UNIVERSITY OF TWENTE.

FOUNDATIONS OF INFORMATION RETRIEVAL
INTRODUCTION & HISTORY

Theo Huibers
Djoerd Hiemstra
Dolf Trieschnigg

WELCOME EVERYONE!

Overview

1) Who's who?
   - BSc before (UT / EU / Elsewhere)
   - MSc. Now? (CS / HMI / other)
   - Programming experience?

2) What will we do this course?

3) What is Information Retrieval? + History

SEARCH ENGINE TECHNOLOGY

Name: Theo Huibers
From: Thaesis & University of Twente

Name: Djoerd Hiemstra
From: Searsia & University of Twente

Name: Dolf Trieschnigg
From: Nedap & University of Twente

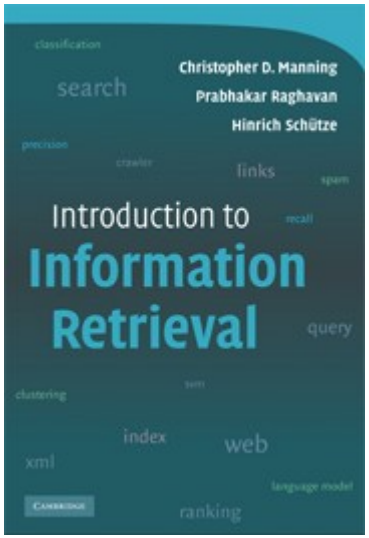
PROGRAM ON CANVAS https://canvas.utwente.nl/courses/1778

Course summary:

Date	Details	
Wed, 5 Sep 2018	FIR Lecture 1, Welcome (Chapter 1, Boolean Retrieval)	10:45 to 12:30
Mon, 10 Sep 2018	FIR Discussion Meeting	8:45 to 10:30
Tue, 11 Sep 2018	Jupyter Notebook 1, Getting Started	due by 23:59
Wed, 12 Sep 2018	FIR Lecture 2, Conceptual Indexing (Chapter 2, The term vocabulary & postings; Chapter 3, Dictionaries & tolerant retrieval; Chapter 6, Scoring & term weighting)	10:45 to 12:30
Mon, 17 Sep 2018	FIR Discussion Meeting	8:45 to 10:30
Tue, 18 Sep 2018	Jupyter Notebook 2, Index and UI	due by 23:59
Wed, 19 Sep 2018	FIR Lecture 3, Retrieval Models (Chapter 6, The Vector Space Model; Chapter 9, Relevance feedback & query expansion; Chapter 11, Probabilistic Information Retrieval; Chapter 12, Language Models; Chapter 21, Link Analysis)	10:45 to 12:30
Mon, 24 Sep 2018	FIR Discussion Meeting	8:45 to 10:30
Tue, 25 Sep 2018	Jupyter Notebook 3, TREC Baseline Run	due by 23:59
Wed, 26 Sep 2018	FIR Lecture 4, Evaluation (Chapter 8, Evaluation in Information Retrieval)	10:45 to 12:30
Mon, 1 Oct 2018	FIR Discussion Meeting	8:45 to 10:30
Tue, 2 Oct 2018	Jupyter Notebook 4, Calculate Evaluation Measures	due by 23:59
Wed, 3 Oct 2018	FIR Lecture 5, Technical Indexing (Chapter 4, Index Construction; Chapter 5, Index Compression; Chapter 7, Computing scores in a search system)	10:45 to 12:30
Mon, 8 Oct 2018	FIR Discussion Meeting	8:45 to 10:30
Tue, 9 Oct 2018	Jupyter Notebook 5, Models & Analyzers	due by 23:59
Wed, 10 Oct 2018	FIR Lecture 6, Guest Lecture by Thijs Westerveld (IR for children at WizeNoze)	10:45 to 12:30
Mon, 15 Oct 2018	FIR Discussion Meeting	8:45 to 10:30
Tue, 16 Oct 2018	Jupyter Notebook 6 Part 1, Multi-field Index, & Your own experiments	due by 23:59
Wed, 17 Oct 2018	FIR Lecture 7, Guest Lecture by Daan Odijk (IR and Personalisation at RTL Nieuws en Videoland)	10:45 to 12:30
Tue, 23 Oct 2018	Jupyter Notebook 6, Part 2 and 3	due by 23:59

STUDY MATERIAL

Christopher Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, *Cambridge University Press*. ISBN 0521865719, 2008.
http://informationretrieval.org



COMMUNICATION: UT MASTODON

A single channel for announcements + questions.

Privacy settings per post.

Students that are not in the course can help.

Connect to more than 1.5 million users

UT will not sell your data / who ads.

Moderated (report harassment, please).

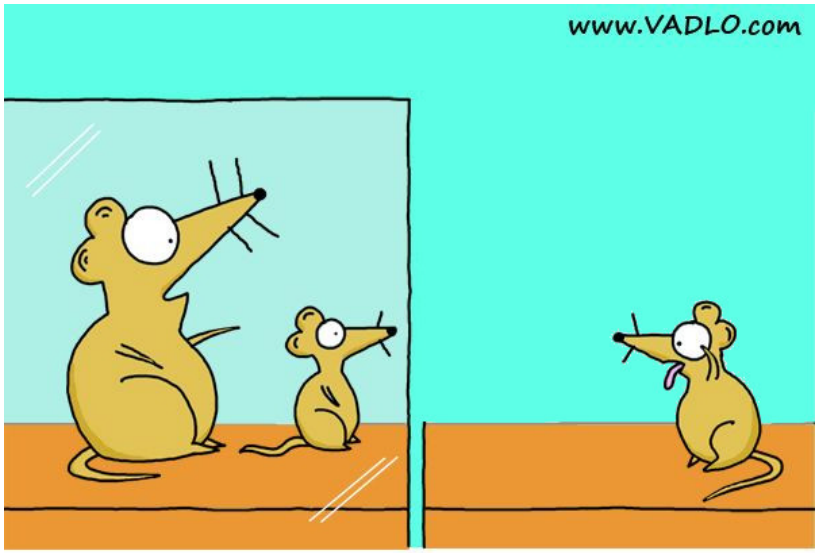
Posts are easily searched by #FIR.

SCIENCE + PRACTICE

- Science
  1) *Concepts*
  2) *Models*
  3) *Experimental evaluation*

- Practice
  1) *Systems*
  2) *Programming*
  3) *Experimental evaluation*

- Lab rats vs. wild rats!



"Don't play with him, he is **Wild Type**."

SCIENCE + PRACTICE

- Traditionally, scientists use their own prototype search engines:
  - Smart
  - Okapi
  - Terrier
  - Lemur/Indri
- Practitioners use professional engines
  - Elasticsearch
  - Solr
  - Lucene
- But … things start to converge!

INFORMATION RETRIEVAL

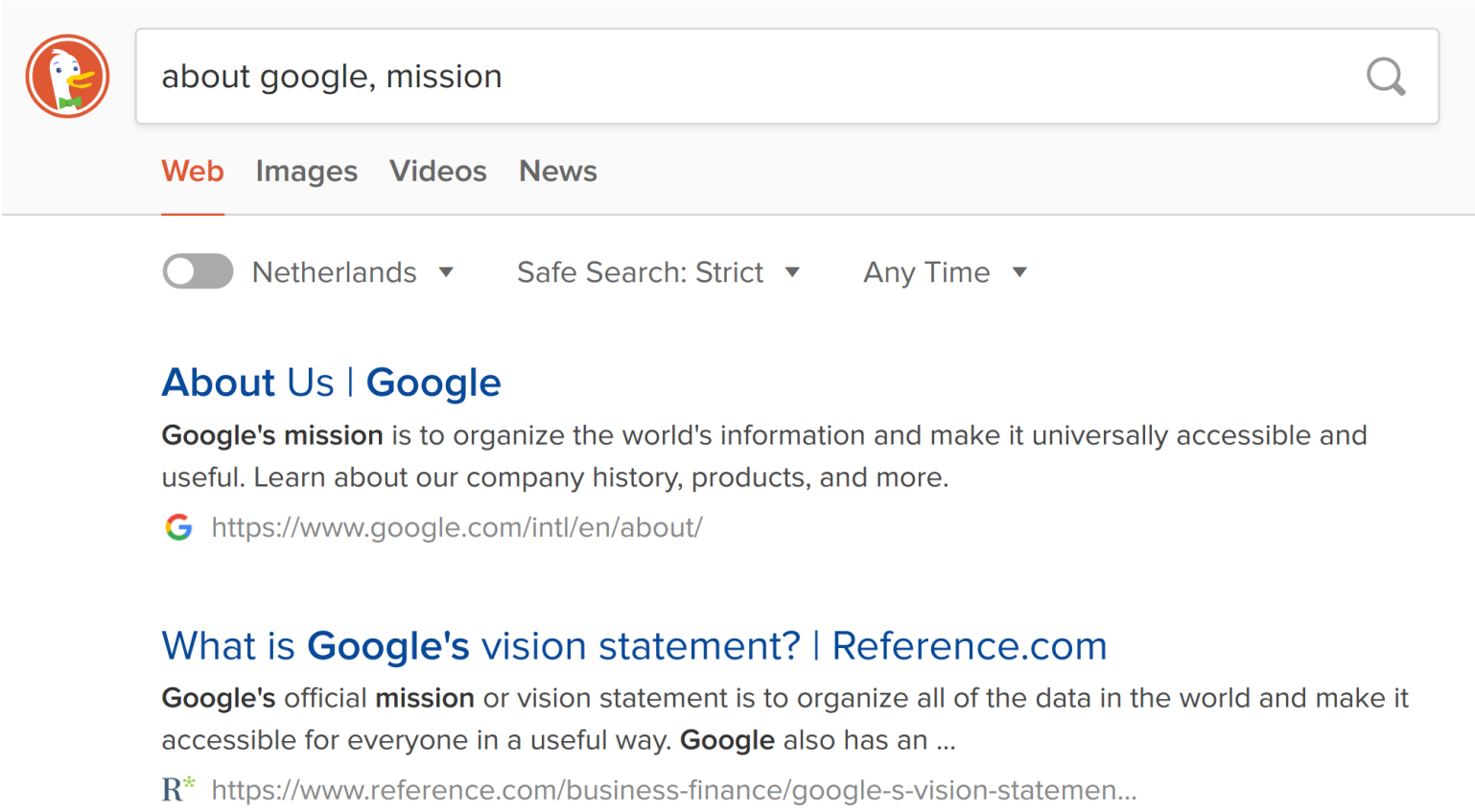
Information Retrieval (IR) is the scientific discipline that studies computer-based search tools.

- *How to distinguish a scientist from a practitioner?*

WHAT IS INFORMATION RETRIEVAL?

= web search !!!

WHAT IS INFORMATION RETRIEVAL?



MISSION

“Organize the world’s information and make it universally accessible and useful.”

What other organisations have this mission?

WHO ELSE?

- Libraries ?
- Scopus, Web of Science, … ?
- Twitter / Facebook ?
- Netflix ?
- Amazon ?
- iTunes / Spotify ?
- Medium ?
- U. Twente Search ?

- (Google books)
- (Goolge Scholar)
- (Google Plus)
- (Google’s YouTube)
- (Goole shopping)
- (Google Play Music)
- (Google Blogger)
- (Google Custom search)

A HISTORY OF “ORGANIZING THE WORLD’S INFO” (pre-history of IR)

- The Library of Alexandria
  - Built: 3rd century BC by Ptolemy I
  - Over 400,000 Papyrus scrolls
  - Visited by a.o. Euclid, Archimedes, …
  - Burned down as Romans conquested Greeks/Egypt

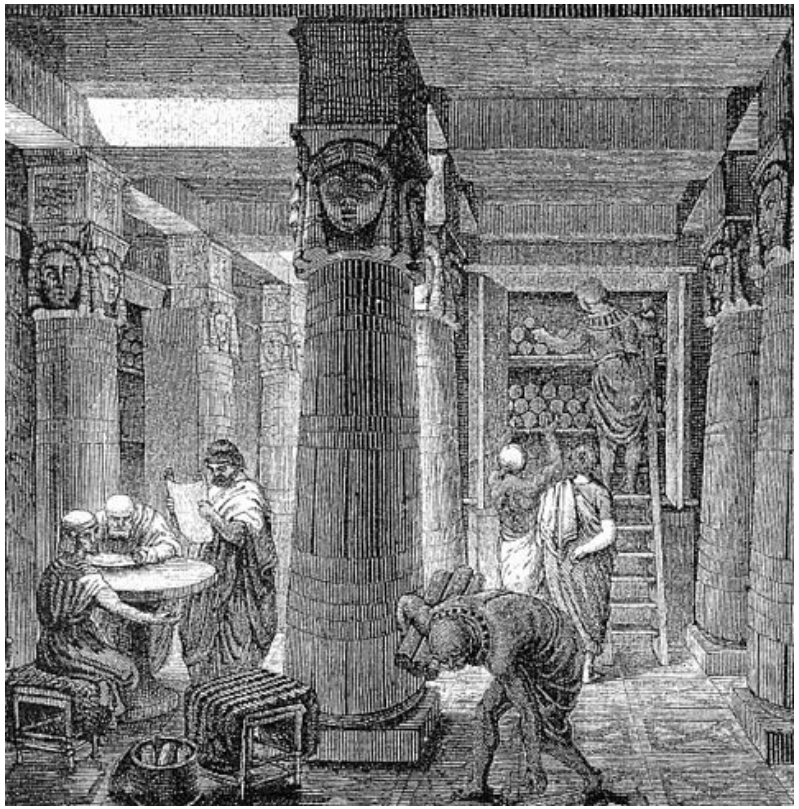


Image from Wikipedia

THE LIBRARY OF ALEXANDRIA

How did Archimedes find the right (relevant) scroll among 400,000 Papyrus scrolls ?

THE LIBRARY OF ALEXANDRIA

- **Callimachus**: poet, critic and scholar at the Library of Alexandria
- Made the **Pinakes**: considered to be the first library catalog.
- It divided works in:
  - genres & categories:
    rhetoric, law, epic, tragedy, comedy, lyric poetry, history, medicine, mathematics, natural science, miscellanies, ...
  - each category was alphabetized by author.

Image: allpostersimages.com

PRE-HISTORY: STANDARDS

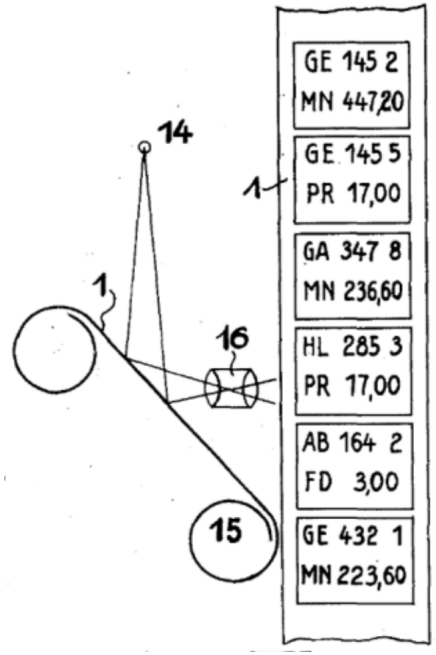
- Melvil Dewey's Decimal Classification (1876)

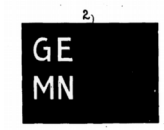
  Hierarchical numbering scheme made up of ten classes, each divided into ten divisions, each having ten sections.
  Decimals create further divisions:

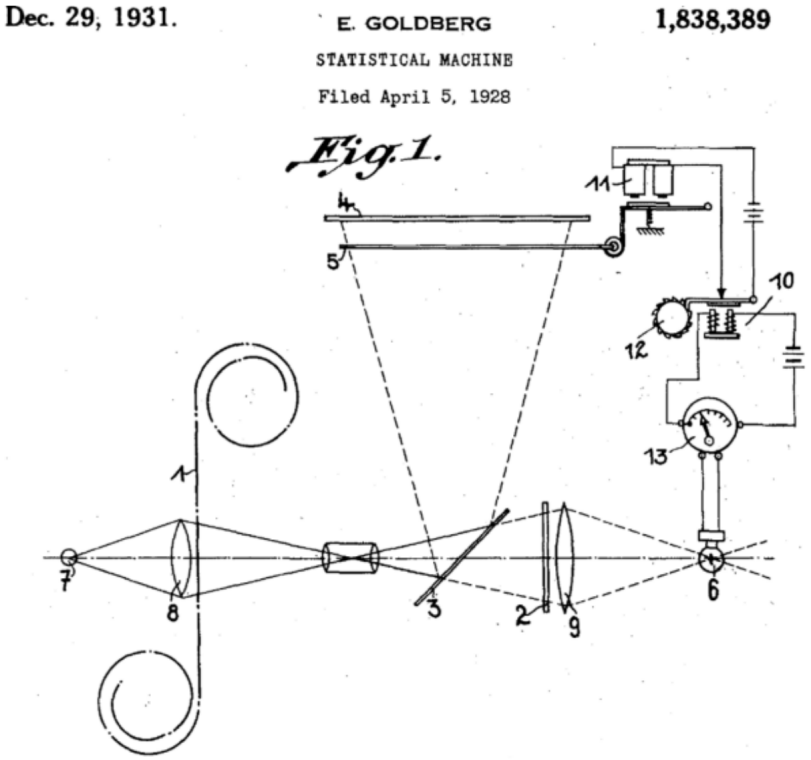
  500 Natural sciences and mathematics
  510 Mathematics
  516 Geometry
  516.3 Analytic geometries
  516.37 Metric differential geometries
  516.375 Finsler Geometry

PRE-HISTORY: FIRST MACHINES

- Emanuel Goldberg's Microfilm Search "Statistical Machine" (patent 1931)







PRE-HISTORY: FIRST MACHINES

- Emanuel Goldberg's Microfilm Search
  "Statistical Machine" (patent 1931)

  "Here it can be seen that catalogue entries were stored on a roll of film (No. 1 of the figure). A query (2) was also on film showing a negative image of the part of the catalogue being searched for; in this case the 1 st and 6 th entries on the roll. A light source (7) was shone through the catalogue roll and query film, focused onto a photocell (6). If an exact match was found, all light was blocked to the cell causing a relay to move a counter forward (12) and for an image of the match to be shown via a half silvered mirror (3), reflecting the match onto a screen or photographic plate (4 & 5)."

HISTORY: FIRST MACHINES

- Calvin Mooers coined the name “Information Retrieval” (1950)

“The problem under discussion here is machine searching and retrieval of information from storage according to a specification by subject... It should not be necessary to dwell upon the importance of information retrieval before a scientific group such as this for all of us have known frustration from the operation of our libraries – all libraries, without exception.”

HISTORY: STANDARDS

- Mortimer Taube (1952)
  “Unit terms”: a proposal to index items by a list of keywords.

1910 - 1965

HISTORY: EVALUATION

- Cyril Cleverdon (1960s)

- First empirical evaluation of information retrieval systems
- Measures: Precision & Recall
- Showed that using all keywords from abstract outperform manual indexing (!)

HISTORY: RANKING

- Many researchers argued that *ranking* is essential

Hans Peter Luhn (1957)
Similarity based in term frequencies (tf)

Karen Sparck-Jones (1972)
Specificity based on inverse document frequency (idf)

Gerard Salton (1975)
based on tf x idf

Keith van Rijsbergen (1975)
Information Retrieval: first popular scholarly book

HISTORY: TEXT RETRIEVAL CONFERENCE (TREC)

- Development of standard reusable test collections based on Cleverdon's work (1992)
- Organized by Donna Harman and later Ellen Voorhees

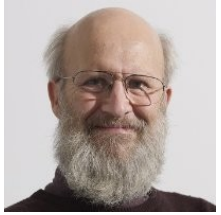
HISTORY: EFFICIENCY & COMPRESSION

- Ian Witten, Alistair Moffat, and Timothy Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, 1994



HISTORY: RANKING & MODELS

- Modern ranking models



Stephen Robertson (1994)
BM25 (with Steve Walker)

Bruce Croft (1998)
Language Models (with Jay Ponte)
(independently discovered by Djoerd Hiemstra and Miller, Leek & Schwartz)

Larry Page (1998)
Google PageRank (with Sergey Brin)

HISTORY: RANKING & MODELS

- Recent developments

  Machine Learning for IR:
  - “learning to rank”
  - “(deep) neural IR”

  Question answering
  - “conversational search”

FURTHER READING

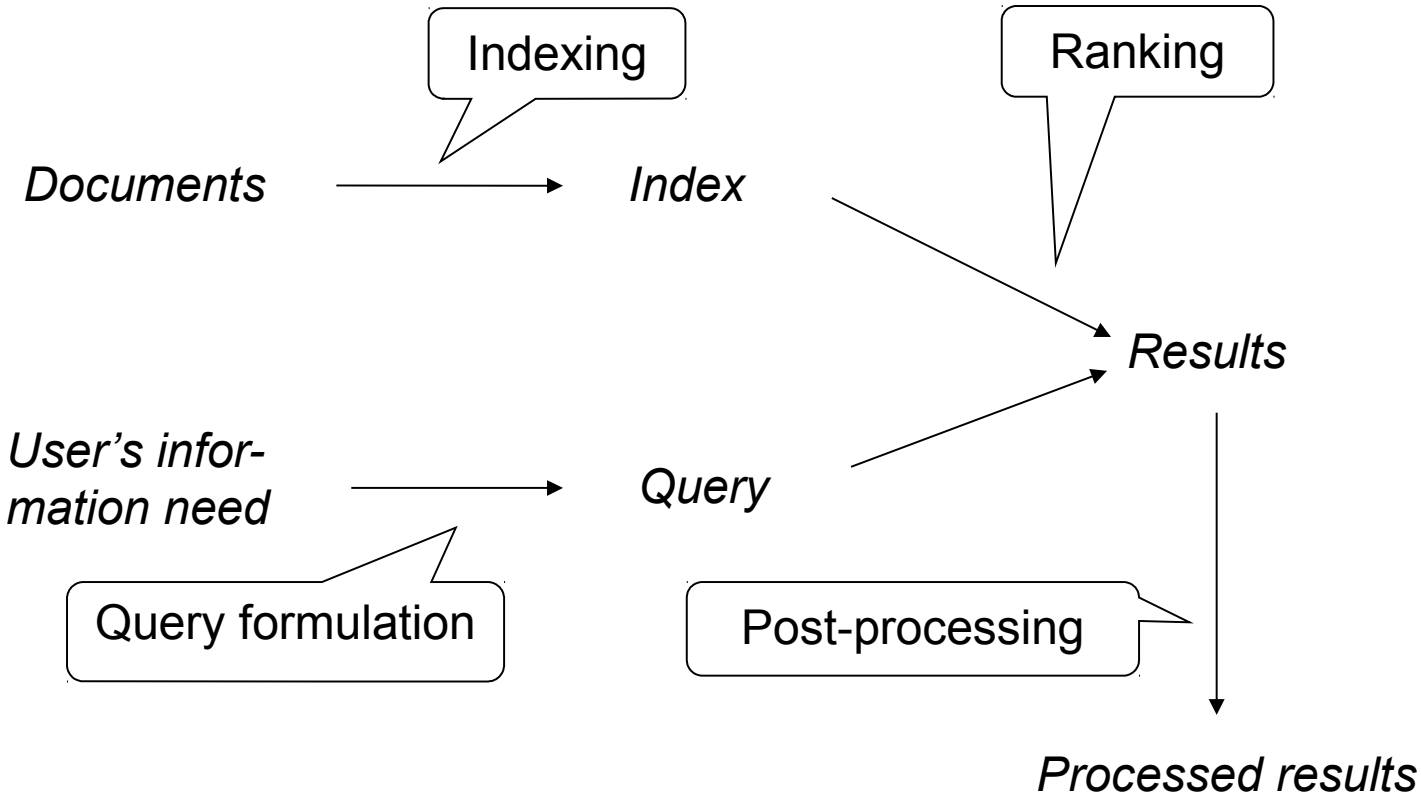
Mark Sanderson and Bruce Croft,
The History of Information Retrieval Research,
Proceedings of the IEEE, Volume 100, 2012
http://marksanderson.org/publications/my_papers/IEEE2012.pdf

WHAT IS INFORMATION RETRIEVAL?

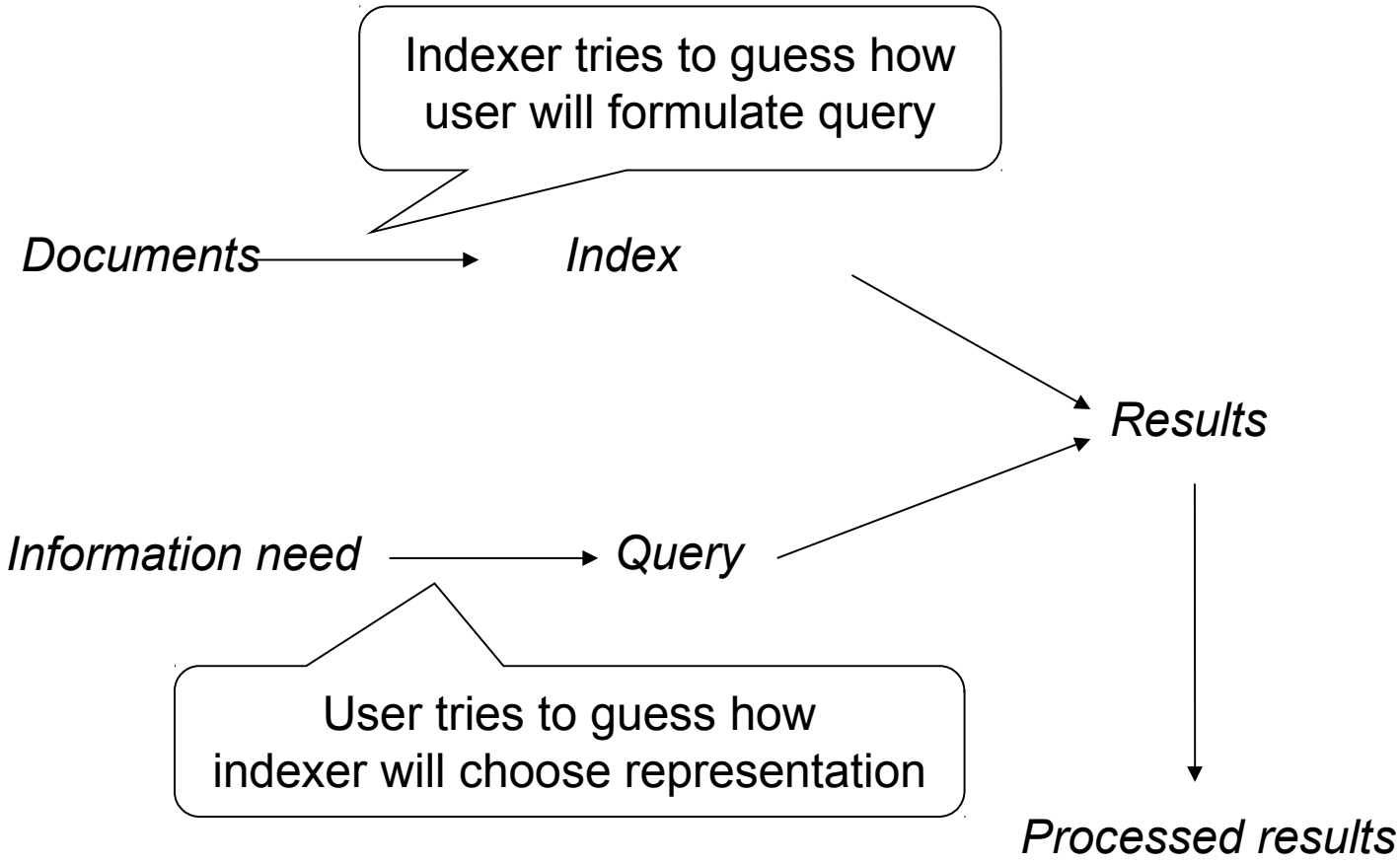
General characteristics:

- Users with an information need
- Documents
  - provide information, and
    (units part of bigger sources: sections, videos, scenes)
- A connection between the two

GRAPHICAL REPRESENTATION OF IR



THE PREDICTION GAME



ANOTHER VIEW

- Information retrieval is search for *similarity*:
  - between a document and a query
  - between documents in a collection (clustering)
  - between users (collaborative filtering)

VARIANTS

- Pull: ad-hoc requests, like WWW-searches
  - collection static, query dynamic
- Push: filtering, like personalised news service or spam filter
  - collection dynamic, query static

MORE THAN TEXT

- Texts
  - journal articles, press releases, WWW pages, …
- Pictures
- Audio
  - music, speeches, sounds for medical or engineering purposes, …
- Video
- Any combination

For example: Image Retrieval Systems

IR RESEARCH

Research in IR is concerned with the design of better IR systems

Theory

↕ ↓ top-down approach

Systems

↕ ↑ bottom-up approach

Experiments

Overview

- What is information retrieval?
- Approaches
- Performance
- Sources
- Course overview

Approaches: indexing

Traditionally, two styles:

- *Manually* by trained indexers, taking terms from pre-defined list (thesaurus)
- *Automatically* by deriving *features* like
  - words, word stems, phrases from texts
  - graphical features (colour distribution, texture etc.) from images
  - how about sounds, how about videos, how about smells?

Approaches: query formulation

- Traditionally by hand
- Formulating a good query is difficult!
- Increasing attention to automated aids for query formulation
  - natural-language queries
  - relevance feedback
  - personalisation
  - recommender systems

Approaches: query formulation

Other dimensions:

- Query in Italian, answer in Dutch
- Query by example: natural-language fragment, part of a picture
- Spoken query
- More expressive query languages (e.g., a description logic)
- Conversational systems

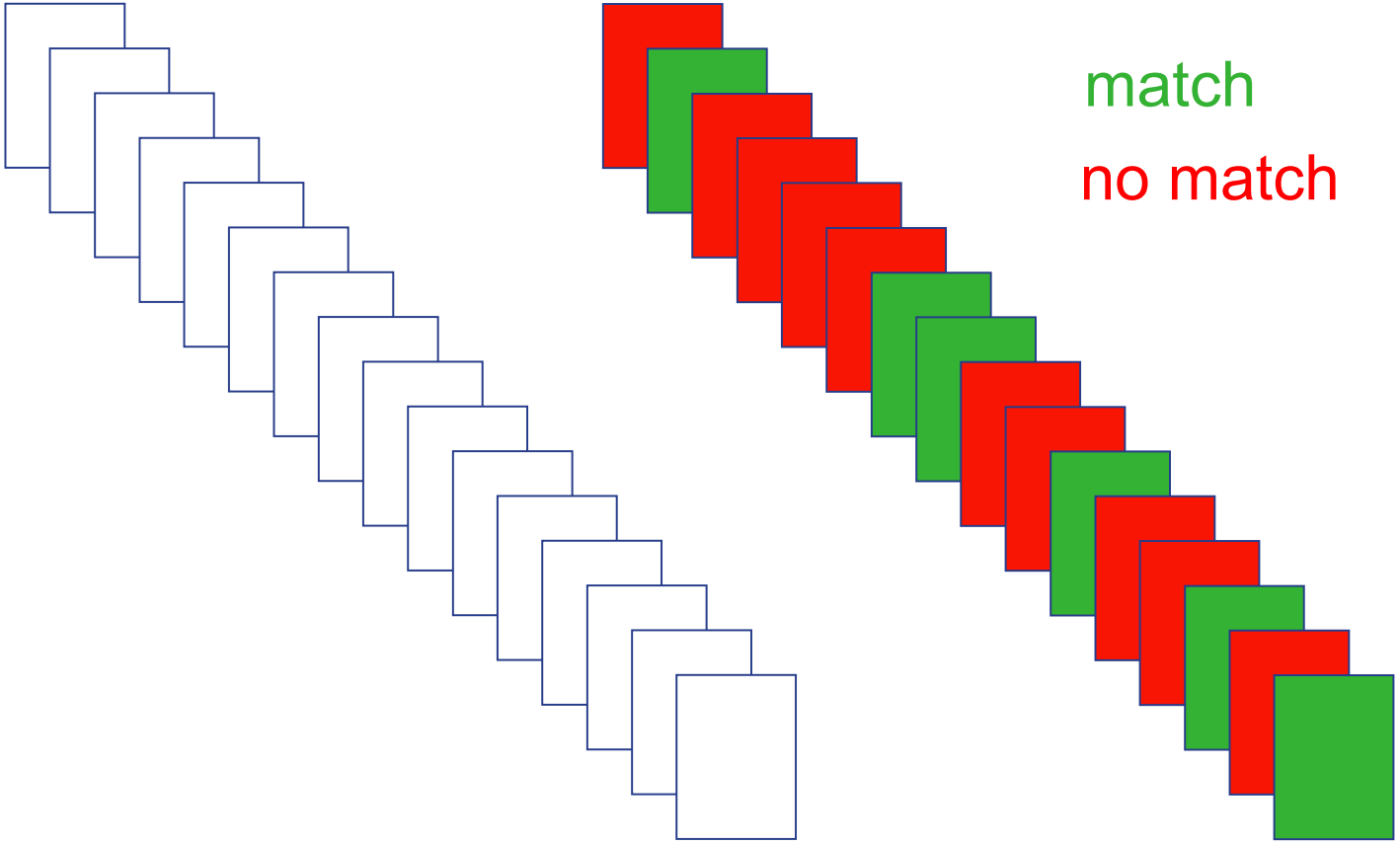
Approaches: ordering engine

Two basic approaches:

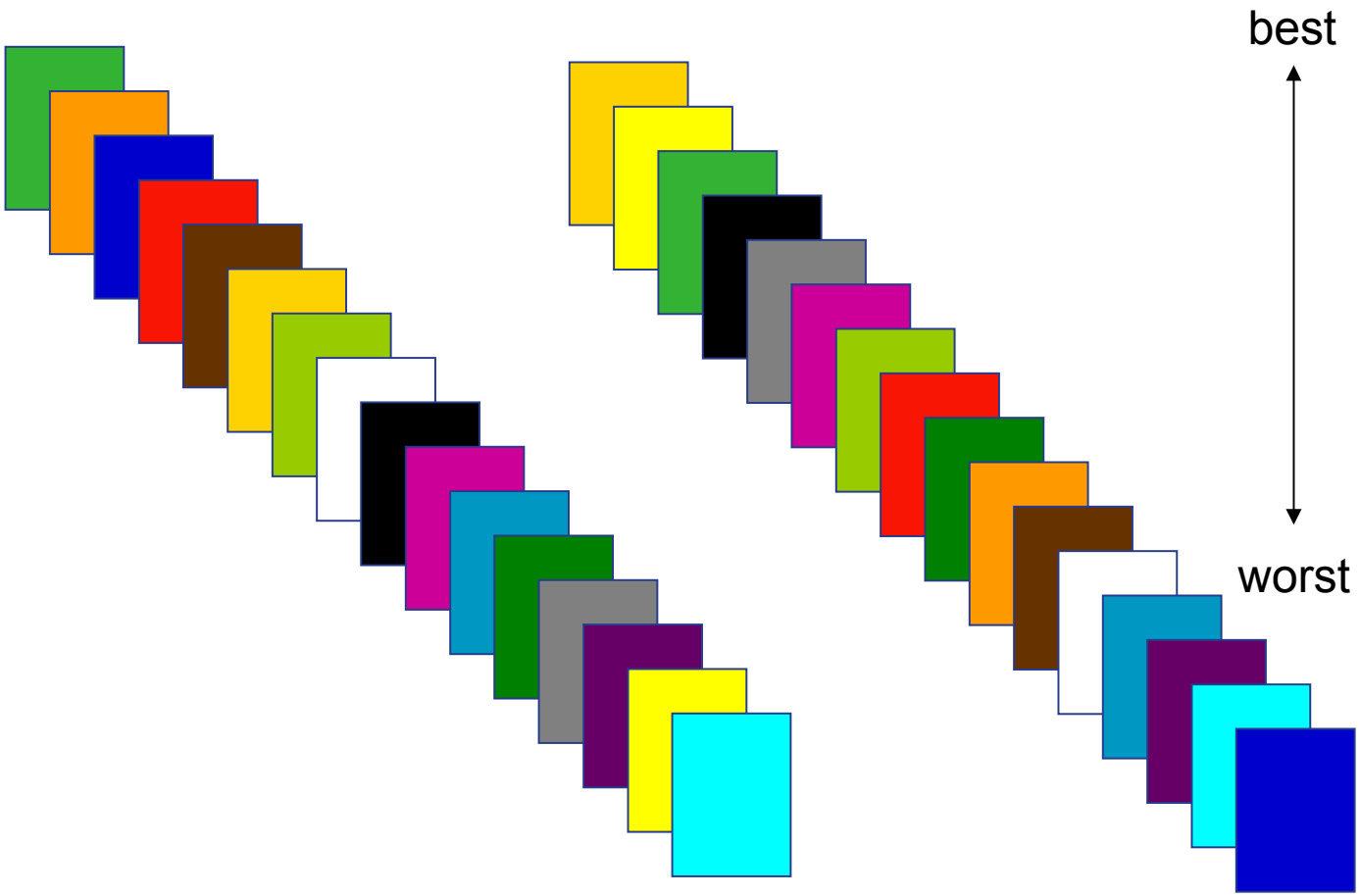
- *Matching* imposes a dichotomy on the collection
- *Ranking* rank-orders the entire collection

- N.B. The set $\{A, B\}$ is a dichotomy of set C iff $A \cap B = \emptyset$ and $A \cup B = C$

Matching



Ranking



Approaches: presentation

- The item as it is found in the collection
- Part of the document: a section, a paragraph, audio fragment
- A summary
- An answer to the question you posed (question-answering systems)

Overview

- What is information retrieval?
- Why information retrieval?
- Approaches
- Performance
- Sources
- Course overview

Performance

- Important decision: which system is better?
- Has large economic impact
- Compare Google's market value
- A good IR system can make the difference between winning or losing e.g.
  - a contract
  - a legal case

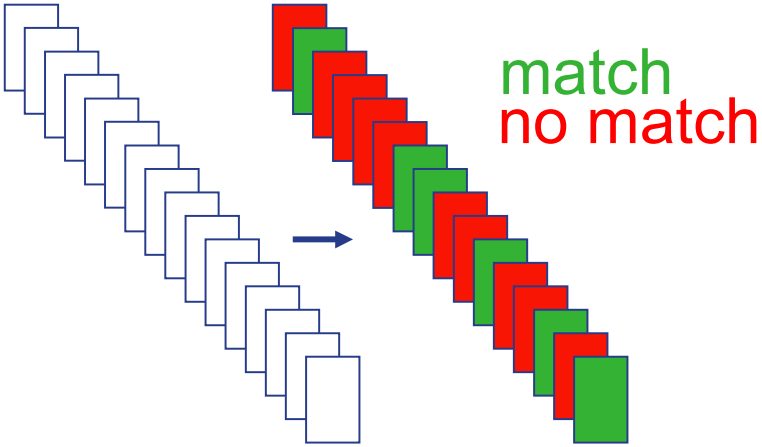
Measuring performance

Theory of measurement in IR is difficult, for example:

- Which queries are a representative sample of the population of all queries?
- Does a good measurement mean that the user is satisfied?
- What about queries that can only be answered by *combinations* of items?

Performance: matching as example

- *Match / no match* is a system decision
- *Relevant / not relevant* is a user decision
- Gives rise to familiar quadrant (compare medical tests)



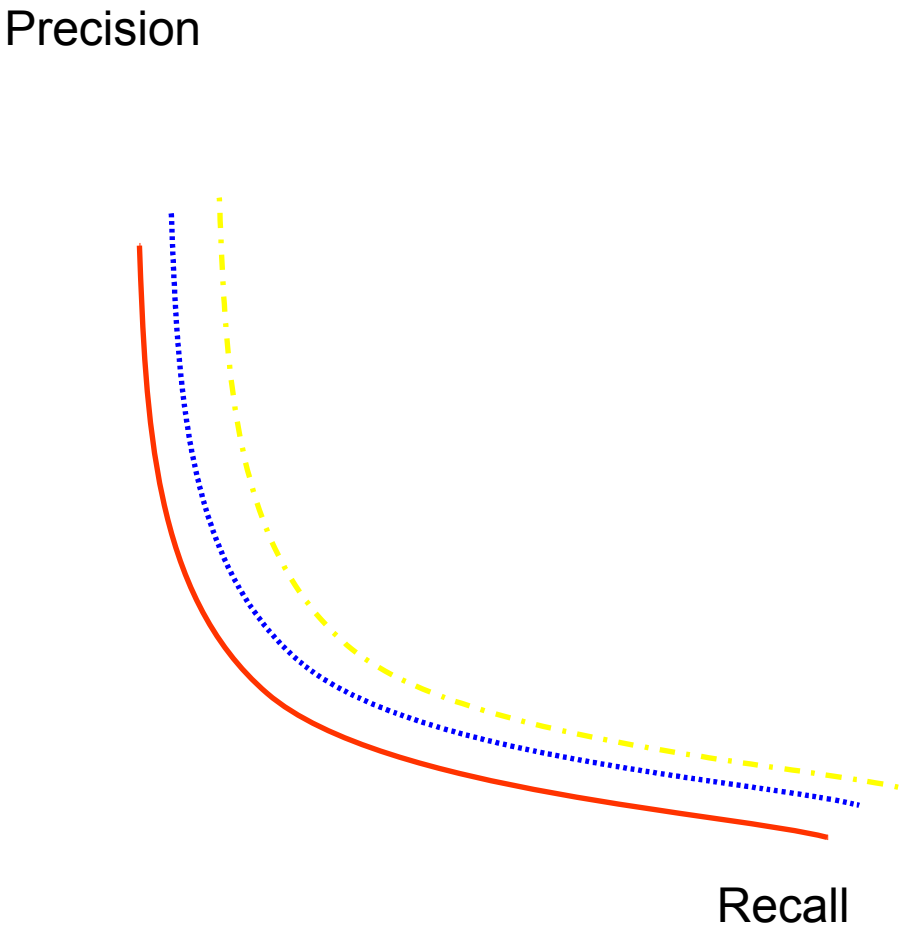
Performance for matching

	System says: Match	System says: No match
User says: Relevant	True positives *(#TP)*	False negatives *(#FN)*
User says: Not relevant	False positives *(#FP)*	True negatives *(#TN)*

$$Recall = \frac{\#TP}{\#TP + \#FN} \qquad Precision = \frac{\#TP}{\#TP + \#FP}$$

Performance for matching

- “Fact of life”: improving recall typically decreases precision.



Measuring performance: TREC

- Yearly competition, held in November
- Idea: demonstrate your system on unknown queries for a known, very large collection
- System with the best recall-precision performance “wins”
- Pro:
  - State of the art known
  - Competition incentive for improvement
  - Forum for exchange of ideas
- Con:
  - Test environment sets constraints on what can be done and what not

Overview

- What is information retrieval?
- Why information retrieval?
- Approaches
- Performance
- Sources
- Course overview

Sources: journals

- Information Retrieval
- Journal of the American Society for Information Science and Technology
- Information Processing & Management
- ACM Transactions on Information Systems

Sources: conferences

- ACM SIGIR conference
- WSDM: Web Search and Data Mining
- ICTIR: Int. Conf. On Theory of Information Retrieval (Amsterdam!)
- CHIR: Conf. Human Interaction IR
- ACM International Conference on Digital Libraries
- ACM Conference on Information, Knowledge and Management
- Text REtrieval Conference (not peer-reviewed but a kind of contest)
- WWW: World Wide Web Conference

The University of Twente provides access to all mentioned journals and conference proceedings.

END OF THE INTRODUCTION

Next:

- Elasticsearch and real data (TREC genomics)
- Preparing the docker image
- Introduction to Elasticsearch