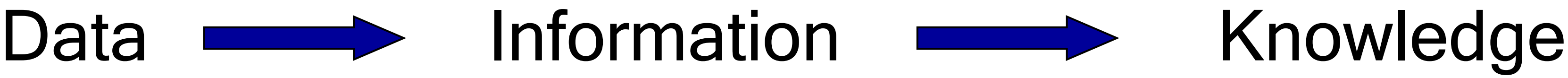
What will we cover?

- Organisation + Introduction
- Indexing and TF-IDF
- Retrieval Models I: VSM and BM25
- Retrieval Models II: Probabilistic IR, LM, DFR, Theory
- Retrieval Models III: Pagerank and others
- Retrieval Models IV: Relevance feedback
- Evaluation: Precision & Recall
- Semantic Search, DB+IR
- Classification and Summarisation
- *Visual Information Retrieval
- *Social Media Mining
- ……

Not necessarily in that order

Motivation and Applications

- There is too much unstructured “Data”
- Data is not the same as information



Data → Information → Knowledge

- Information is worthless if it cannot be found (extracted from raw data)
- Knowledge=Processed information retrieved from data

Motivation and Applications

How many times in one day (in average) do you use Google/Bing?

- IR is pervasive in daily life
- IR is fundamental for technological developments (and your coursework!)

Blended teaching this year

- Live online lectures
  - 2 hours lecture every week
  - Our main teaching delivery
  - Recording will be uploaded after the lecture finishes

- Labs on campus
  - 2 hours every week, every week from w2 to w11

Today

- Overview of Lectures and Labs in Weeks 1-12
    - Coursework
    - Important Dates
    - Assessment
- First lecture: Introduction to Information Retrieval (IR)

Course Work

- Assignment 1: Read, discuss and present research papers:
    - Teams of 2-4 summarise 2-3 research papers.
    - Starting references are provided, you search for further references.
    - Ideally identify the gap or difference between statements in the papers.

- Assignment 2: Search engine design:
    - Same teams of 2-4 design a search engine.
    - A document describing your search engine design.

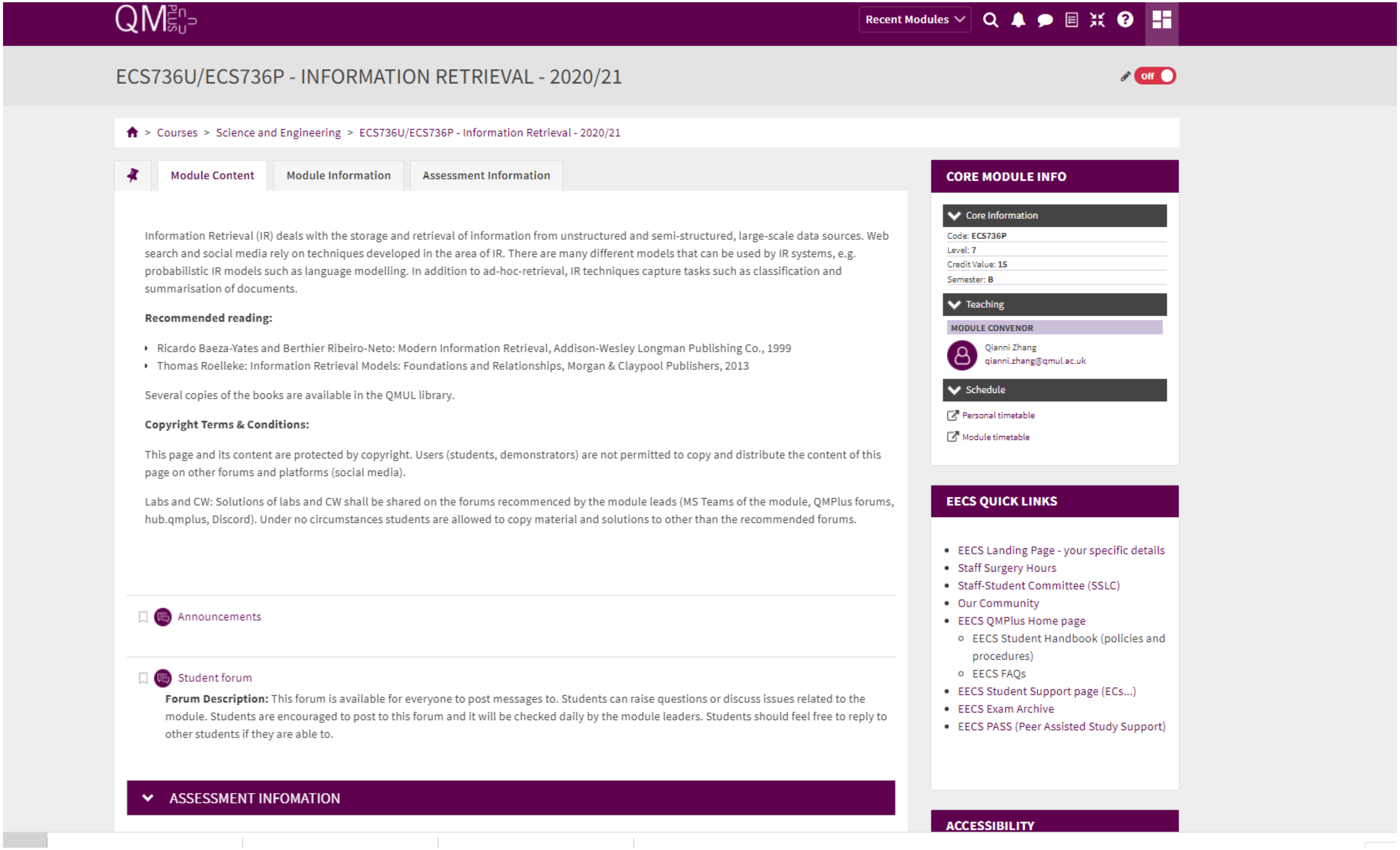
- Assignment 3: Design and Develop a search engine:
    - Same teams of 2-4 design and develop a search engine.
    - Technology: Python, Lucene (Java APIs) and/or any other tool (Gate, Lemur, Terrier, …, Elasticsearch, Solr)

Team work

- Coursework assignments are done in the same team of 2-4 students
- Deadline for building your group: **Monday week 3**
- Any student who has not chosen their group by then will be allocated to a group at random

Resources

All support materials for this course are available at: https://qmplus.qmul.ac.uk/course/view.php?id=15508



ECS736U/ECS736P - INFORMATION RETRIEVAL - 2020/21

Courses > Science and Engineering > ECS736U/ECS736P - Information Retrieval - 2020/21

Information Retrieval (IR) deals with the storage and retrieval of information from unstructured and semi-structured, large-scale data sources. Web search and social media rely on techniques developed in the area of IR. There are many different models that can be used by IR systems, e.g. probabilistic IR models such as language modelling. In addition to ad-hoc-retrieval, IR techniques capture tasks such as classification and summarisation of documents.

Recommended reading:

- Ricardo Baeza-Yates and Berthier Ribeiro-Neto: Modern Information Retrieval, Addison-Wesley Longman Publishing Co., 1999
- Thomas Roelleke: Information Retrieval Models: Foundations and Relationships, Morgan & Claypool Publishers, 2013

Several copies of the books are available in the QMUL library.

Copyright Terms & Conditions:

This page and its content are protected by copyright. Users (students, demonstrators) are not permitted to copy and distribute the content of this page on other forums and platforms (social media).

Labs and CW: Solutions of labs and CW shall be shared on the forums recommenced by the module leads (MS Teams of the module, QMPlus forums, hub.qmplus, Discord). Under no circumstances students are allowed to copy material and solutions to other than the recommended forums.

Announcements

Student forum

Forum Description: This forum is available for everyone to post messages to. Students can raise questions or discuss issues related to the module. Students are encouraged to post to this forum and it will be checked daily by the module leaders. Students should feel free to reply to other students if they are able to.

ASSESSMENT INFOMATION

CORE MODULE INFO

Core Information

Code: **ECS736P**

Level: **7**

Credit Value: **15**

Semester: **B**

Teaching

MODULE CONVENOR

Qianni Zhang
qianni.zhang@qmul.ac.uk

Schedule

Personal timetable

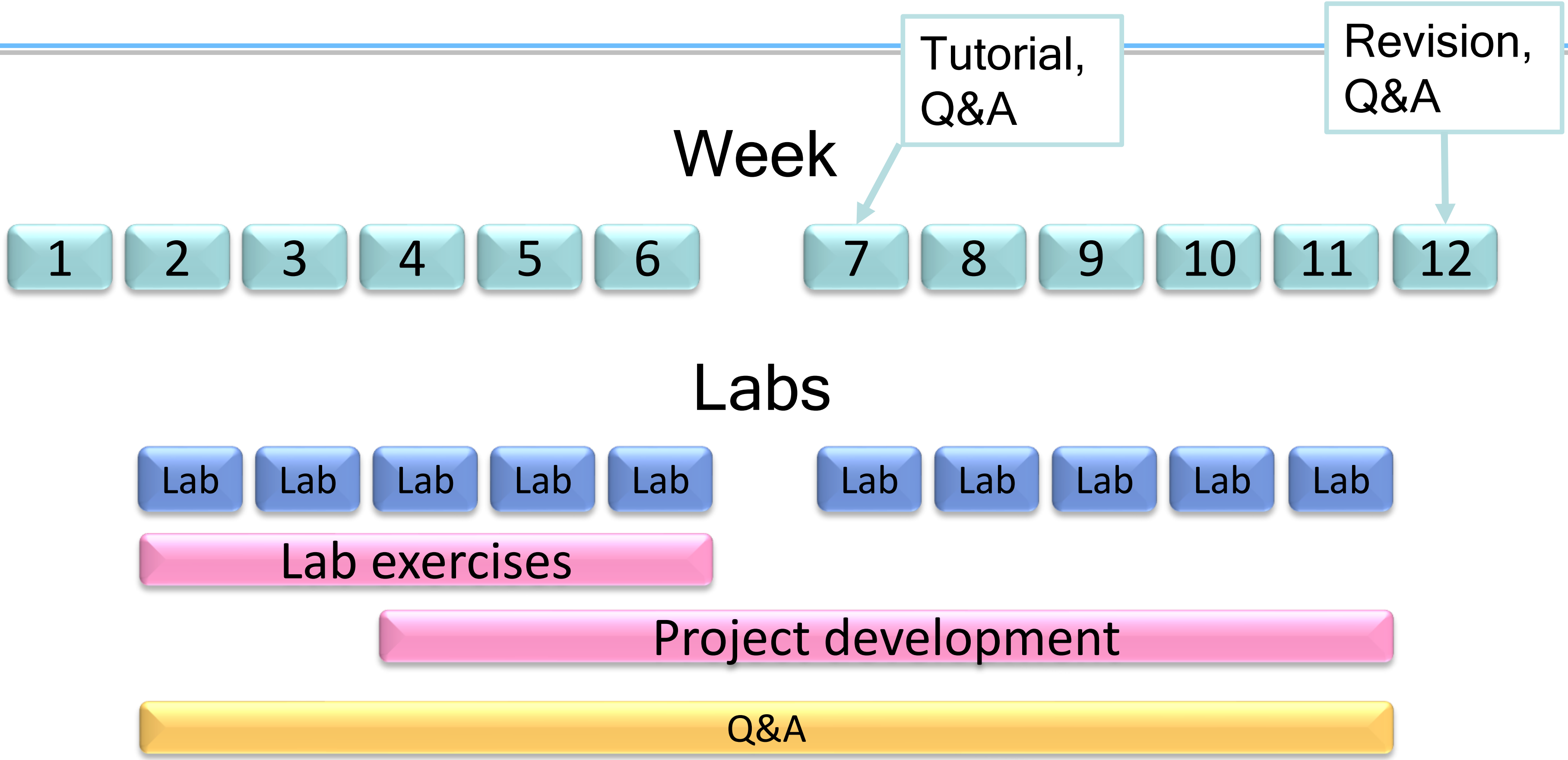
Module timetable

EECS QUICK LINKS

- EECS Landing Page - your specific details
- Staff Surgery Hours
- Staff-Student Committee (SSLC)
- Our Community
- EECS QMPlus Home page
  - EECS Student Handbook (policies and procedures)
  - EECS FAQs
- EECS Student Support page (ECs...)
- EECS Exam Archive
- EECS PASS (Peer Assisted Study Support)

ACCESSIBILITY

Lab work



Tutorial, Q&A

Revision, Q&A

Week

1 2 3 4 5 6 7 8 9 10 11 12

Labs

Lab Lab Lab Lab Lab Lab Lab Lab Lab Lab

Lab exercises

Project development

Q&A

NOTE: The above programme is tentative, check the website for exact details!

Important Dates

- **Week 1-2:**

  Build your team of 2-4.

- **Week 4:**

  Presentation of research papers.

- **Week 8:**

  Submission of your search engine design.

  The labs and lectures enable you to develop the design.

- **Week 12:**

  Presentation and demo of your search engine.

Assessment

- **65%** final exam (open book)
- **35%** coursework
  - 10%: Research paper presentation: slides, video recording
  - 10%: Search engine design: design document
  - 15%: Search engine presentation and demo: slides, source code package, video recording
- Marking: group marks unless strong evidence for individual contributions; submission details see QM+.

Course Work: Topics and Starting References

- Term-weighting approaches in automatic text retrieval (Salton and Buckley, IP&M, 1988)
- Indexing by latent semantic indexing (S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, JASIS, 1990)
- Inference Networks for Document Retrieval (H.R. Turtle, and B. Croft, SIGIR 1990)
- Towards an information logic (C. J. van Rijsbergen, SIGIR, 1989)
- Okapi at TREC-3 (S. E. Robertson, S. Walker, M. M. Hancock-Beaulieu, and M. Gatford, TREC-3, 1994)
- Pivoted document length normalisation (A. Singhal, C. Buckley, and M. Mitra, SIGIR, 1996)
- Self-indexing inverted files for fast text retrieval (A. Moffat and J. Zobel, ACM TOIS, 1996)
- Advantages of query biased summaries in information retrieval (A. Tombros and M. Sanderson, SIGIR, 1998)
- A language modelling approach to information retrieval (Ponte and Croft, SIGIR, 1998)
- The anatomy of a large-scale hypertextual web search engine (Brin and Page, WWW7, 1998)

Course Work: Topics and Starting References

- A study of smoothing methods for language models applied to ad hoc information retrieval (C. Zhai and J. Lafferty, SIGIR, 2001)
- Combining document representations for known-item search (P. Ogilvie, and J. Callan, SIGIR, 2003)
- Stuff I've seen: A system for personal information retrieval and re-use (S. Dumais, E. Cutell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins, SIGIR, 2003)
- Parsimonious language models for information retrieval, (Hiemstra etal, SIGIR 2004)
- A General Matrix Framework for Modelling Information Retrieval (T. Roelleke, T. Tsikrika, and G. Kazai, IP&M, 2006)
- Harmony Assumptions in Information Retrieval and Social Networks (T. Roelleke etal, Computer Journal 2015)

YOU CAN PROPOSE YOUR OWN STARTING PAPERS ON INDEXING OR RETRIEVAL.
http://www.sigir.org/resources.html

Recommended reading

- Lecture handouts
- Books:
  - Introduction to Information Retrieval (C. D. Manning)
  - Modern Information Retrieval (R. Baeza-Yates and Berthier Ribeiro-Neto)
  - Information Retrieval Models: Foundations and Relationships (T. Roelleke)

- Introductory texts which covers much of the course
- Books are “recommended”
- Not compulsory, but worth considering if you wish to find out more details about the topics covered in the course
- Online version available and in the library

Module team

Lecturer and MO: Dr Qianni Zhang (qianni.zhang@qmul.ac.uk)

- Module organisation, admin
- Lecturing
- Assessment …

Teaching fellow: Bilal Hassan (b.hassan@qmul.ac.uk)

- Assistance in module organisation
- Tutorials
- Assessment …

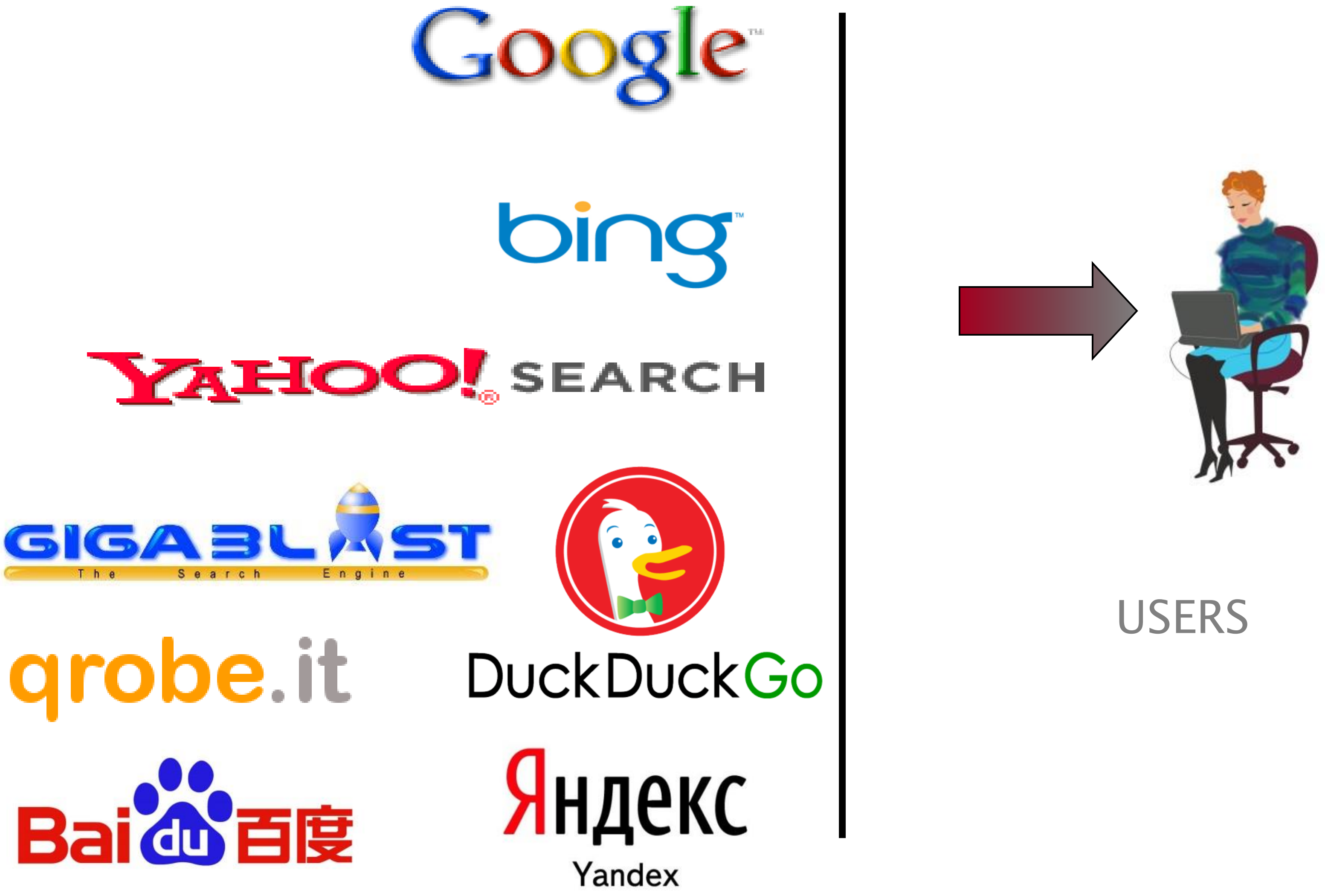
Demonstrators: lab supervision, Q&A, assessment, …

- Mr Edgar Giussepi Lopez Molina (e.g.lopezmolina@qmul.ac.uk)
- Miss Weiwei Cui (w.cui@qmul.ac.uk)
- Mr Ji Lin (j.lin@qmul.ac.uk)
- Mr Kit Bransby (k.m.bransby@qmul.ac.uk)
- A few more to join

Introduction to Information Retrieval (IR)

- Terminology
- Information Need
- Retrieval Tasks
- A Conceptual Model for IR
- Document and Document Representation
- Queries
- Best-match retrieval
- History
- Topics in IR
- Information Retrieval vs Information Extraction vs Web Search
- Important forums (Conferences and Journals)

Text-based systems



USERS

Database and the web?

What are some limitations of Database Systems?

A (Simple) Database Example

Student Table

Student ID	Last Name	First Name	Department ID	email
1	Maryam	Karimzadehgar	CS	mkarimz2@qmul.ac.uk
2	Peters	jordan	EE	kj@qmul.ac.uk
3	Smith	Chris	EE	sc@qmul.ac.uk
4	Smith	John	CS	Sj@qmul.ac.uk

Department Table

Department ID	Department
EE	Electronic Engineering
CS	Computer Science

Course Table

Course ID	Course Description
cs736	Information Technology
ee750	Communication

Enrollment Table

Student ID	Course ID	Grades
1	cs736	90
1	ee750	75
2	cs736	95
2	ee750	80
3	cs736	60
4	ee750	77

Databases vs. IR

- Format of data:
  - DB: Structured data. Clear semantics based on a formal model.
  - IR: Mostly unstructured. Free text.
- Queries:
  - DB: Formal (like SQL)
  - IR: often expressed in natural language (keywords search)
- Result:
  - DB: exact result
  - IR: Sometimes relevant, often not

Terminology

- **General:** Information Retrieval, Information Need, Query, Retrieval Model, Retrieval Engine, Search Engine, Relevance, Relevance Feedback, Evaluation, Information Seeking, Human-Computer-Interaction, Browsing, Interfaces, Ad-hoc Retrieval, Filtering

- **Related:** Document Management, Knowledge Engineering

- **Expert:** term frequency (TF), document frequency, inverse document frequency (IDF), vector-space model (VSM), probabilistic model, BM25 (Best-Match Version 25), DFR (Divergence from Randomness), page rank, stemming, precision, recall

Information Need

- Example of an information need in the context of the world wide web:

Find all documents (*information!*) about universities in the UK that

(1) offer master degrees in Information Retrieval and

(2) are registered with ACM SIGIR.

The information (*the document!*) should include full curriculum, fees, student campus, e-mail and other contact details.

- Formal representation of an information need = **Query**

Information Retrieval: An Informal Definition

Representation, storage, organisation and access of **information**

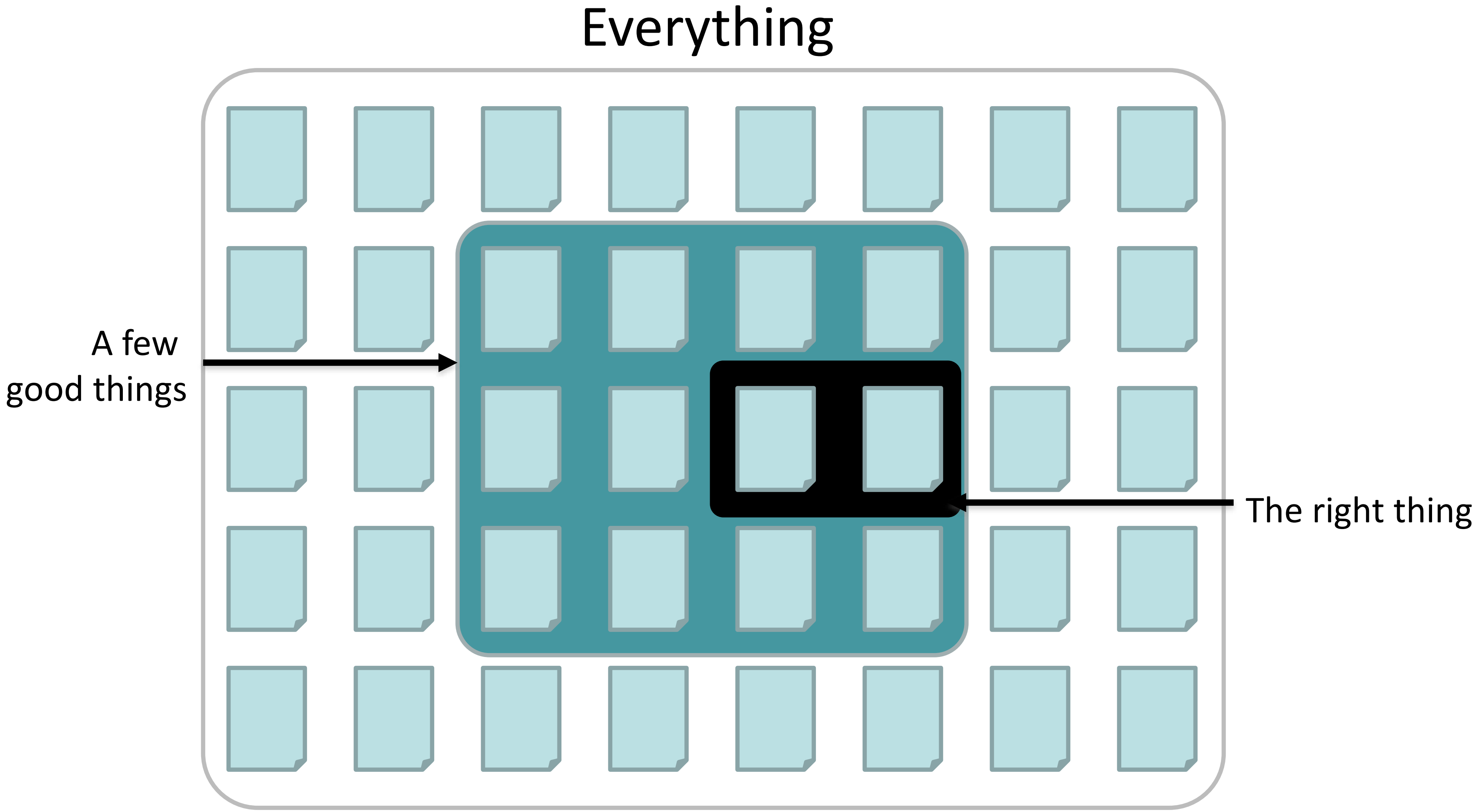
(information items, information objects, documents).

Find relevant (useful) information

Goal of an IR system

- **Recall**: Retrieve all relevant documents (e.g. legal)
  - Retrieve as few non-relevant documents as possible.
- **Precision**: Retrieve the most relevant documents (e.g. web)
  - Retrieve relevant documents before non-relevant documents.

Scope of Information Needs


Everything

A few good things

The right thing

Information Retrieval / Data Retrieval

	Information Retrieval	Data Retrieval
Matching	vague	exact
Model	probabilistic	deterministic
Query language	natural	artificial
Query specification	incomplete	complete
Items wanted	relevant	all (matching)
Error handling	insensitive	sensitive

What is IR?

- Goal: Find the documents most relevant to a given Query

- Dealing with notions of:
  - Collection of documents
  - Query (User’s information need)
  - Notion of Relevancy

Types of Information Needs

- Retrospective (Ad-hoc Querying)
  - “Searching the past”
  - Different queries posed against a static collection
- Prospective (Filtering)
  - “Searching the future”
  - Static query posed against a dynamic collection
  - Time dependent

Retrospective Searches (I)

- Topical search

  Identify positive accomplishments of the Hubble telescope since it was launched in 1991.

  Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.

- Open-ended exploration

  Who makes the best chocolates?

  What technologies are available for digital reference desk services?

Retrospective Searches (II)

- Known-item search

  Find Qianni Zhang’s homepage.

  What’s the ISBN number of “Modern Information Retrieval”?

- Question answering

“Factoid”	Who discovered Oxygen? When did Hawaii become a state? Where is Ayer’s Rock located? What team won the World Series in 1992?
“List”	What countries export oil? Name U.S. cities that have a “Shubert” theater.
“Definition”	Who is Aaron Copland? What is a quasar?

Prospective “Searches”

- Filtering
  - Make a binary decision about each incoming document
- Routing / Multi-label Classification
  - Sort incoming documents into different bins

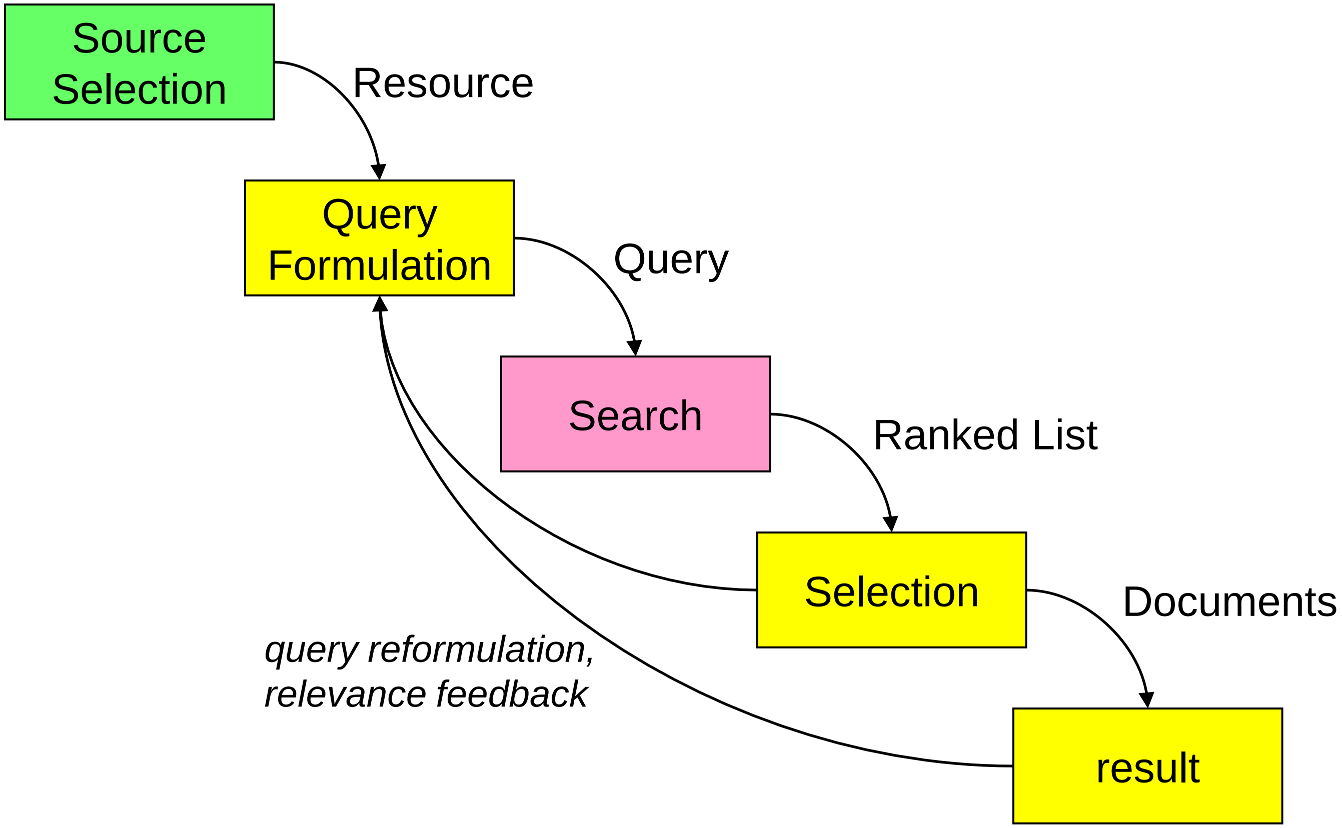
Relevance and Types of Relevance

- How well information addresses your needs
  - Harder to pin down than you think!
  - Complex function of user, task, and context
- Types of relevance:
  - Topical relevance: is it about the right thing?
  - Situational relevance: is it useful?

What Types of Documents / Information / Media?

- Text (Documents)
- XML and structured documents
- Images
- Audio (sound effects, songs, etc.)
- Video
- Source code
- Applications/Web services

The Information Retrieval Cycle



Source Selection

Resource

Query Formulation

Query

Search

Ranked List

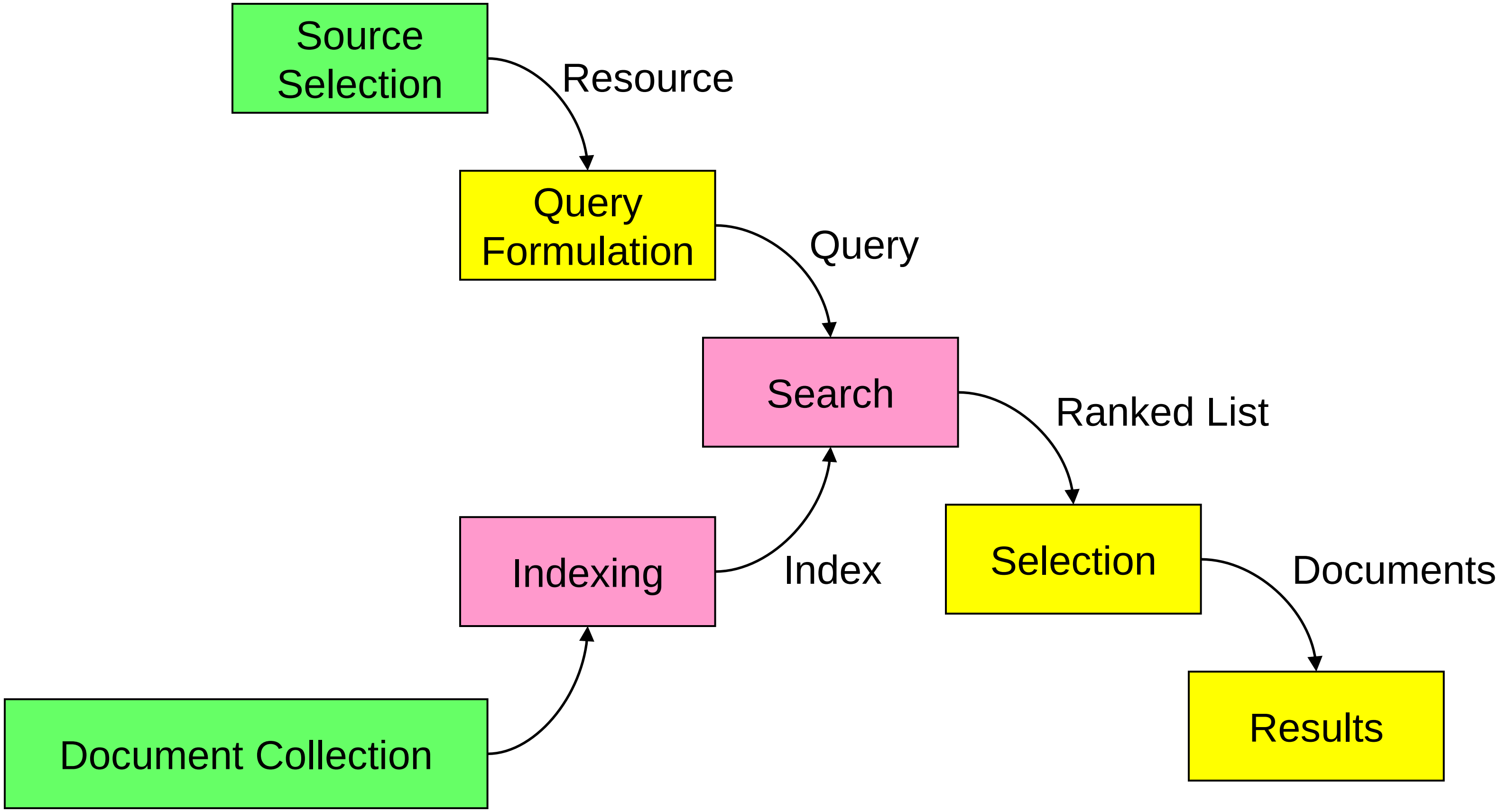
Selection

Documents

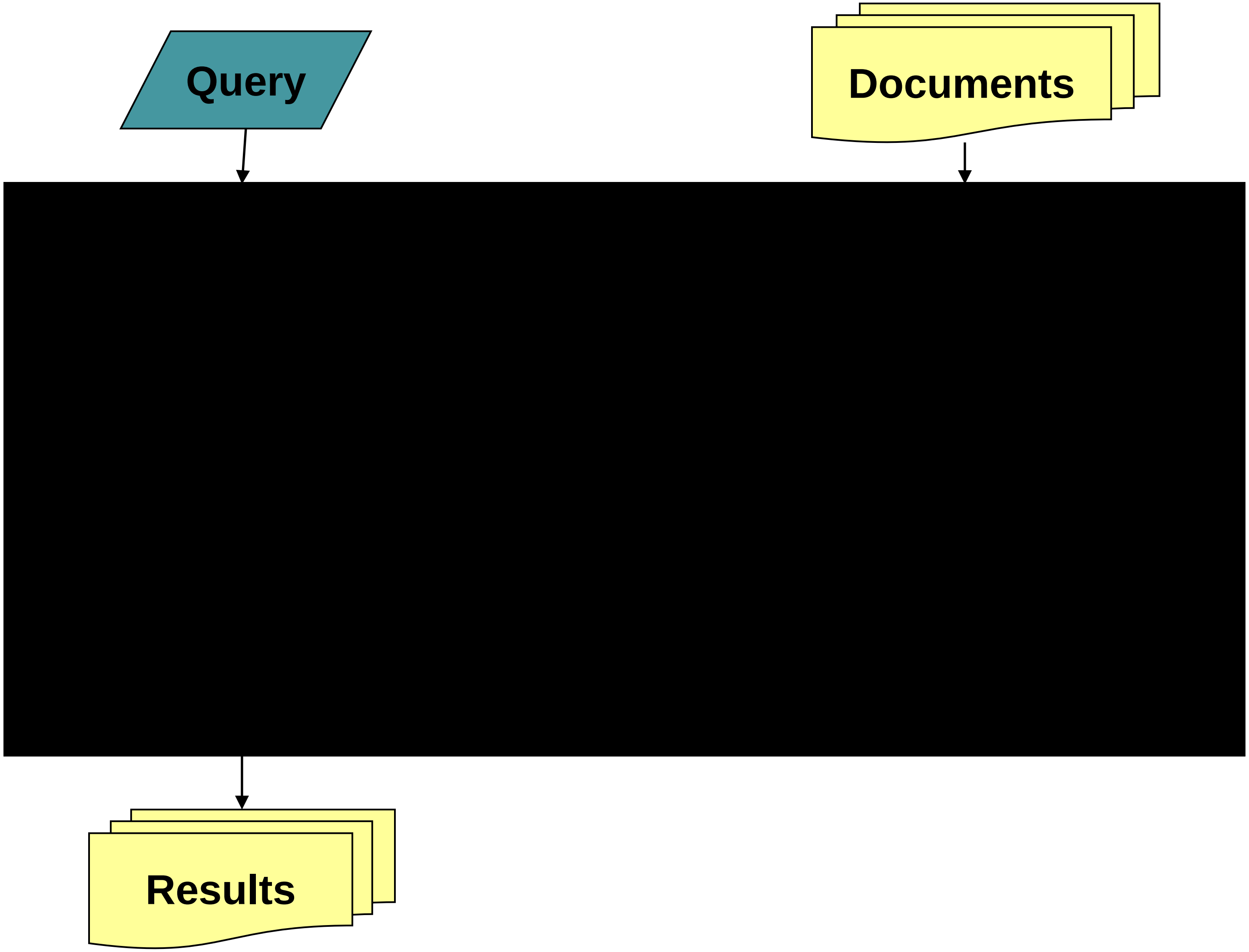
result

query reformulation, relevance feedback

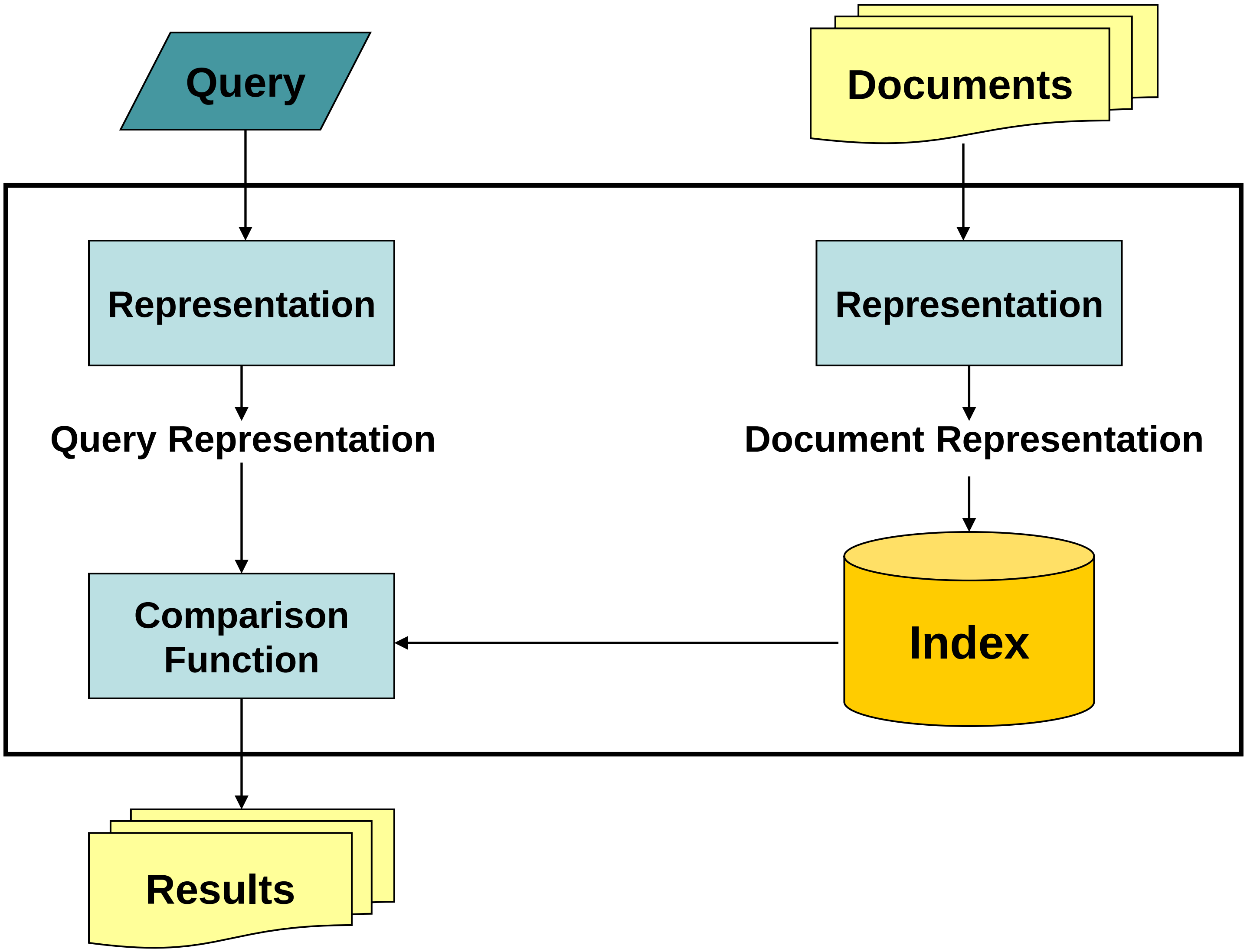
Search Process



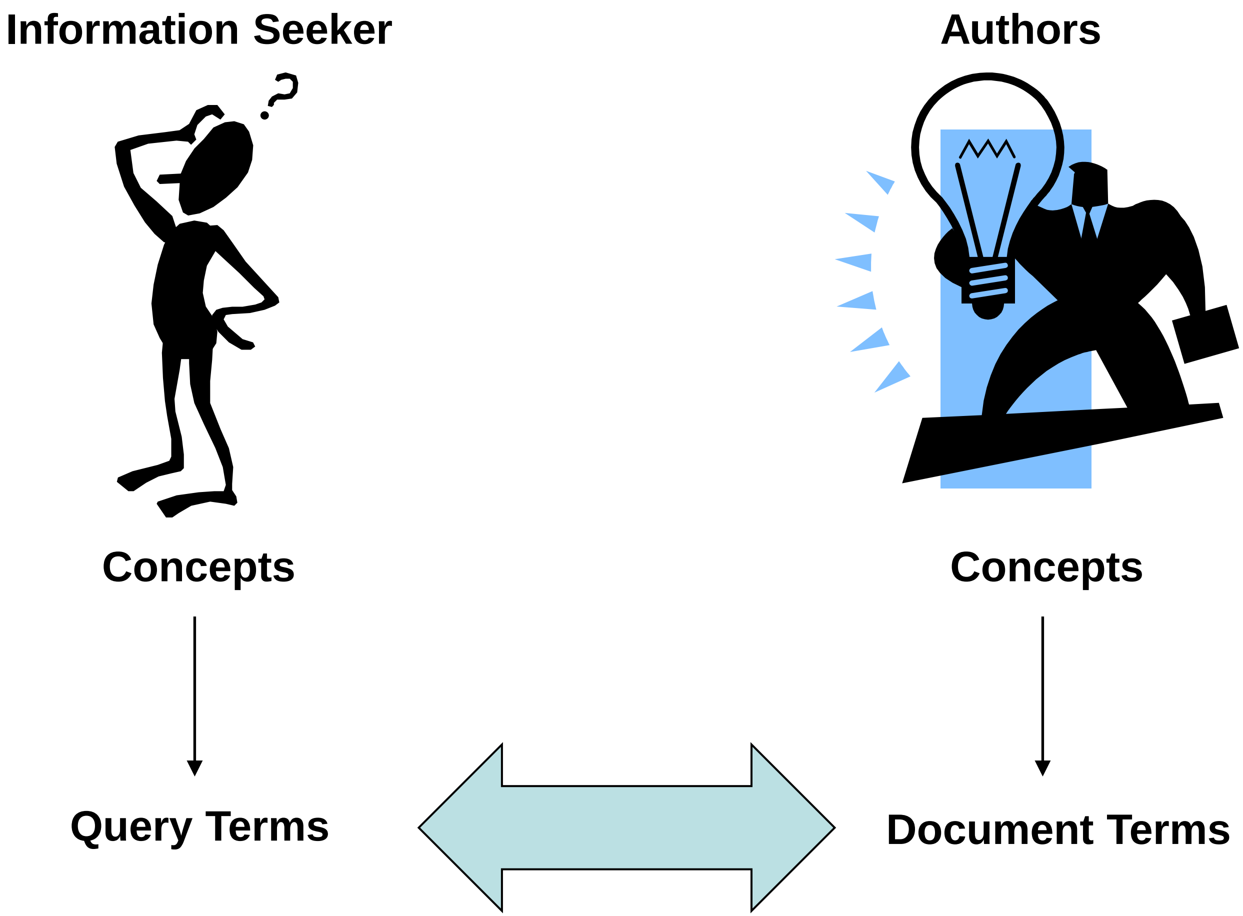
The IR Black Box



Inside The IR Black Box



The Central Problem in IR



Information Seeker

Authors

Concepts

Concepts

Query Terms

Document Terms

Do these represent the same concepts?

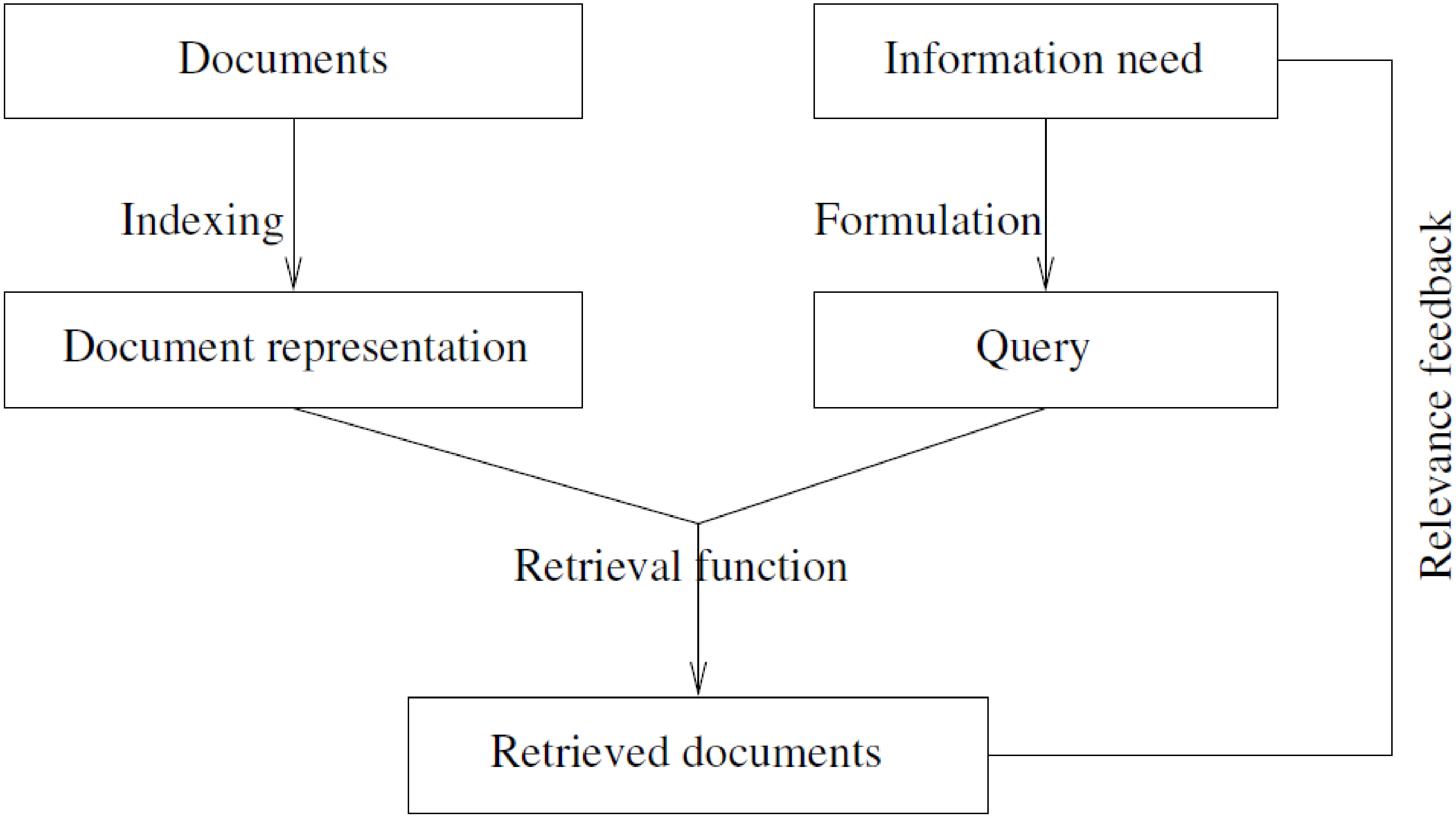
Why IR is hard? Because Language is hard!!!

Ambiguity: e.g. Good Friday
Synonymy: e.g. classify/categorise
Polysemy: e.g. hamburger, bank
Morphology: e.g. sailing, sailor, sails
Paraphrase: different text, same meaning
Anaphora: e.g. he, she, it
Pragmatics: children vs grown-ups

How do we represent documents?

- Remember: computers don’t “understand” anything!
- “Bag of words” representation:
  - Break a document into words
  - Disregard order, structure, meaning, etc. of the words
  - Simple, yet effective!

A Conceptual Model for IR



Documents and Document Representations

Documents

- Unit of retrieval
- A passage of free text
  - composed of text, strings of characters from an alphabet
  - composed of natural language:

    newspaper articles, journal paper, dictionary definition, e-mail messages
  - size of documents:

    arbitrary, newspaper article vs journal article vs e-mail

- Sub-document can also be a unit of retrieval (passage, XML element, answer to a question)

Documents and Document Representations

Document Representation

- Free-text representation: extracted directly from text, good performance in broad domains.
- Controlled vocabulary representation: most concise representation, good performance in narrow domains with limited number of (expert) users.

- Full-text representation: most complete representation, optimal performance, huge resource requirements.
- Reduced (partial) content representation: stopwords, stemming, noun phrases, compression.

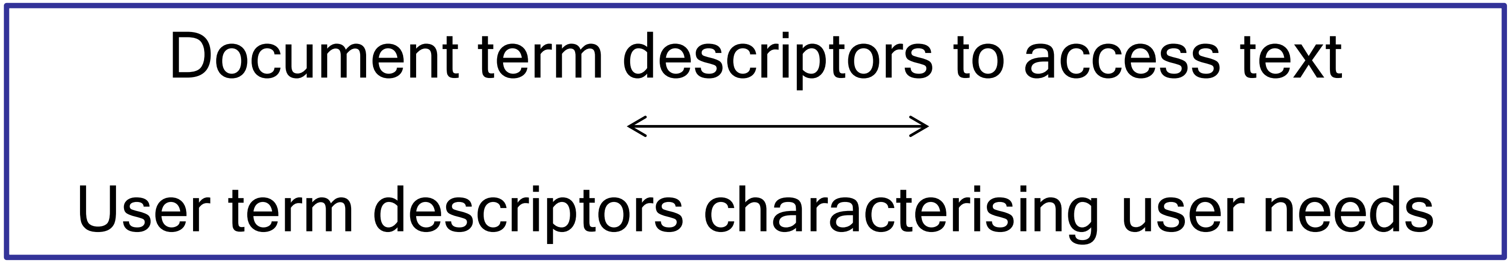
- Structure representation: chapter, section, paragraph.
- Semantic representation: actors, employees, workedWith.

Queries

- Information Need
- Simple queries
  - composed of two or three, perhaps of dozen of keywords
  - e.g. as in web retrieval
- Boolean queries
  - ‘neural network AND speech recognition’
  - e.g. as in online catalog and patent search
- Context queries
  - proximity search, phrase queries
  - e.g. ‘neural’ and ‘network’ distance at most 5 words

Best-Match Retrieval

- Compare the terms in a document and query
- Compute "similarity" between each document in the collection and the query based on the terms they have in common
- Sorting the document in order of decreasing similarity with the query
- The outputs are a ranked list and displayed to the user - the top ones are more relevant as judged by the system



Document term descriptors to access text

User term descriptors characterising user needs

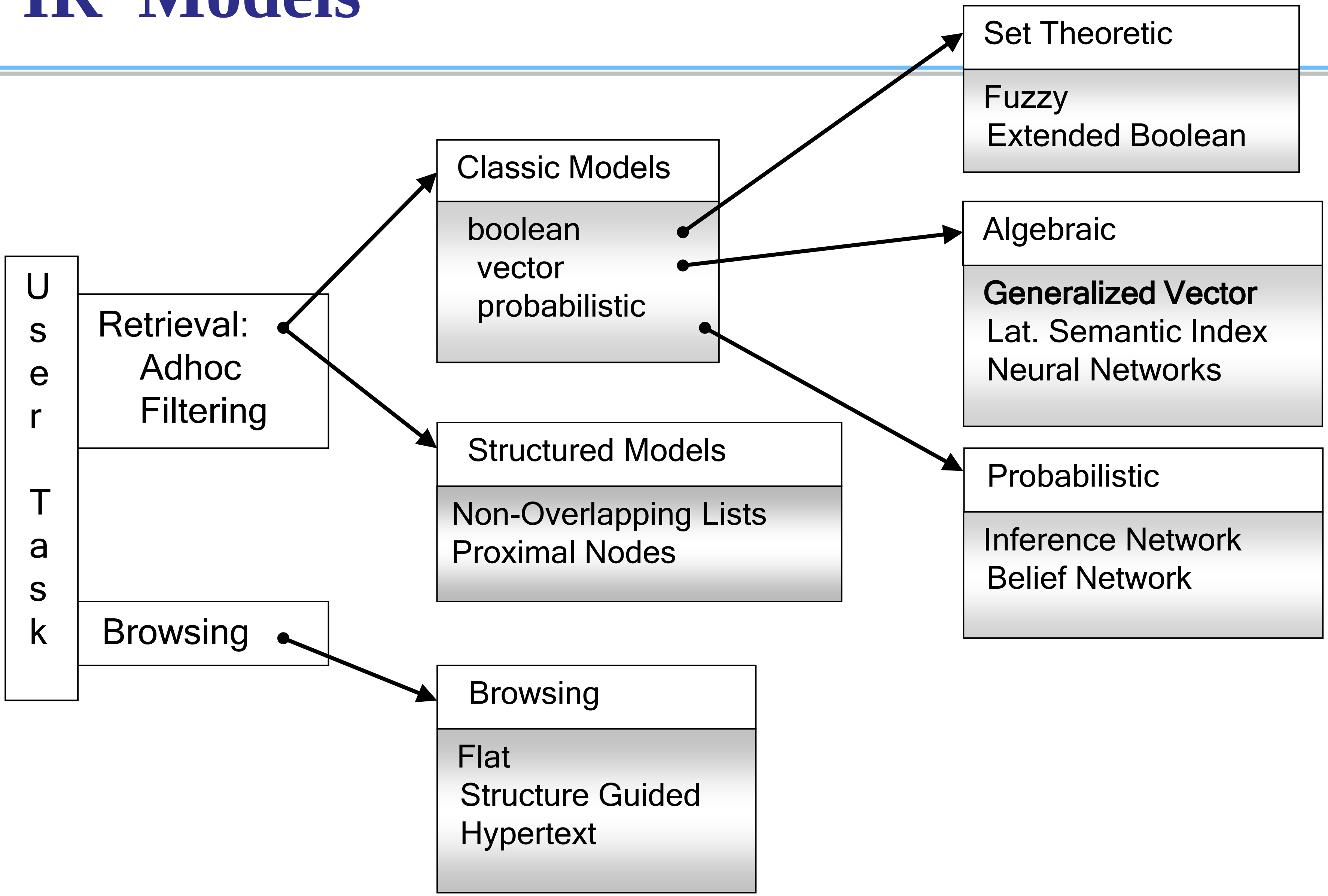
Tasks of IR

- Index the documents in the collection (offline)
- Process the query
- Measure Similarity and compute ranking scores
  - Find documents most closely matching the query (relevant documents)
- Display results
  - E.g., user may refine the query (feedback)

Similarity Models

- Boolean model
- Vector-space model
- Probabilistic model
- Language modelling

IR Models



User Task

Retrieval:
Adhoc
Filtering

Browsing

Classic Models

boolean
vector
probabilistic

Structured Models

Non-Overlapping Lists
Proximal Nodes

Browsing

Flat
Structure Guided
Hypertext

Set Theoretic

Fuzzy
Extended Boolean

Algebraic

Generalized Vector
Lat. Semantic Index
Neural Networks

Probabilistic

Inference Network
Belief Network

Search Output

- What now?
  - User identifies relevant documents for “delivery”
  - User issues new query based on content of result set
- What can the system do?
  - Assist the user to identify relevant documents
  - Assist the user to identify potentially useful query terms

Selection Interfaces

- One dimensional lists
  - What to display? title, source, date, summary, ratings, ...
  - What order to display? similarity score, date, alphabetic, ...
  - How much to display? number of hits
  - Other aids? related terms, suggested queries, …
- Two+ dimensional displays
  - Clustering, projection, contour maps, VR
  - Navigation: jump, pan, zoom

Query Expansion / Enrichment

- Relevance feedback
  - User designates “more like this” documents
  - System adds terms from those documents to the query
- Manual reformulation
  - Initial result set leads to better understanding of the problem domain
  - New query better approximates information need
- Automatic query suggestion

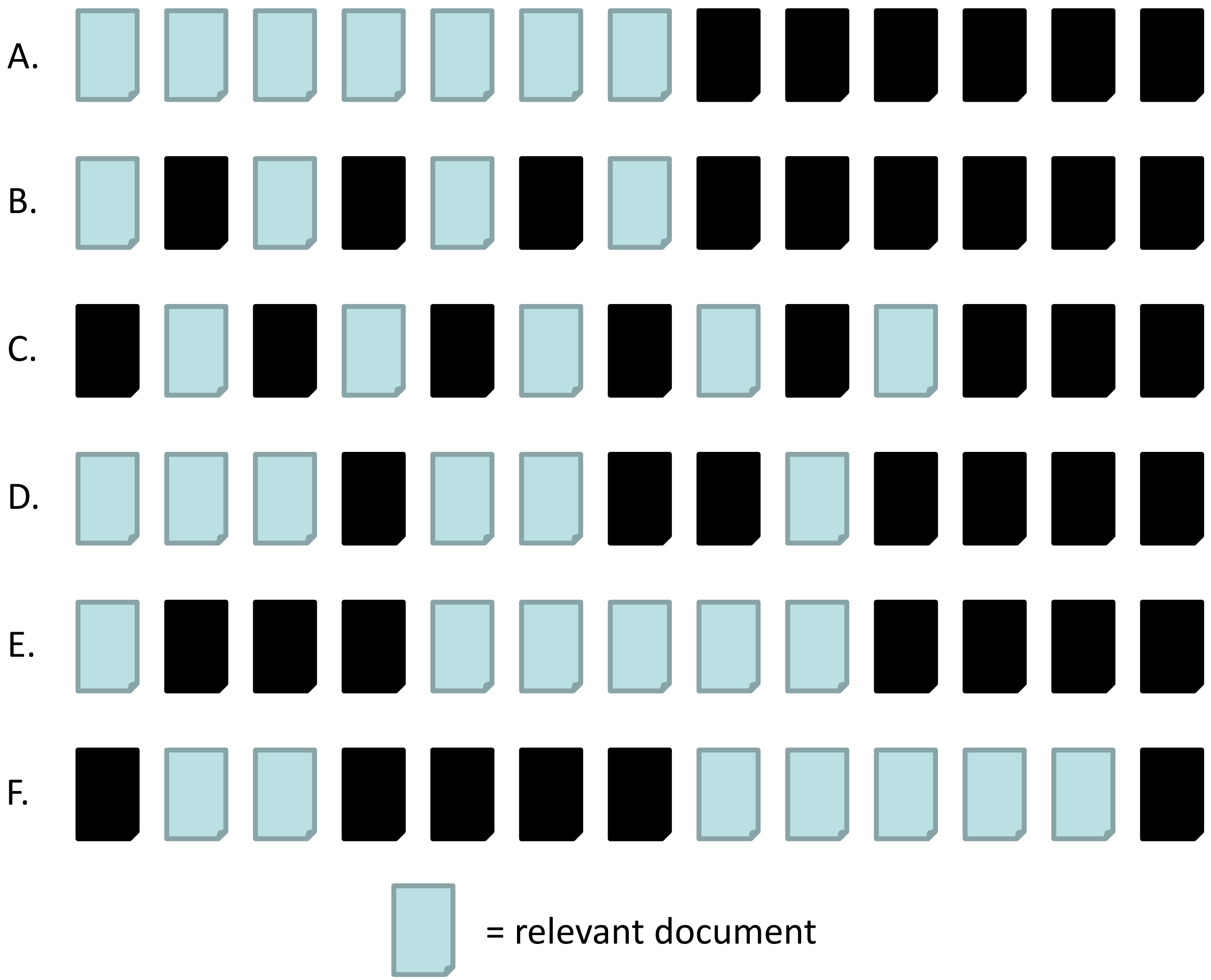
Evaluating IR Systems

- User-centered strategy
  - Recruit several users
  - Observe each user working with one or more retrieval systems
  - Measure which system works the “best”
- System-centered strategy
  - Given documents, queries, and relevance judgments
  - Try several variants of the retrieval method
  - Measure which variant is more effective

Good Effectiveness Measures

- Capture some aspect of what the user wants
- Have predictive value for other situations
- Easily replicated by other researchers
- Easily compared

Which is the Best Rank Order?



History

- **Manual IR in libraries:** manual indexing; manual categorisation
- **70ies and 80ies:** Automatic IR in libraries
- **90ies:** IR on the web and in digital libraries

Success factors: Response time, coverage, interactivity, low (no!) costs, precision-oriented (you do not "feel" the recall)

precision ≈ correctness, recall ≈ completeness

(Some) Topics in IR

- Retrieval models (ranking function, learning to rank, machine learning)
- Text processing (“Indexing”): NLP / understanding (language models)
- Interactivity and users
- Efficiency, compression, MapReduce, Scalability
- Distributed IR (data fusion, aggregated search, federated search)
- Multimedia: image, video, sound, speech
- Evaluation including crowd-sourcing
- Web retrieval and social media search
- Cross-lingual IR (FIRE), Structured Data (XML),
- Digital libraries, Enterprise Search, Legal IR, Patent Search, Genomics IR

References and more information

Conferences:

SIGIR, CIKM, SPIRE, FQAS, BCS-IRSG (ECIR), RIAO, SAC-IAR, IIIX, EDCL, JCDL, IRF, ICTIR

http://www.sigir.org/events/events-upcoming.html

Journals:

TOIS, IP&M, IR, JDOC, JASIST

http://www.sigir.org/resources.html