Web Retrieval
Evaluation

Frank Hopfgartner
Institute for Web Science and Technologies

Intended Learning Outcomes

universität koblenz
weiter:denken

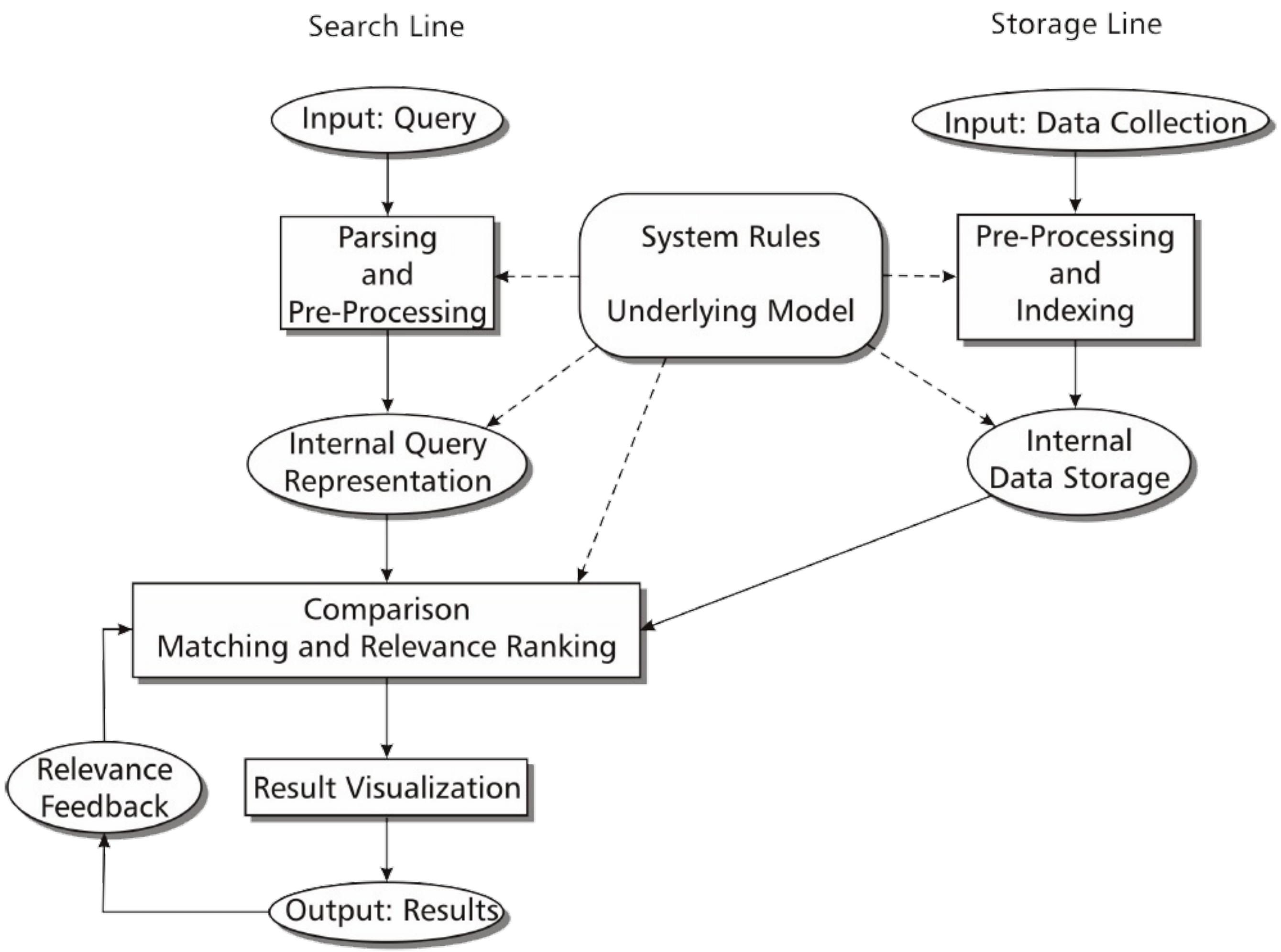
- At the end of this lecture, you are expected to
  - understand how to evaluate an IR system
  - understand the difference between evaluation measures that ignore the ranking and those that consider the ranking

Outline

- IR System Architecture
- Motivation (Why should we evaluate?)
- What should we evaluate?
- How should the evaluation be conducted?
- Evalution Metrics
- Further Evaluation Approaches

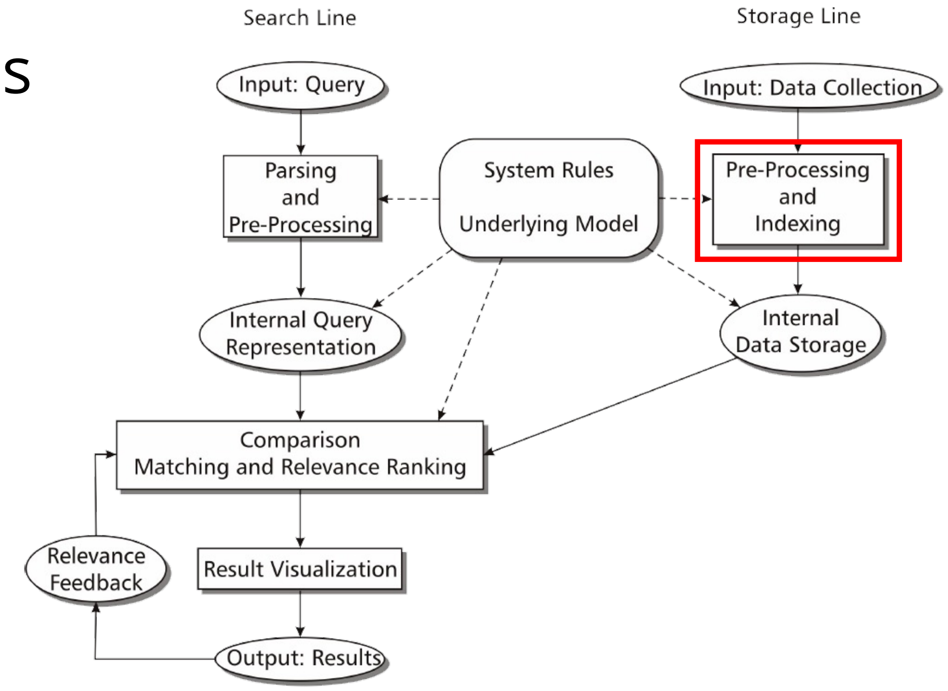
IR System Architecture

IR System Architecture



Preprocessing and indexing

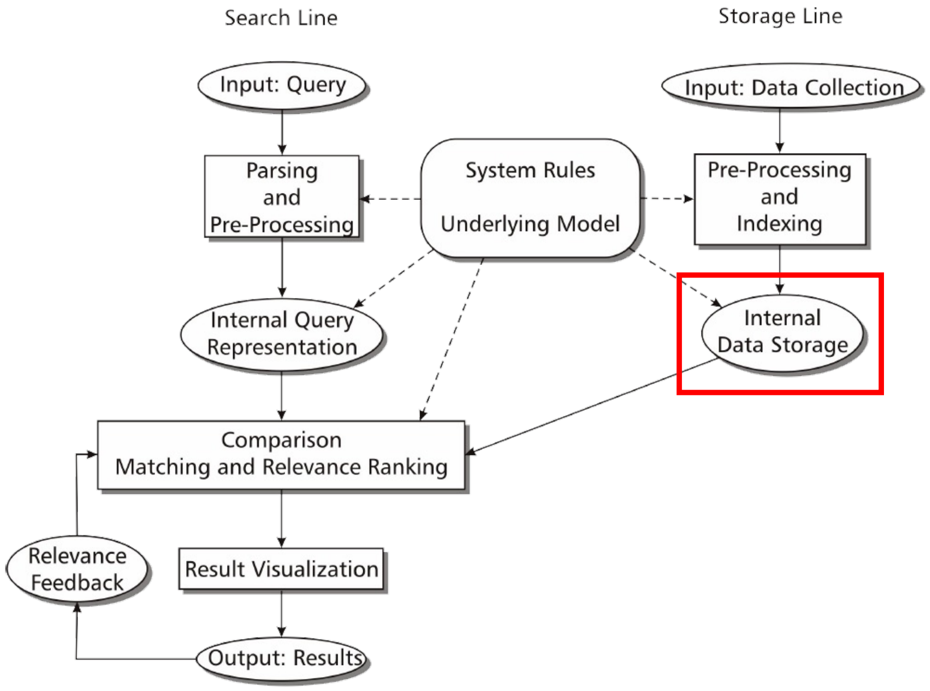
- Transforms raw data into an internal format
- For documents:
  - Interpretation of character sequences
  - Recognition of
    - Words and phrases
    - Sentence structure
    - Part-of-speech
  - Syntatical analysis
  - Morphological analysis
  - statistical and linguistic methods



Internal data storage

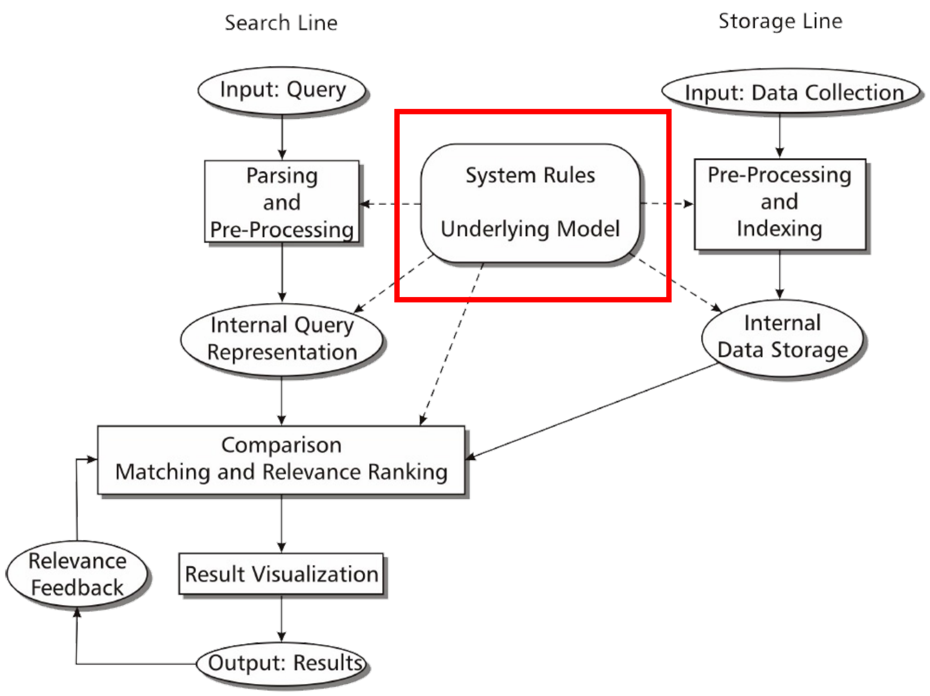
universität koblenz
weiter:denken

- Store the data so that
  - its content can be described accurately
  - efficient access is guaranteed
  - storage space is kept to a minimum
- Data storage types:
  - Inverted index
  - Suffix trees
  - …



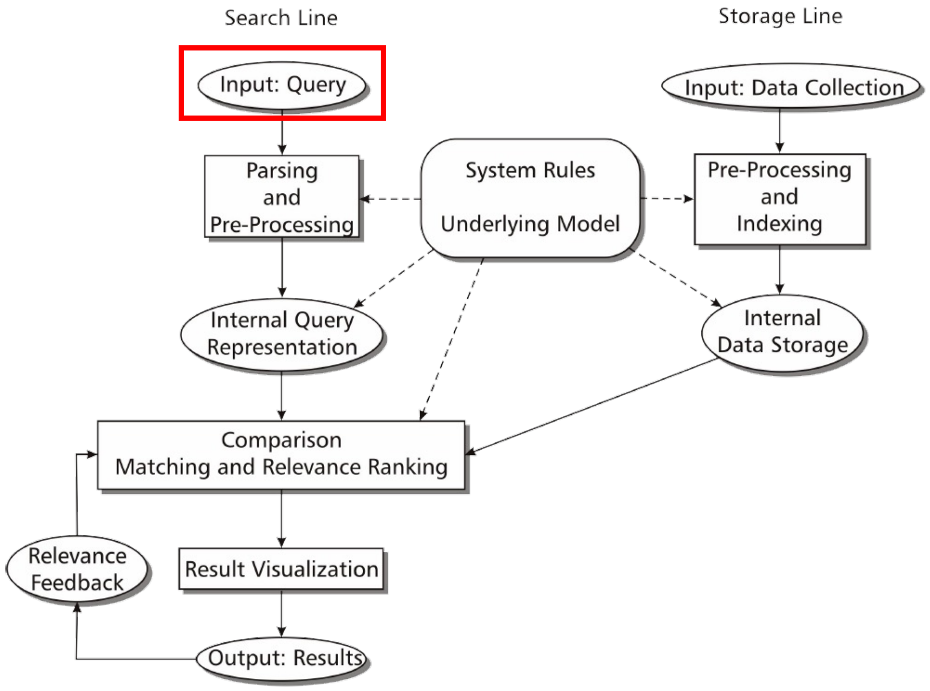
Underlying Model

- Fundamental component
- Framework for the representation of
  - queries
  - objects and
  - their relations
- Models
  - Boolean
  - Vector Space
  - Probabilistic, ...



Queries

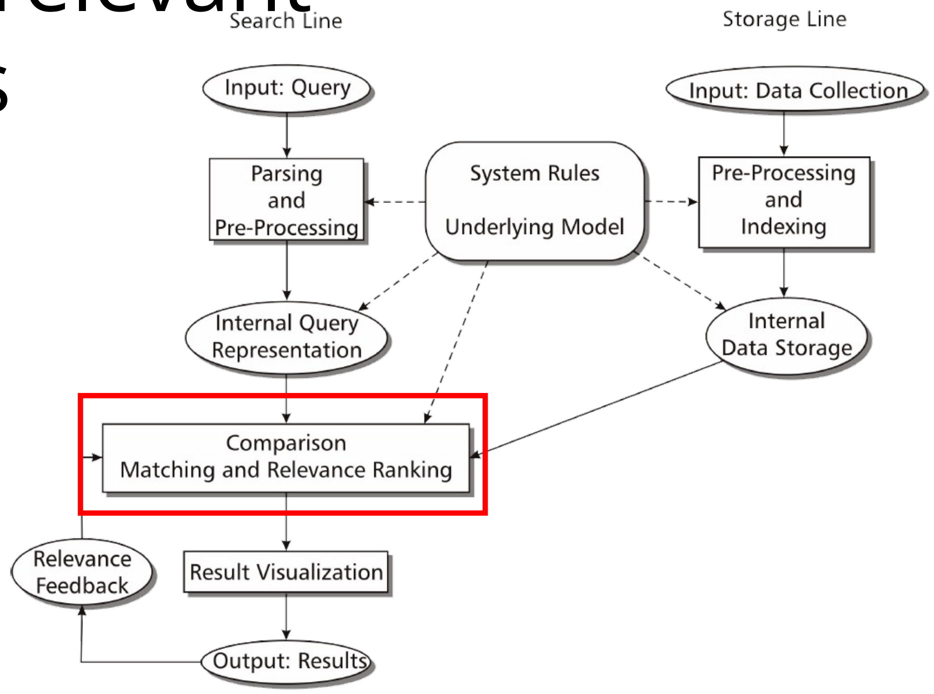
- Types:
  - Natural language
  - stylised natural language
  - Boolean
  - Form based (GUI)
- e.g., Boolean:
  - Terms
  - Operators
  - AND, OR, NOT



Matching and relevance ranking

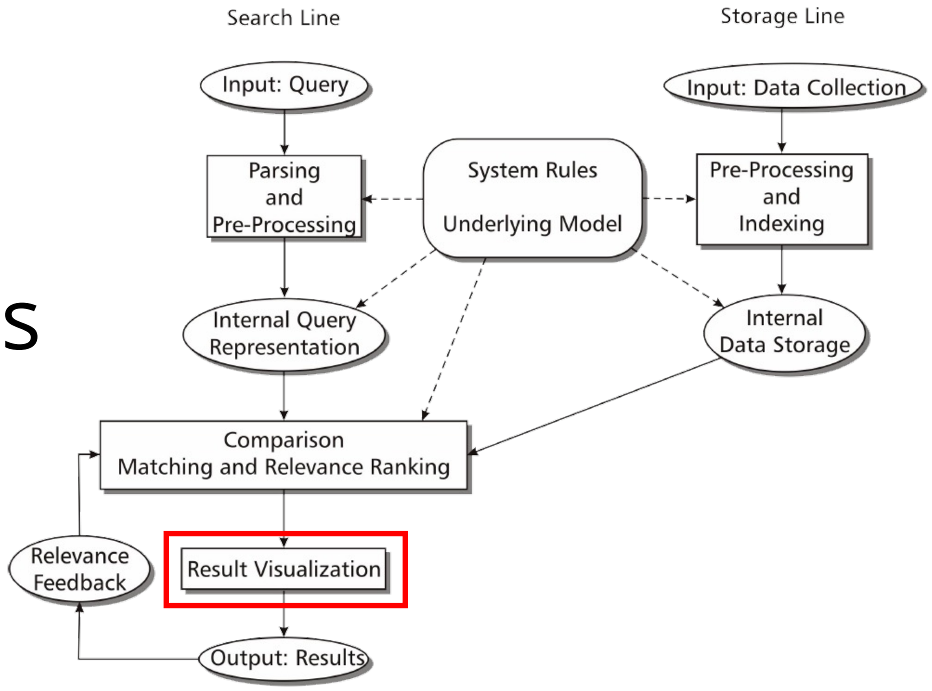
universität koblenz
weiter:denken

- Searches in internal data storage for documents that match to query
- A relevance matrix separates relevant from non-relevant documents
- Sorting, e.g.,
  - chronological
  - based on appearance of search term
  - based on popularity



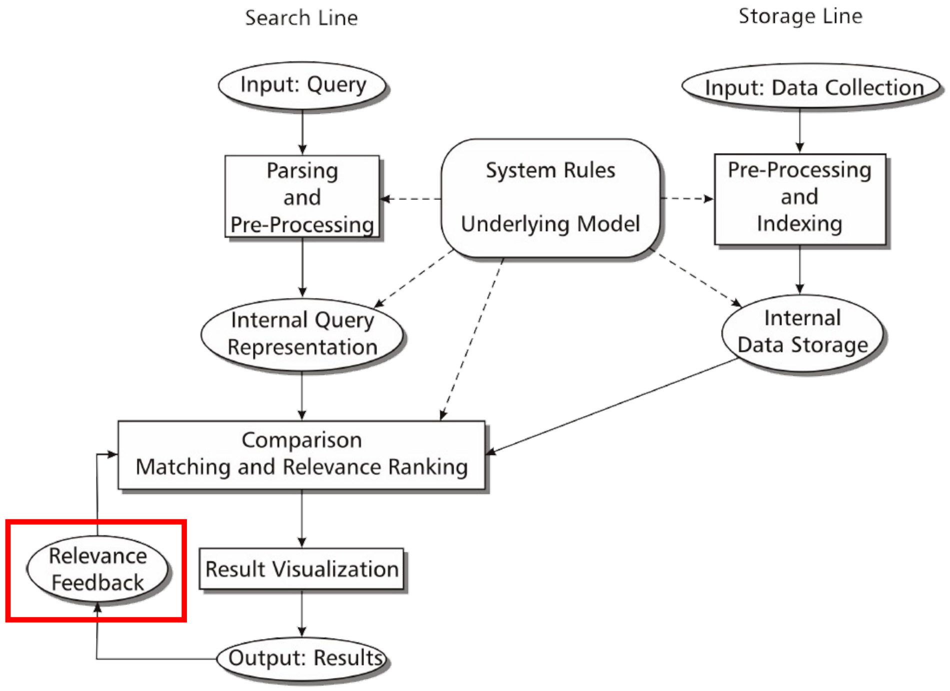
Interface and visualisation

- Interaction with user
- Accepts requests
- Visualises results
  - sorted lists
  - information per document
  - illustrates similarities
- Deals with interactions such as
  - Relevance feedback
  - Query refinement
  - Filtering



Relevance feedback

- Filter
  - Reduces the result set
  - Filter criteria are metadata
  - Date
  - Domain
  - File type
  - …



Motivation to evaluate IR systems

Let’s remind ourselves

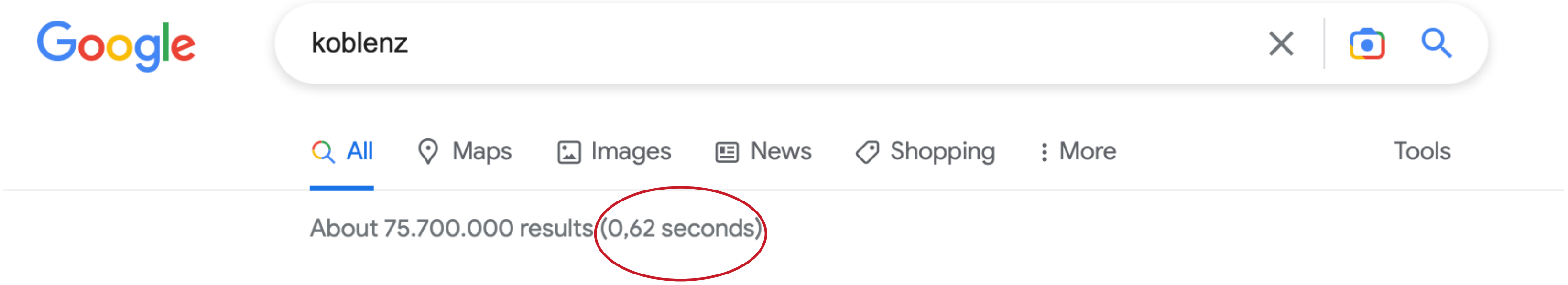
universität koblenz
weiter:denken

- The goal of an IR system is to satisfy users’ **information needs**
- An information need is an individual or group’s desire to locate and obtain information to **satisfy** a conscious or unconscious need
- Satisfaction is the **opinion of the user** about the IR system

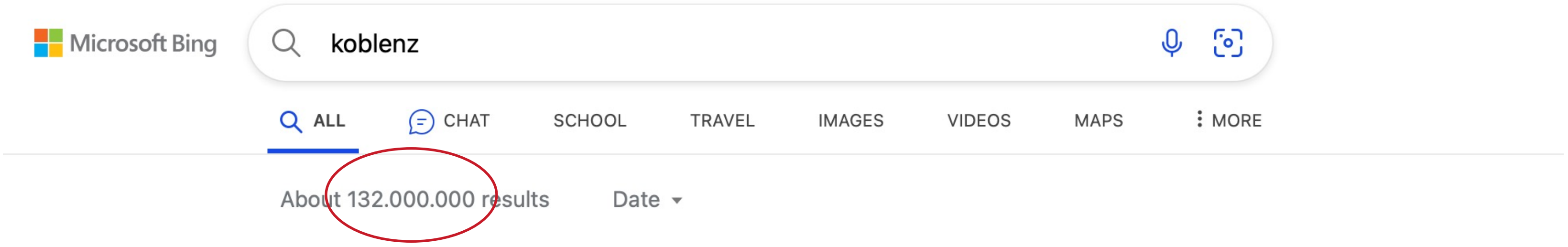
What influences your opinion of a search engine?

universität koblenz
weiter:denken

How fast it responses to your query?



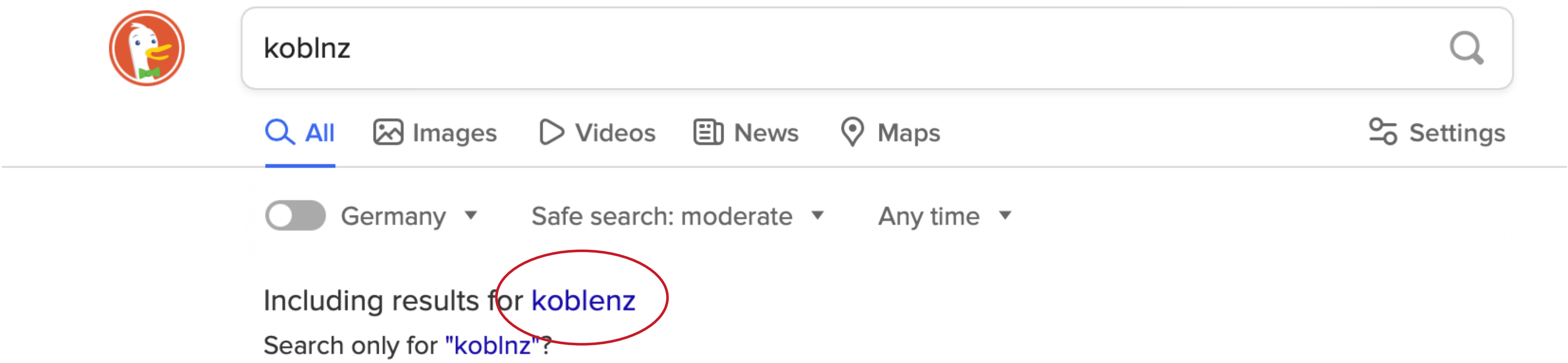
How many documents it can return?



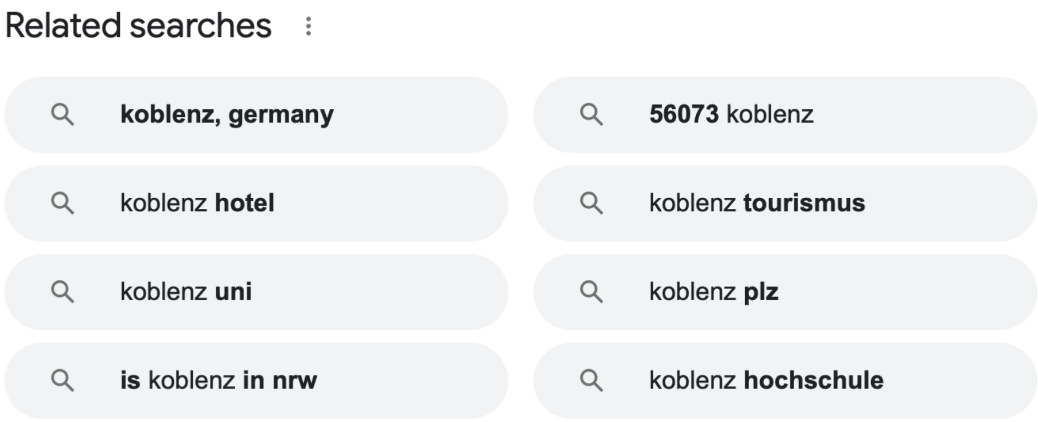
What influences your opinion of a search engine?

universität koblenz
weiter:denken

Can it correct spelling mistakes?



Can it suggest related terms?



What influences your opinion of a search engine?

universität koblenz
weiter:denken

- How well it supports user interaction
- Whether the user is satisfied with the results
- How easily users can use the system
- Whether the system helps users carry out tasks
- Whether the system impacts on the wider environment
- …

These all point to different aspects of IR systems.
We have to perform an evaluation to find the best one.

IR Evaluation

universität koblenz
weiter:denken

- To **evaluate** means to *"ascertain the value or amount of something or to appraise it"*

- *"IR evaluation is the systematic determination of merit of something using criteria against a set of standards"* (Harman, 2011)
  - Require a *systematic approach* for conducting evaluation
  - Need to identify suitable *criteria* for evaluating search
  - Need to *compare* against some *standard* (i.e. *comparative evaluation*)

- Measuring performance of search systems essential
  - Benchmark current performance
  - Quantify impact of changes
  - ...

What should we evaluate?

Let's look at another definition of IR

"Information Retrieval (IR) deals with the representation, storage, organization of, and access to information items."

This suggests that there are many aspects of an IR system that we could evaluate.

R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, Second Edition, 2011.

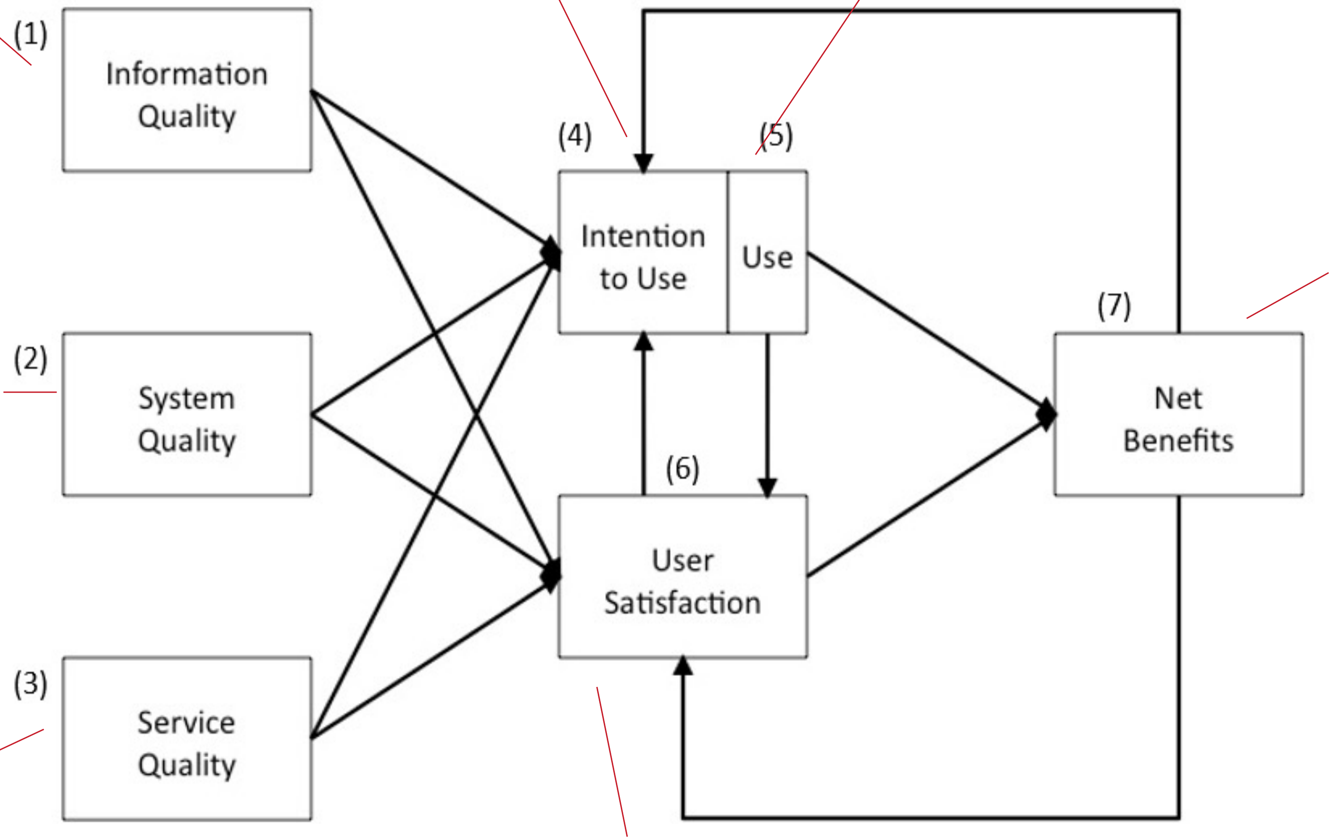
IR Evaluation from an Information Systems success perspective

Quality of data that system ingests or produces

Degree to which the system is intended to be used (i.e. perceived use)

Degree to which system and its components is actually used (e.g. usage analytics) and could be used to capture search process

Search quality, including: Coverage, Presentation of outputs, Response, Precision, Recall and User effort (Lancaster, 1966). Can assess the search process and outputs



Quality of support users receive

User's level of satisfaction when using the system (e.g. during task-based studies or for general use)

Outcomes of using the system and information found and impact on:

- Individuals
- Groups
- Organisation
- Externally

May include assessing:

- Enjoyment
- Task performance
- Usefulness
- Learning effects
- ……

W. H. Delone and E. R. McLean. Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, 3(1):60-95, 1992. https://doi.org/10.1287%2Fisre.3.1.60

Levels of IR evaluation

universität koblenz
weiter:denken

- Evaluation of retrieval systems tends to focus on either the system (algorithms) or the user
- Saracevic (1995) distinguishes six levels of evaluation for information systems that include IR systems

IR system evaluation

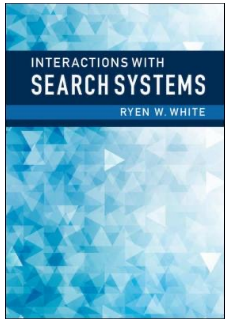
- Engineering
- Input
- Processing

Focus of IR evaluation mainly here (test collections and batch-mode lab-style evaluation)

Interactive IR (IIR) evaluation / Human information behaviour

- Output
- Use and user
- Social

Criteria: What to measure



In the 1960's, Cyril Cleverdon suggested the following

- Coverage
- Time lag
- Presentation
- Effort
- Recall
- Precision

In 2016, Ryan White suggested the following

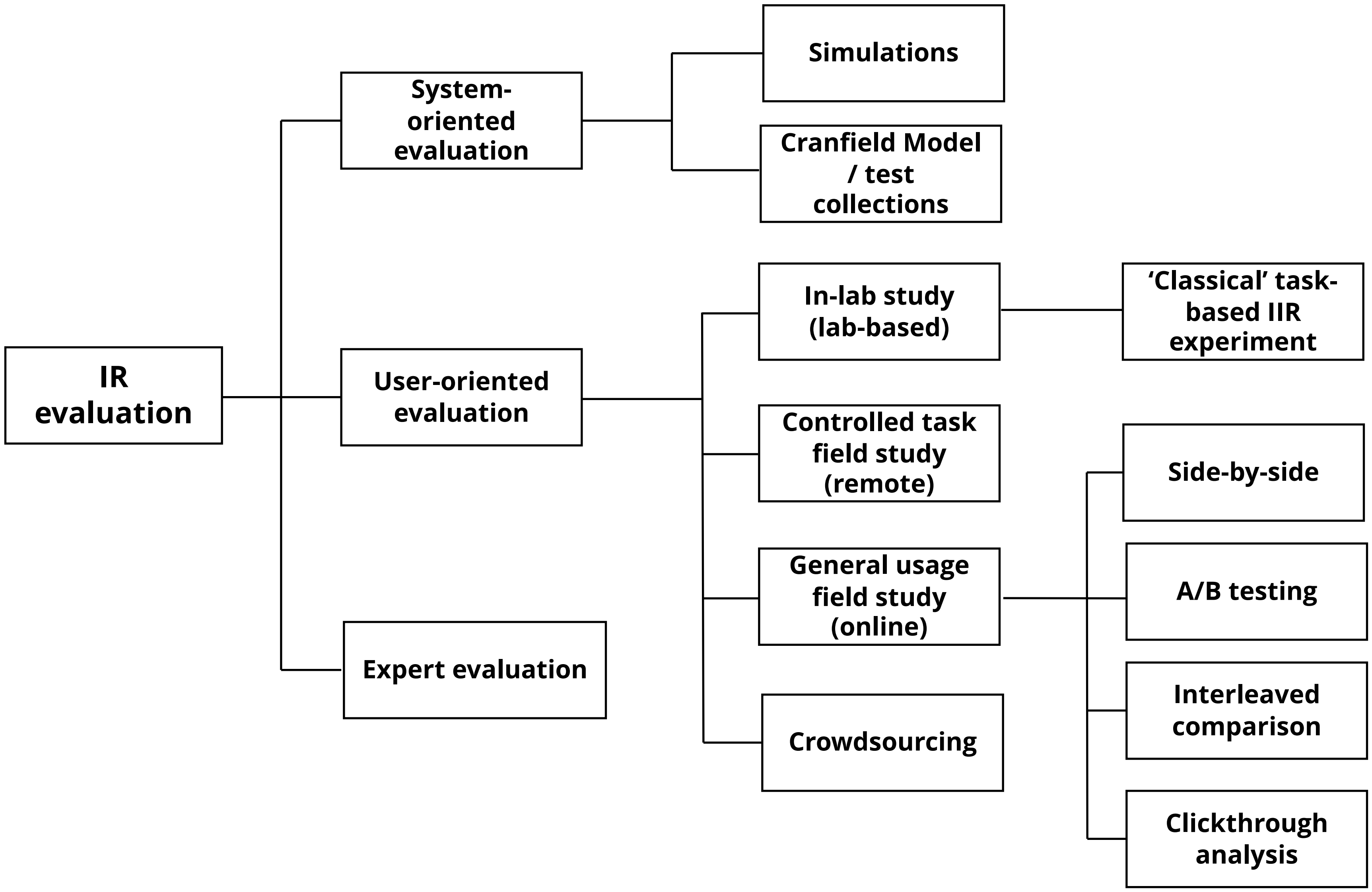
Search outcomes

- Relevance (precision, effort, etc.)
- Novelty / diversity
- Success
- Satisfaction
- Support for creativity
- Adoption and retention

Search process

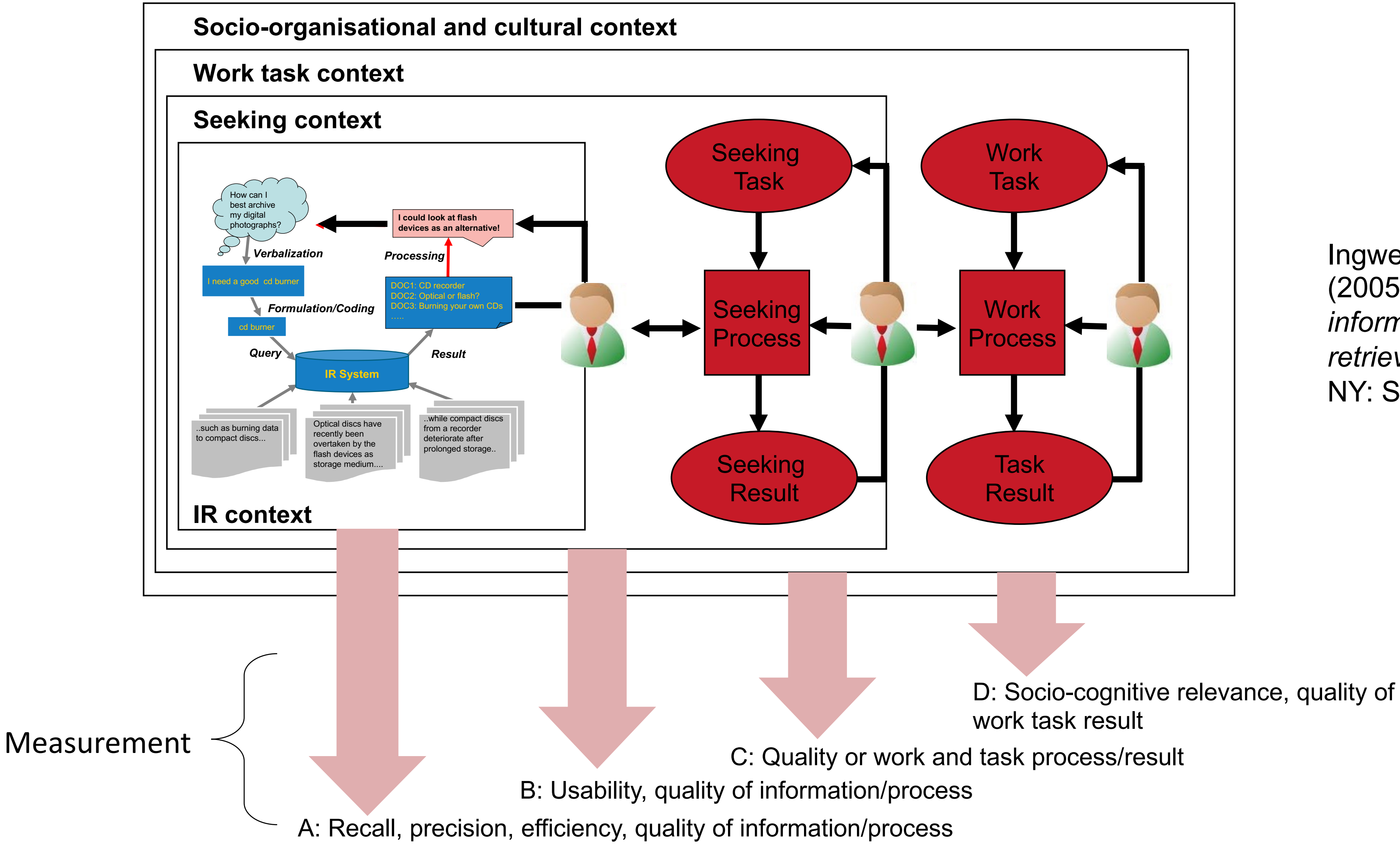
- Learning
- Efficiency
- Cognitive load
- Serendipity
- Enjoyment
- Frustration
- Engagement

Evaluation Methodologies



Context of IR evaluation

universität koblenz
weiter:denken



Ingwersen: & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*, New York, NY: Springer-Verlag

How should the evaluation be conducted?

Planning an evaluation

universität koblenz
weiter:denken

When preparing an evaluation key questions include (Saracevic, 2000)

- **Why** conduct the evaluation?
  (i.e., the goal/purpose of the evaluation)
- **What** should be evaluated?
  (i.e., the success criteria to be used)
- **How** should the evaluation be conducted? (i.e., the evaluation methodology)
- For **whom** to evaluate?
  (i.e., the stakeholder of the evaluation)

Relevance

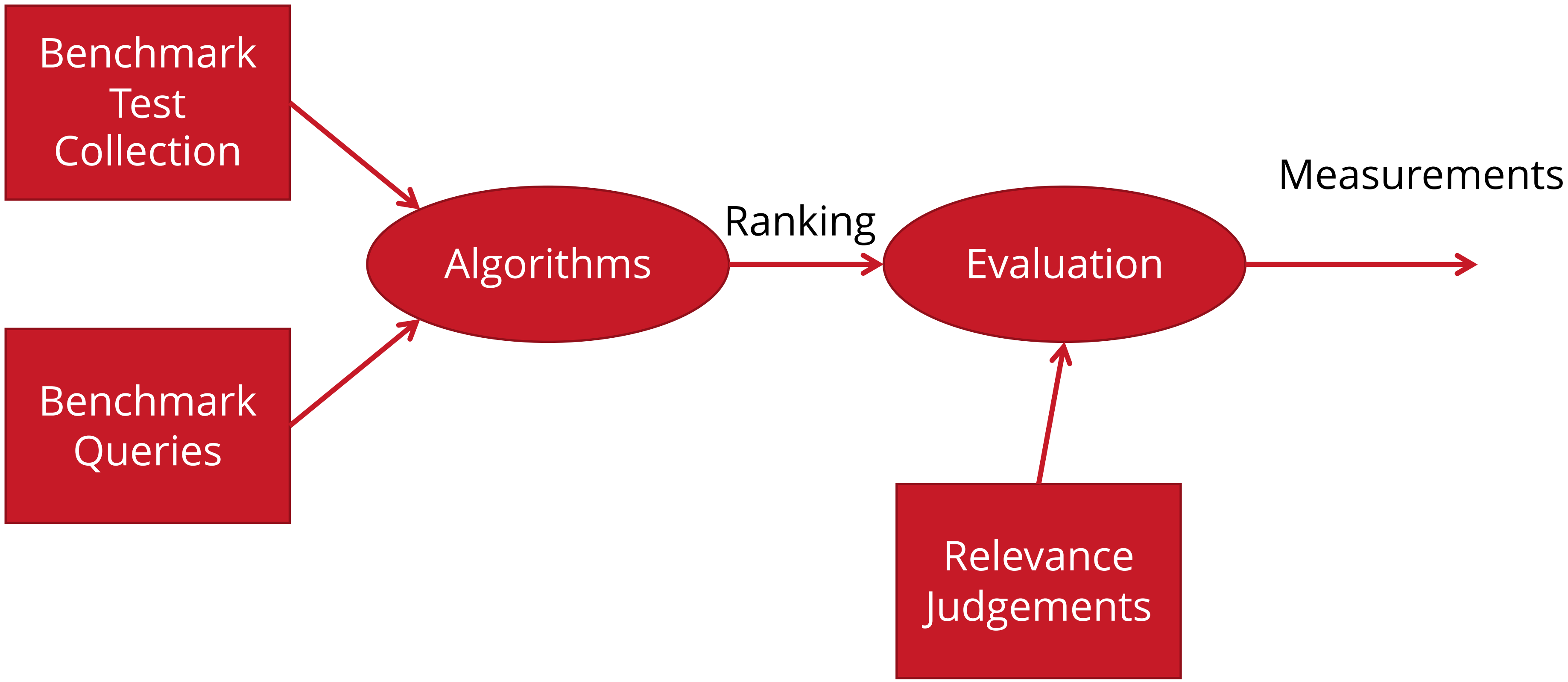
- The retrieved resource is relevant if it is appropriate to the information need (not a query). Otherwise, it is non-relevant
- Types
  - Actual relevance: hard to estimate
  - Subjective relevance/ Pertinence: Relevance to a particular user
  - Objective relevance: External assessor(s)
  - System relevance: determined by an IR system
    - RSV (Retrieval Status Value)

How to evaluate an IR system?

universität koblenz
weiter:denken

- Given a test collection consisted of
  - A collection of resources, e.g. documents
  - A set of informations needs
    - Topics that are expressible as queries
  - A set of relevance judgements
    - typically a binary assessment being of either relevant or nonrelevant
  - Assessors
- Evaluate retrieval effectiveness
  - One assessor per resource/information need
  - Binary assessment
  - No agreement among assessors is required

How to evaluate an IR system?



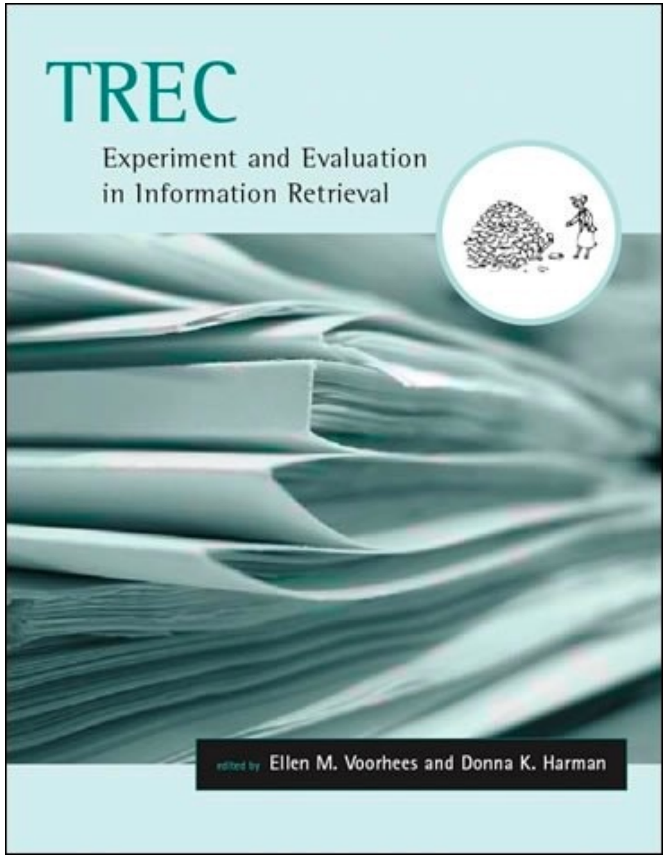
Collection

universität koblenz
weiter:denken

- The assessments are called gold standards or ground truth
- The outcome of the evaluation is highly variable for different resources and information needs.
  - The test collection should be of reasonable size

TREC Experiment and Evaluation in IR

"This book provides a comprehensive review of TREC research, summarizing the variety of TREC results, documenting the best practices in experimental information retrieval, and suggesting areas for further research.

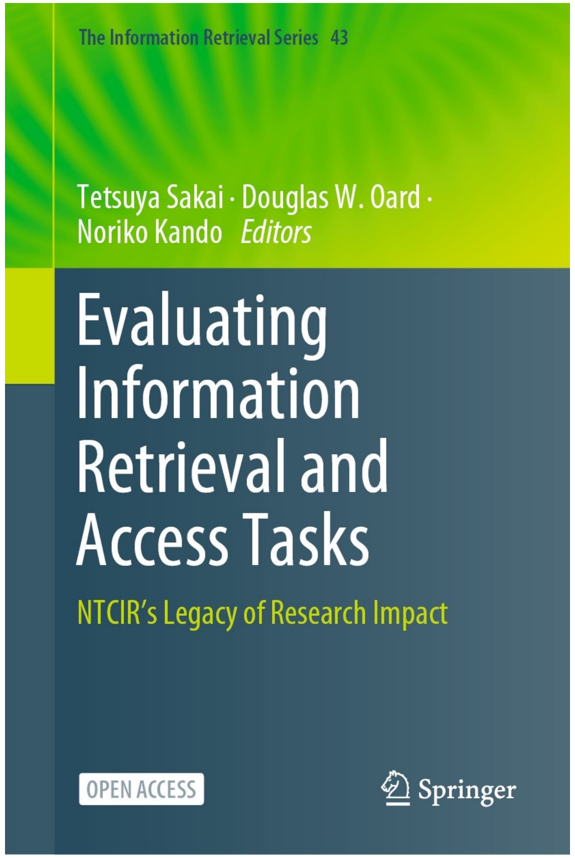


E.M. Voorhees and D. K. Harman. TREC Experiment and Evaluation in Information Retrieval. MIT Press, 2005.

NII Testbeds and Community …

universität koblenz
weiter:denken

> *This open access book summarizes the first two decades of the NII Testbeds and Community for Information access Research (NTCIR). NTCIR is a series of evaluation forums run by a global team of researchers and hosted by the National Institute of Informatics (NII), Japan.*

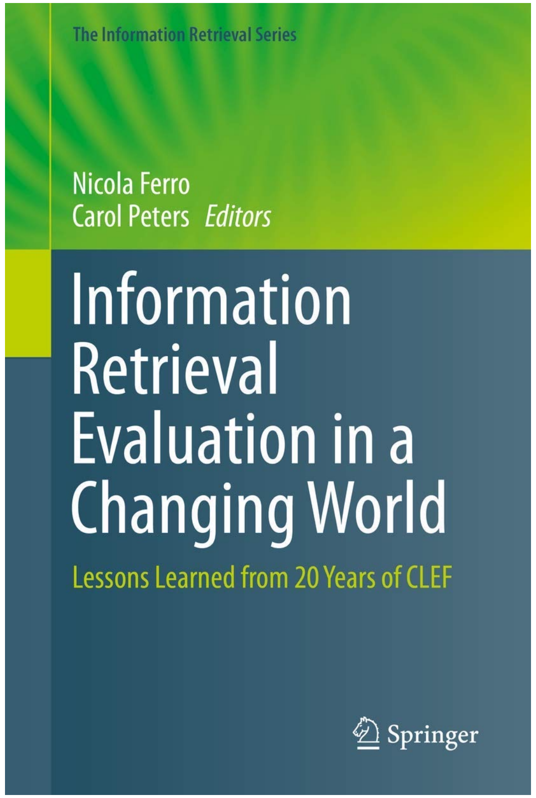


Sakai, Oard, and Kando. Evaluating Information Retrieval and Access Tasks, Springer Verlag, 2021. Available open access at https://doi.org/10.1007/978-981-15-5554-1

CLEF Initiative

universität koblenz
weiter:denken

"This volume celebrates the twentieth anniversary of CLEF - the Cross- Language Evaluation Forum for the first ten years, and the Conference and Labs of the Evaluation Forum since – and traces its evolution over these first two decades.



Ferro and Peters. Information Retrieval Evaluation in a Changing World. Lessons Learned from 20 years of CLEF, Springer Verlag, 2019. Available at https://doi.org/ 10.1007/978-3-030-22948-1

Forum for IR Evaluation

INFORMATION RETRIEVAL SOCIETY OF INDIA
IRSI

FIRE 2024
Forum for Information Retrieval Evaluation

- Home
- Organization
- Contact Us

- Archives
- Data
- Resources
- Past Proceedings
- FIRE
  - 2023

Welcome

The 16th meeting of *Forum for Information Retrieval Evaluation 2024* will be held in India. FIRE started in 2008 with the aim of building a South Asian counterpart for TREC, CLEF and NTCIR, and has since evolved continuously to support and encourage research within the information retrieval community. FIRE has adapted to meet the new challenges in multilingual information access and frameworks for large-scale evaluation of information retrieval methods, primarily text.

SPONSORS

To be announced soon.

PUBLICATIONS

To be announced soon.

Source: http://fire.irsi.res.in/fire/2024/home

Example Query/Topic (TREC 8)

<num> Number: 412

<title> airport security

<desc> Description

What security measures are in effect or are proposed to go into effect in airports?

<narr> Narrative

A relevant document could identify a specific airport and describe the security measures already in effect or proposed for use at that airport. Relevant items could also describe a failure of security that was cited as a contributing cause of a tragedy which came to pass or which was later averted. Comparisons between and among airports based on the effectiveness of the security of each are also relevant.

Corpora

- Classical corpora
  - Small, first testing

Corpus	Composition	Docs	Topics
Cranfield	Articles on aerodynamics	1,400	225
MED	Biomedical articles	1,033	30
TIME	News	425	83
CACM	Computing science papers	3,204	52

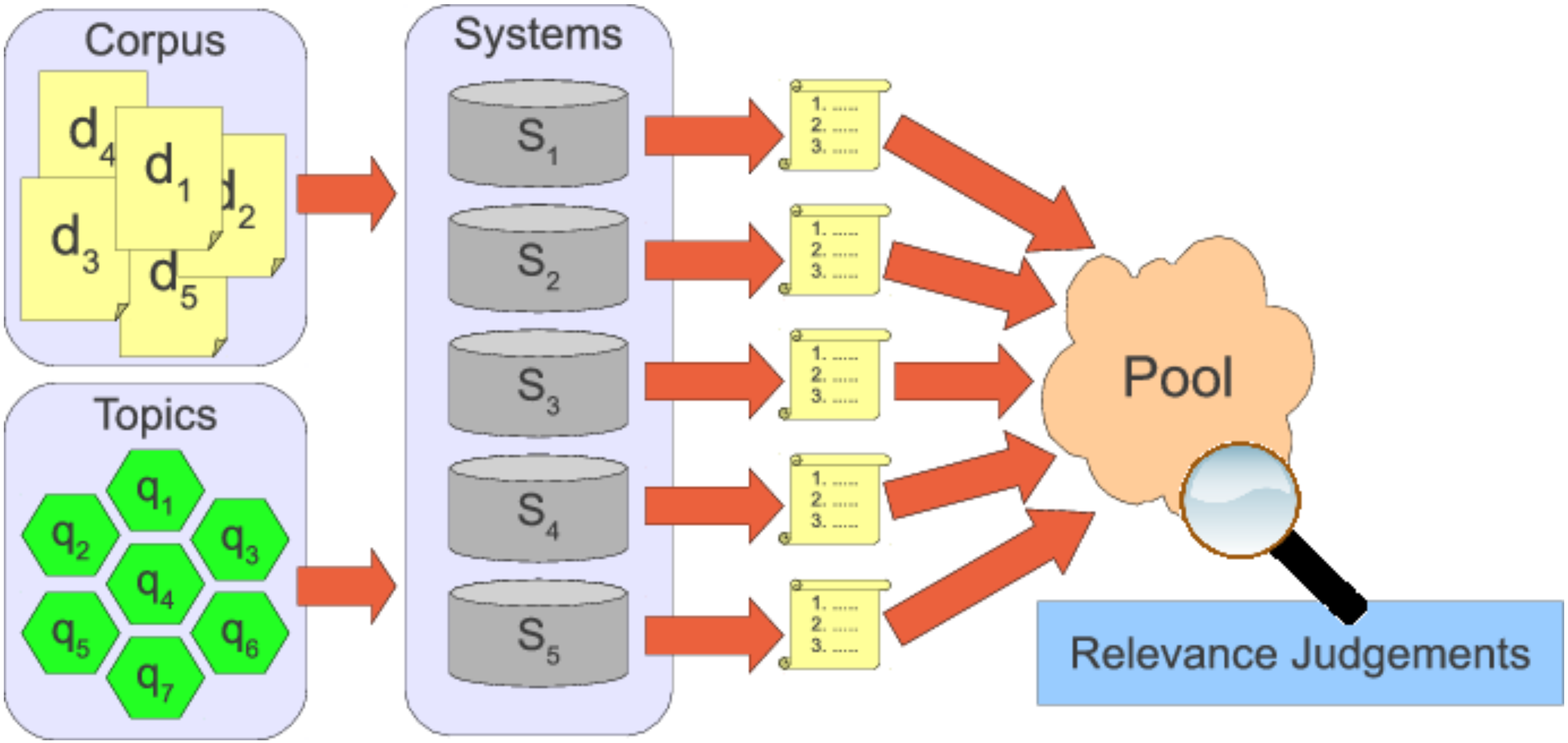
- Modern corpora
  - TREC, CLEF
  - Large, Very large
  - Different tasks
- Reuters CV1, CV2

Creating Relevance Assessments

- Assessor
  - Specialists
  - Computer support
  - Fast document scanning
- Old collections
  - Complete judgements
- But: TREC Terabyte Ad hoc Track 2005
  - 25.000.000 Documents, 50 Topics
  - Required time (theoretic)
    - 40 assessors, 10s / document, 8h /day
    - Total: 29.7 years
- Solution: Pooling

Pooling

universität koblenz
weiter:denken



Crowdsourcing Relevance Judgements

- Use non-professional assessors
  - Massive parallel assessments
  - Established platform: Amazon Mechanical Turk
- Benefits
  - Fast
  - Cheap: 0.01 to 0.05 cents per judgement
- Issues
  - Agreement of assessors
  - Spam
  - User interface

Kazai, G., Kamps, J. & Milic-Frayling, N. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf Retrieval* **16**, 138–178 (2013). https://doi.org/10.1007/s10791-012-9205-0

LLMs for relevance assessments

universität koblenz
weiter:denken

opinion

DOI:10.1145/3624730 Gianluca Demartini et al.

Opinion

Who Determines What Is Relevant? Humans or AI? Why Not Both?

A spectrum of human–artificial intelligence collaboration in assessing relevance.

To measure progress on better methods for Web search, question answering, conversational agents, or retrieval from knowledge bases, it is essential to know which responses are relevant to a user's information need. Such judgments of what is relevant are traditionally obtained by asking human assessors.

With the latest improvements on autoregressive large language models (LLMs) such as ChatGPT, researchers started to experiment with the idea of replacing human relevance assessment by LLMs.[9] The approach is simple: Just ask an LLM chatbot whether a response is relevant for an information need, and it does provide an "opinion."

In recent empirical studies on Web search[3] but also in programming,[7] human–computer interaction,[5] or protein function prediction,[10] it has been shown that LLM-generated opinions often agree with the assessment of humans. Some people readily believe the decision on what is relevant can be outsourced to artificial intelligence (AI) in the form of LLMs, without any human involvement.

However, as we argue here, there are several issues with such a fully automated judgment approach—and these issues cannot be overcome by a technical solution. Rather than continuing with the ongoing quest to study where and how AI can replace humans, we suggest to examine forms of human–AI collaboration for which we lay out a spectrum in this column.

IMAGE CREATED BY ANDRIJ BORYS ASSOCIATES WITH THE ASSISTANCE OF SHUTTERSTOCK.AI

Why Not Just Use LLMs?

There are a number of issues that arise when we let LLMs judge the quality of search results or system-provided answers.

Judgment bias toward a particular LLM. If we use a particular LLM to create relevance judgments to measure system quality, it would likely favor results from systems that use the same or a similar LLM for response generation. Such a bias in the gold standard benchmark can lead to wrong findings when comparing multiple systems for quality.

Bias toward user groups. Bender et al.[1] highlight the severe risk of LLMs to bias against underrepresented user groups. Such bias will likely be reflect-

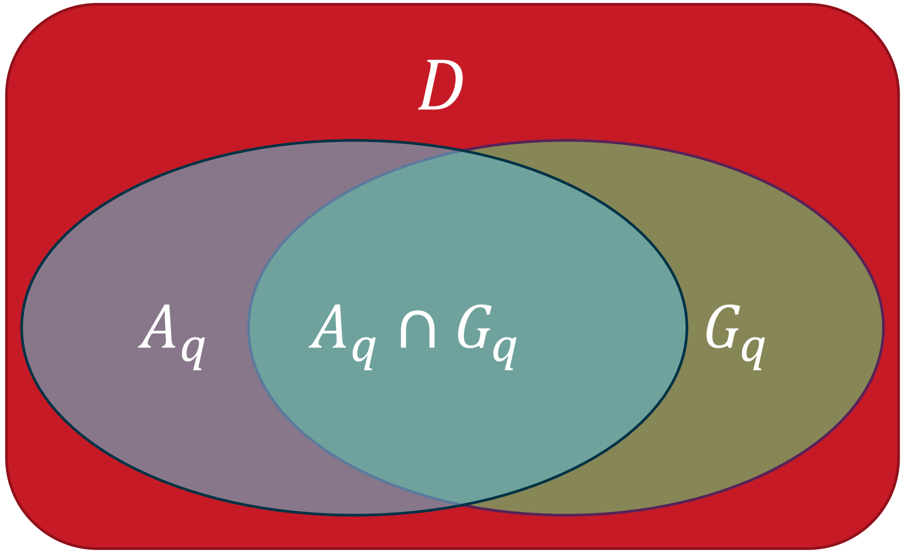
Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2024. Who Determines What Is Relevant? Humans or AI? Why Not Both? Commun. ACM 67, 4 (April 2024), 31–34. https://doi.org/10.1145/3624730

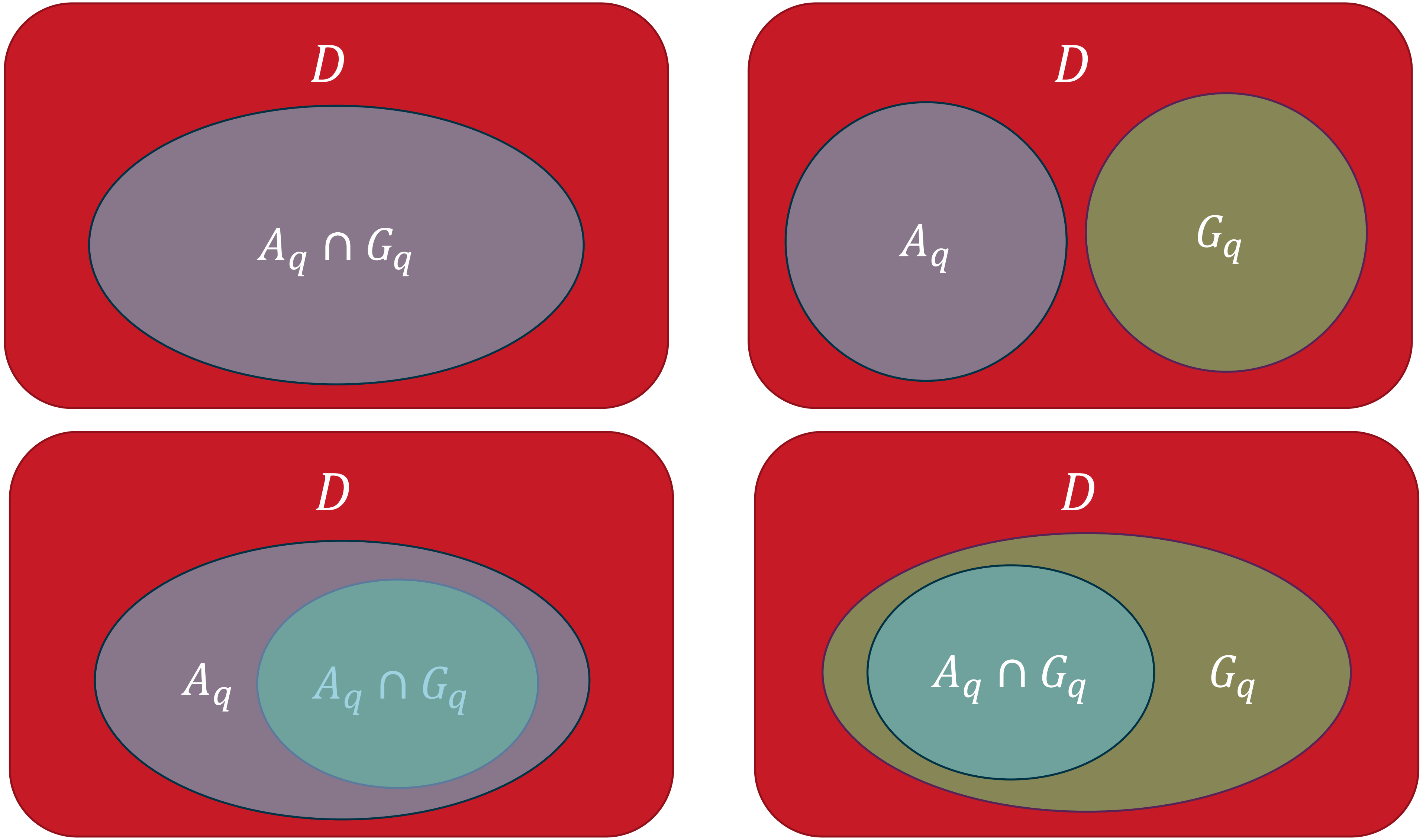
Evaluation Metrics

Metrics ignoring the ranking

A typical retrieval scenario

- $D = \{d_1, d_2, \dots, d_N\}$ is the collection of N resources
- q is the query
- G_q is the gold standard set that corresponds to q
- A_q is the retrieved result given q




D
$A_q \cap G_q$
D
A_q
G_q
D
A_q
$A_q \cap G_q$
D
$A_q \cap G_q$
G_q

Confusion matrix

Each document d is either retrieved or not, and either relevant or not. This induces the following confusion matrix:

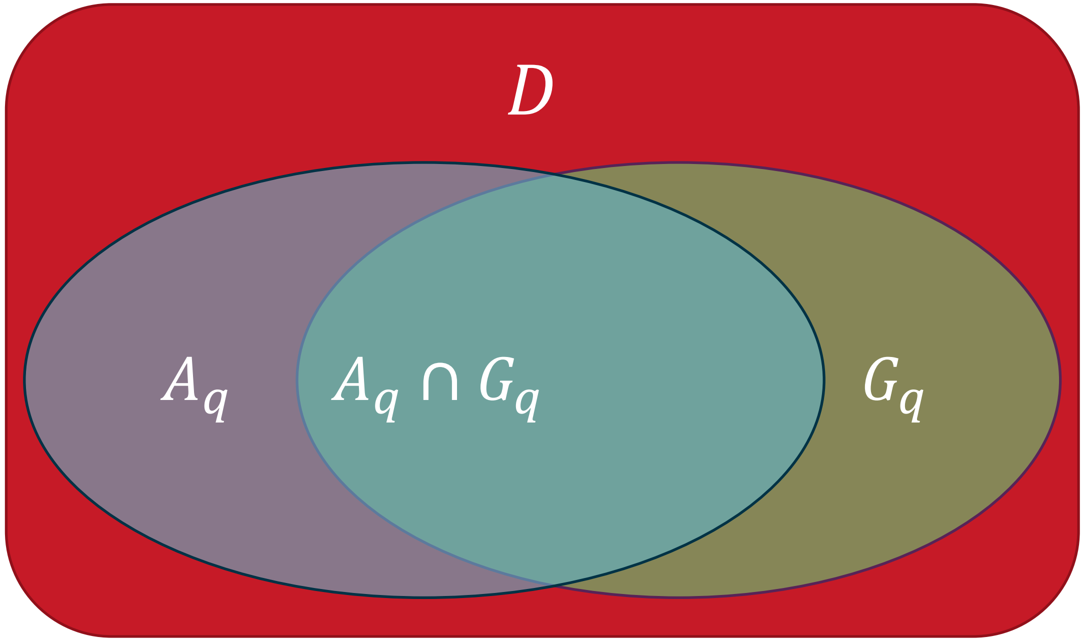
$$\begin{array}{r c} & \begin{array}{cc} \text{relevant} & \text{not relevant} \end{array} \\ \begin{array}{r} \text{retrieved} \\ \text{not retrieved} \end{array} & \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \end{array}$$

$$\begin{array}{r c} & \begin{array}{cc} \text{relevant} & \text{not relevant} \end{array} \\ \begin{array}{r} \text{retrieved} \\ \text{not retrieved} \end{array} & \begin{pmatrix} hits & noise \\ misses & rejected \end{pmatrix} \end{array}$$

Confusion matrix

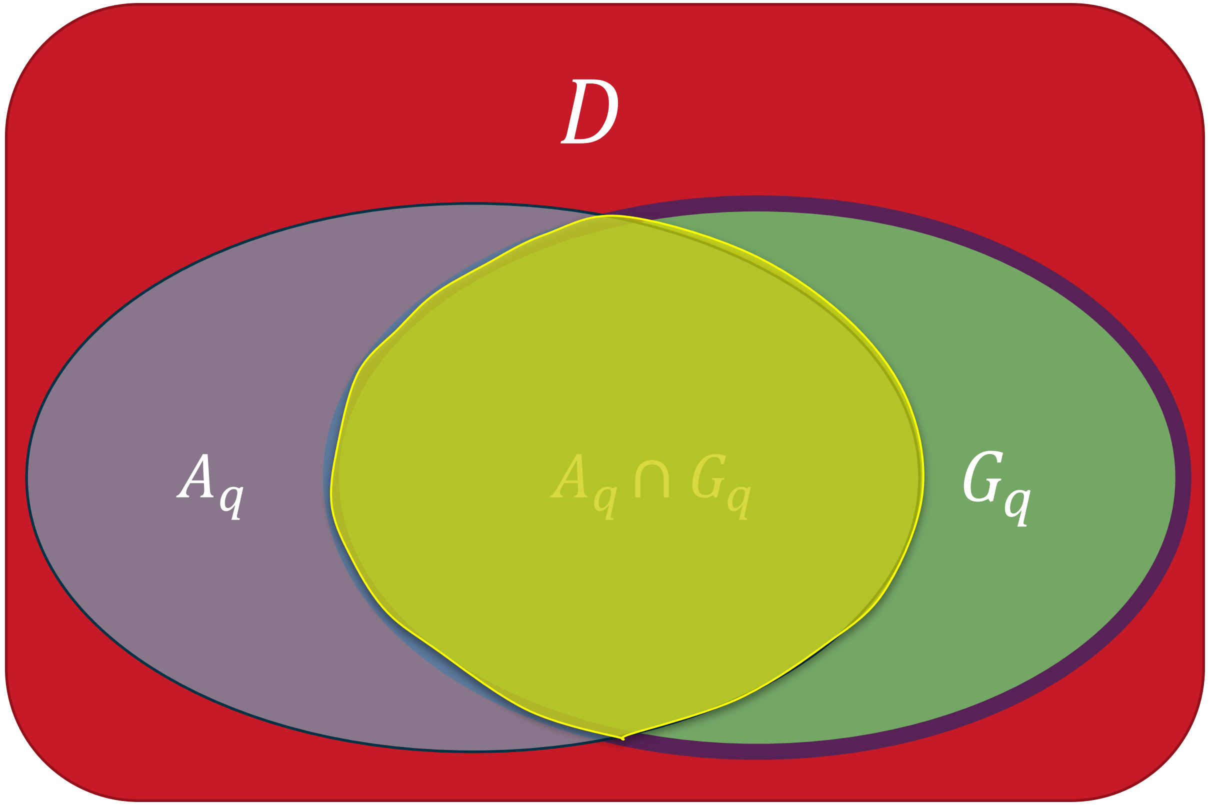
$$\begin{array}{r c} & \begin{array}{cc} \text{relevant} & \text{not relevant} \end{array} \\ \begin{array}{r} \text{retrieved} \\ \text{not retrieved} \end{array} & \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \end{array}$$

- $A_q \cap G_q = TP$
- $G_q = TP + FN$
- $A_q = TP + FP$



Recall

- Among all relevant resources, which fraction is retrieved?
- $r = \frac{|A_q \cap G_q|}{|G_q|}$
- $r = \frac{TP}{TP+FN}$

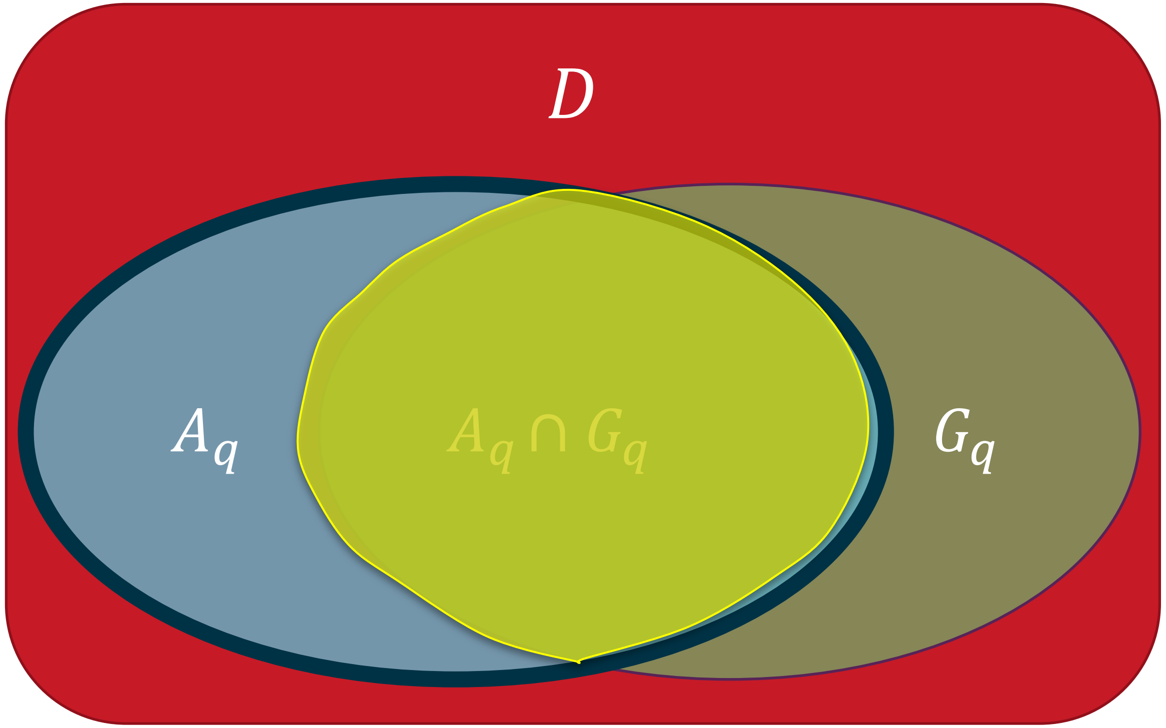


Recall: an example

- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Recall
  - $r = \frac{|A_q \cap G_q|}{|G_q|} = \frac{|\{d_2, d_3, d_8, d_{10}, d_{17}, d_{29}\}|}{|\{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}|} = \frac{6}{8} = 0{,}75$

Precision

- Among all retrieved resources, which fraction is relevant?
- $p = \frac{|A_q \cap G_q|}{|A_q|}$
- $p = \frac{TP}{TP+FP}$



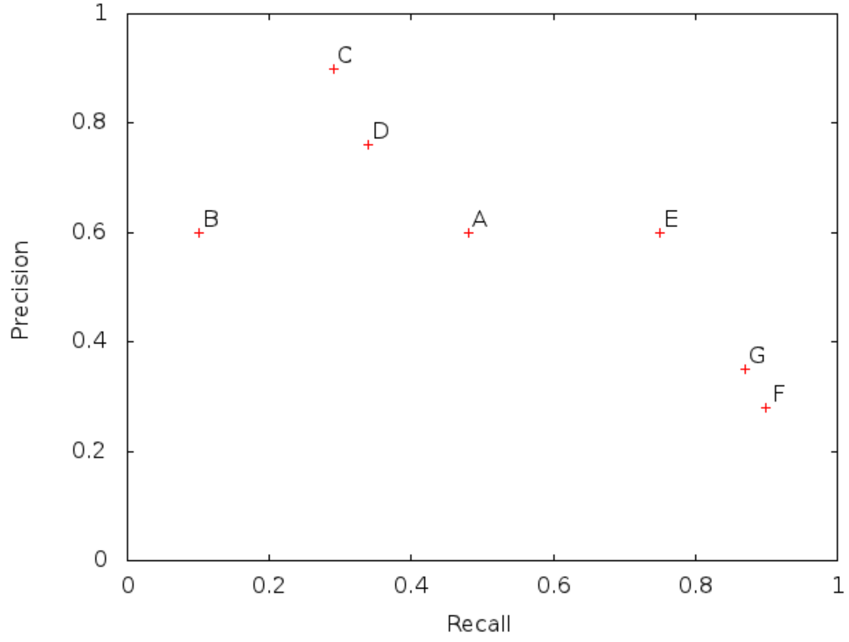
Precision: an example

- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Precision
  - $p = \frac{|A_q \cap G_q|}{|A_q|} = \frac{|\{d_2, d_3, d_8, d_{10}, d_{17}, d_{29}\}|}{|\{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}|} = \frac{6}{10} = 0{,}6$

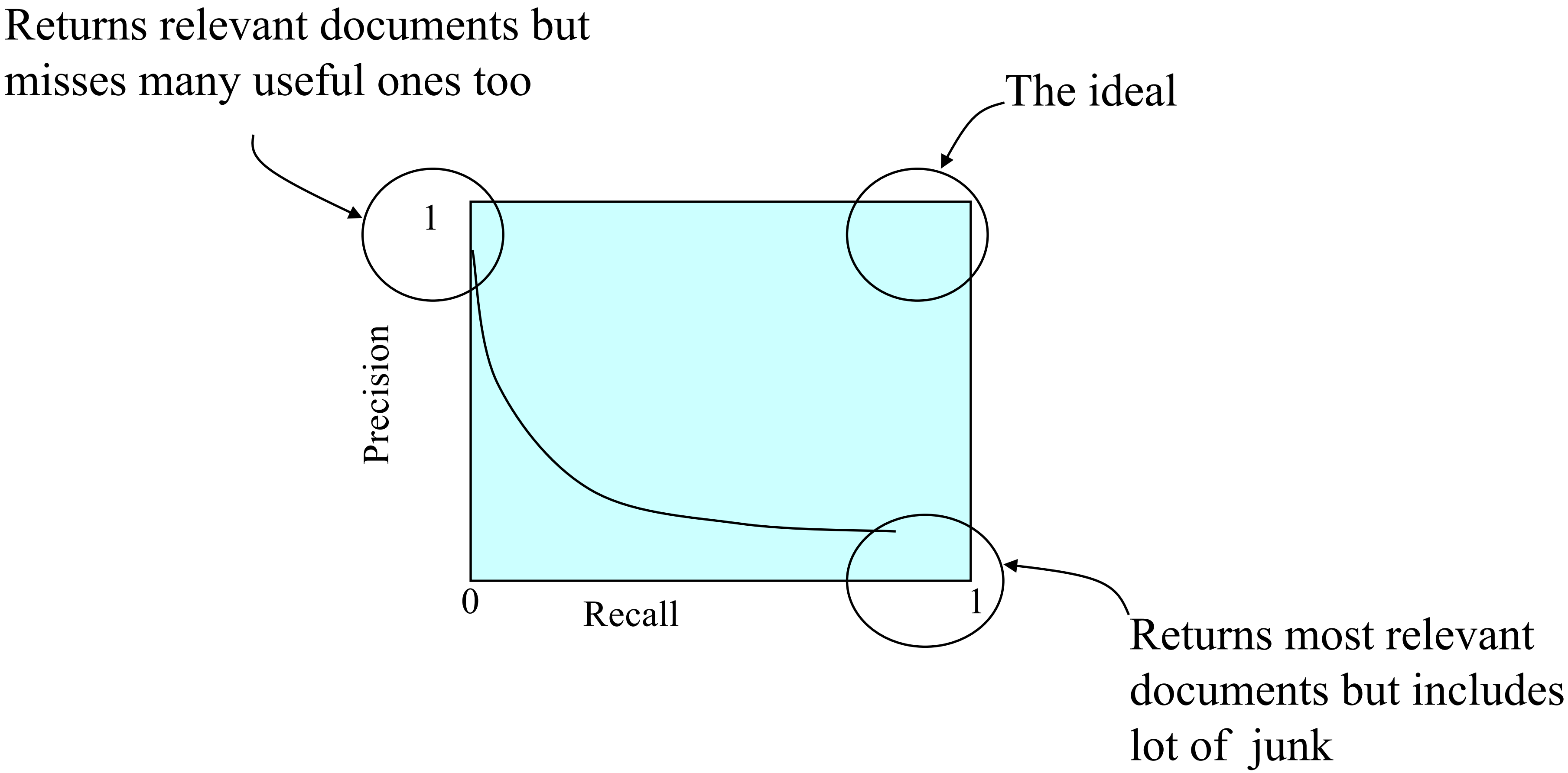
Properties of Precision and Recall

- Range [0,1]
- High values are better
  - Recall of 1 can always be obtained
  - High precision can be influenced
- Values are opposed
- How to compare Systems?
  - Application might dictate preference of recall or precision

System	Recall	Precision
A	0.48	0.60
B	0.10	0.60
C	0.29	0.90
D	0.34	0.76
E	0.75	0.60
F	0.90	0.28
G	0.87	0.35



Trade-offs


Returns relevant documents but
misses many useful ones too
The ideal
1
Precision
0
Recall
1
Returns most relevant
documents but includes
lot of junk

F-Measure

universität koblenz
weiter:denken

- Combines recall and precision (weighted harmonic mean)

$$H_\alpha(r,p) = \frac{1}{\alpha \frac{1}{p} + (1-\alpha)\frac{1}{r}}$$

- Typically formulated as F-Measure:

$$F_\beta = (\beta^2 + 1)\frac{rp}{\beta^2 p + r}$$ by setting $$\alpha = \frac{\beta^2}{\beta^2 + 1}$$

- (Nearly) always used with β=1 : $$F_1 = \frac{2rp}{p+r}$$

- This means that the precision and recall are equaly important
- If precision is more important than recall, we set $\beta < 1$. Otherwise, we set $\beta > 1$

van Rijsbergen

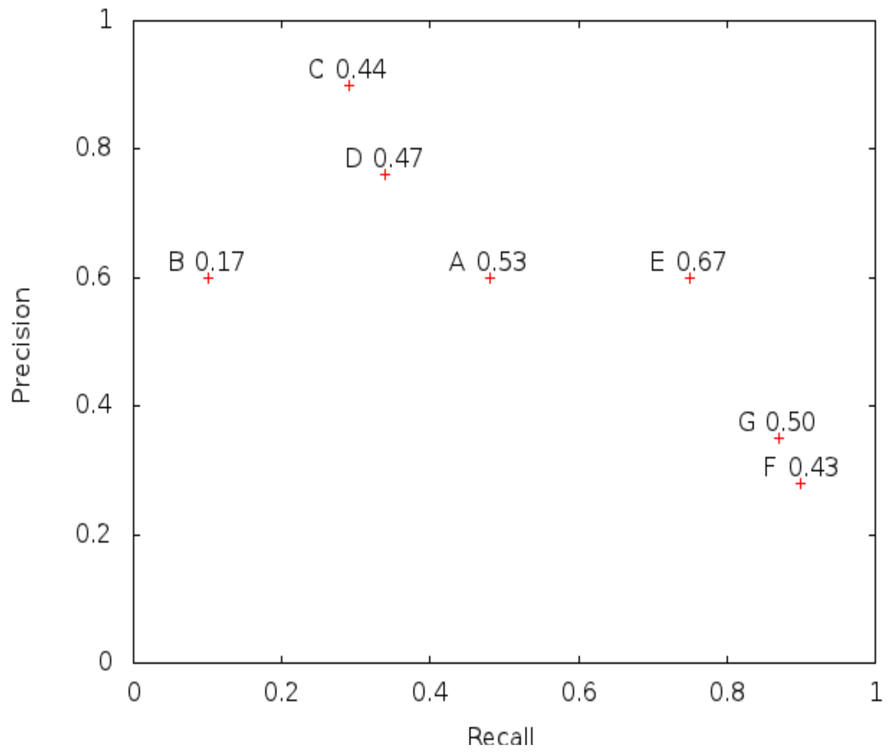
F1-Score: an example

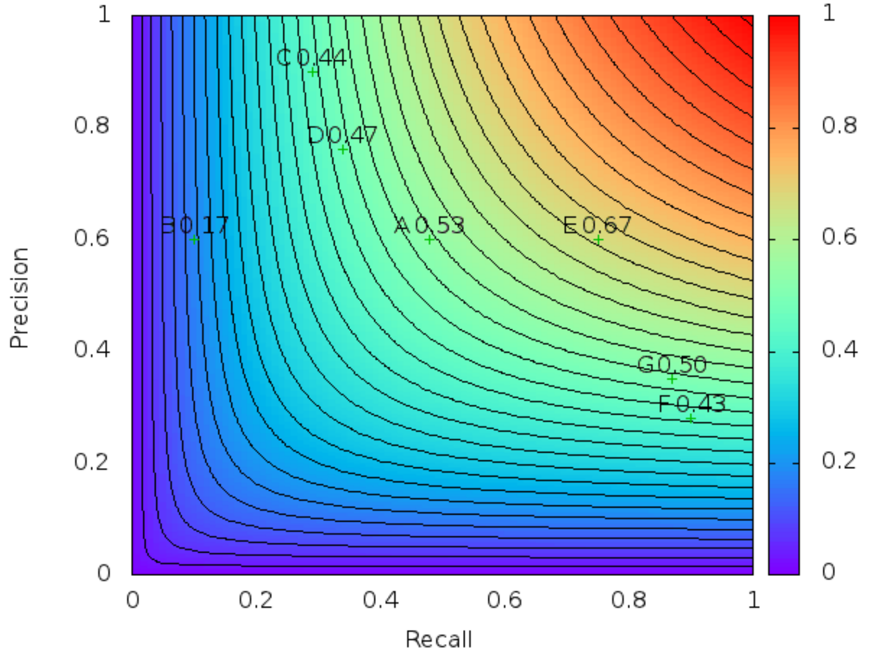
- Given
  - the collection $D = \{d_1, d_2, \ldots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- F1-score
  - $F1 = 2\frac{rp}{p+r} = \frac{2\times0.6\times0.75}{0.6+0.75} = 0.667$

Properties of F1

- Range [0,1]
- High values are better

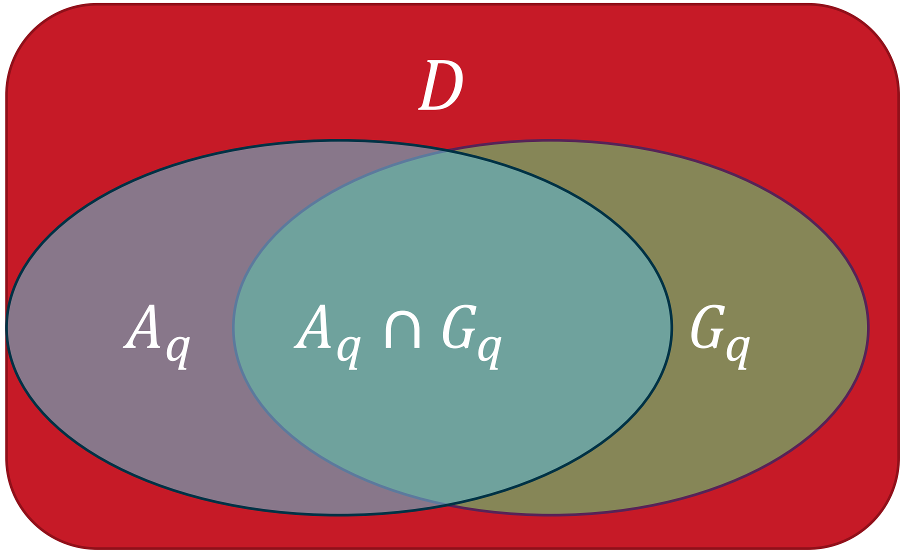
System	Recall	Precision	F1
A	0.48	0.60	0.53
B	0.10	0.60	0.17
C	0.29	0.90	0.44
D	0.34	0.76	0.47
E	0.75	0.60	0.67
F	0.90	0.30	0.43
G	0.87	0.32	0.50





Accuracy

- Accuracy is the fraction of correct decisions
  - $Acc = \frac{TP+TN}{TP+TN+FP+FN}$
  - $= \frac{|A_q \cap G_q| + |D \setminus \{A_q \cup G_q\}|}{|D|}$



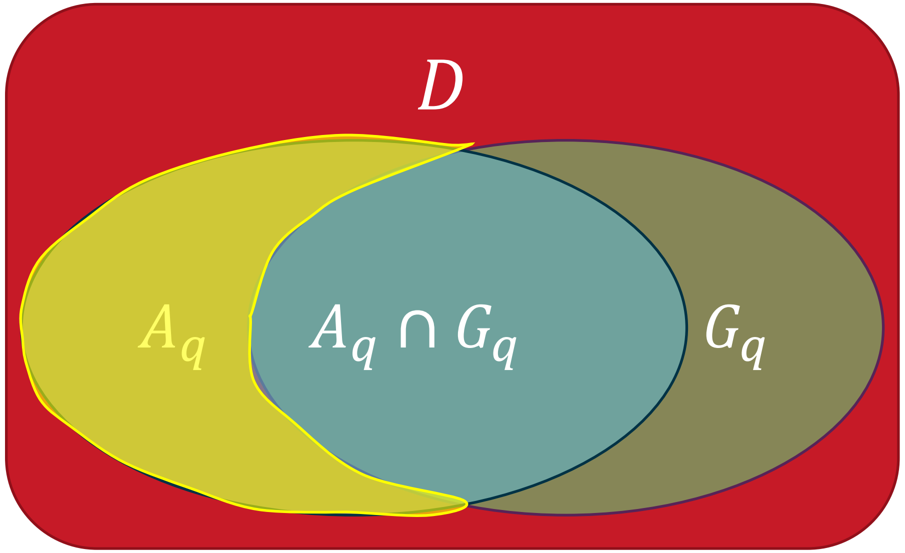
- Considering the size of D, accuracy is not a good measure for IR systems.
- If for every query, a ($a \to |D|$) resources are not relevant, a system which does not retrieve anything will get an accuracy $= a/|D|$

Accuracy: an example

- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Accuracy
  - $Acc = \frac{A_q \cap G_q + D \setminus \{A_q \cup G_q\}}{D} = \frac{|\{d_2, d_3, d_8, d_{10}, d_{17}, d_{29}\}| + |\{d_1, \dots\}|}{\{d_1, d_2, \dots, d_{100}\}} = \frac{6+88}{100} = 0.94$

Fallout

- Fallout is the fraction of the noise that the system exposes to the user
  - $Fallout = \frac{|A_q \setminus G_q|}{|D \setminus G_q|}$



- Considering the size of D, fallout is of little use to evaluate IR systems

Fallout: an example

- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Fallout
  - $Fallout = \frac{|A_q \setminus G_q|}{|D \setminus G_q|} = \frac{|\{d_4, d_7, d_{12}, d_{20}\}|}{\{d_1, d_4, \dots\}} = \frac{4}{92} = 0.043$

Evaluation

- Precision, Recall, F-score are good for evaluating the performance of Boolean retrieval systems (Relevant and Non-relevant)
- They cannot evaluate rankings
- For example, [R,R,N,N] and [N,N,R,R] will be evaluated similarly by these measures
  - R: relevant
  - N: Non-relevant

universität koblenz
weiter:denken

Ranking Aware Metrics

Typical Ranked Retrieval Setting

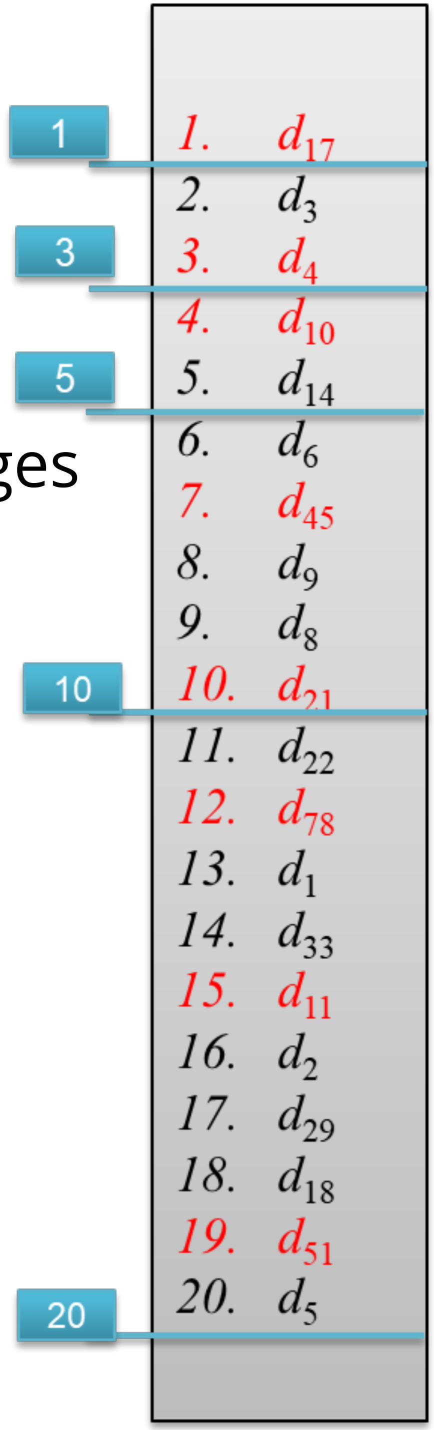
- $D = \{d_1, d_2, \dots, d_N\}$ is the collection of N resources
- q is the query
- G_q is the gold standard set that corresponds to q
- L_q is the ordered retrieved result given q
  - Order of relevance
- Example
  - $G_q = \{d_4,\ d_{10},\ d_{11},\ d_{17},\ d_{21},\ d_{45},\ d_{51},\ d_{78}\}$

G_q	$\{d_4,\quad d_{10},\quad d_{11},\quad d_{17},\quad d_{21},\quad d_{45},\quad d_{51},\quad d_{78}\}$
L_q	$\{d_{17}, d_3, d_4, d_{10}, d_{14}, d_6, d_{45}, d_9, d_8, d_{21}, d_{22}, d_{78}, d_1, d_{33}, d_{11}, d_2, d_{29}, d_{18}, d_{51}, d_5\}$
	$\{d_{17}, d_3, d_4, d_{10}, d_{14}, d_6, d_{45}, d_9, d_8, d_{21}, d_{22}, d_{78}, d_1, d_{33}, d_{11}, d_2, d_{29}, d_{18}, d_{51}, d_5\}$

Precision at k (p@k)

- Fixed cutoff (k) in results list
- Motivation from UI
  - Systems deliver chunks of result list as pages
  - Users rarely go beyond first page
- Determine precision at cutoff (p@k)
- Example

k	# relevant docs	p@k
1	1	1.000
3	2	0.667
5	3	0.600
10	5	0.500
20	8	0.400



universität koblenz
weiter:denken

R-Precision

- Problem of p@k
  - Choice of k?
  - Less than k relevant documents
  - Stability
- R-Precision
  - Flexible cutoff at $|G|$
  - Precision-recall break-even: $|G| = |A|$
- Example
  - $G_q = \{d_4,\ d_{10},\ d_{11},\ d_{17},\ d_{21},\ d_{45},\ d_{51},\ d_{78}\}$
  - $p_R = \frac{4}{8}$

1. d_{17}
2. d_3
3. d_4
4. d_{10}
5. d_{14}
6. d_6
7. d_{45}
8. d_9

$|G|$

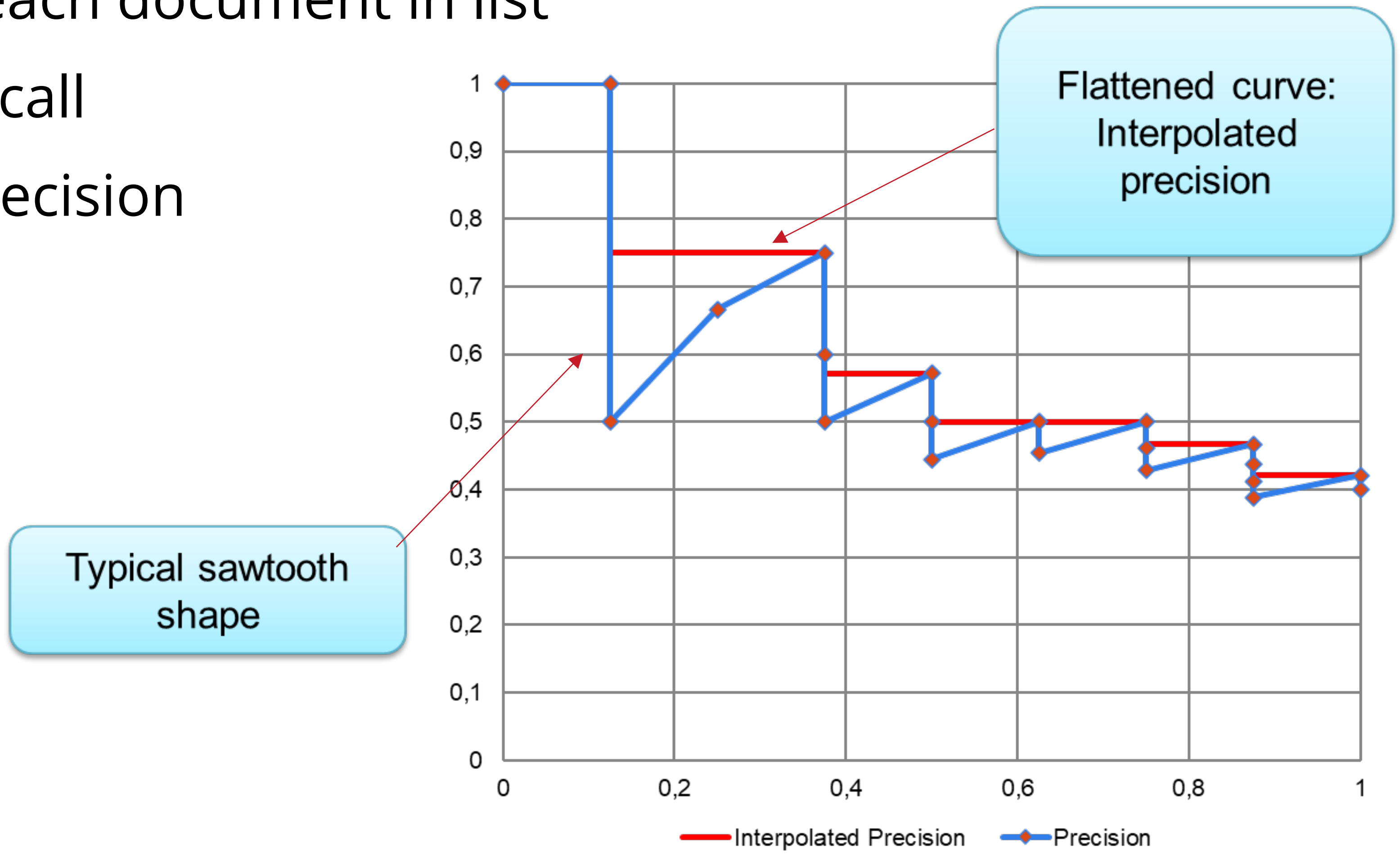
9. d_8
10. d_{21}
11. d_{22}
12. d_{78}
13. d_1
14. d_{33}
15. d_{11}
16. d_2
17. d_{29}
18. d_{18}
19. d_{51}
20. d_5

Precision Recall Graph

- Plot evolution of recall and precision in result list (no function)
- For each document in list

  x: recall

  y: precision

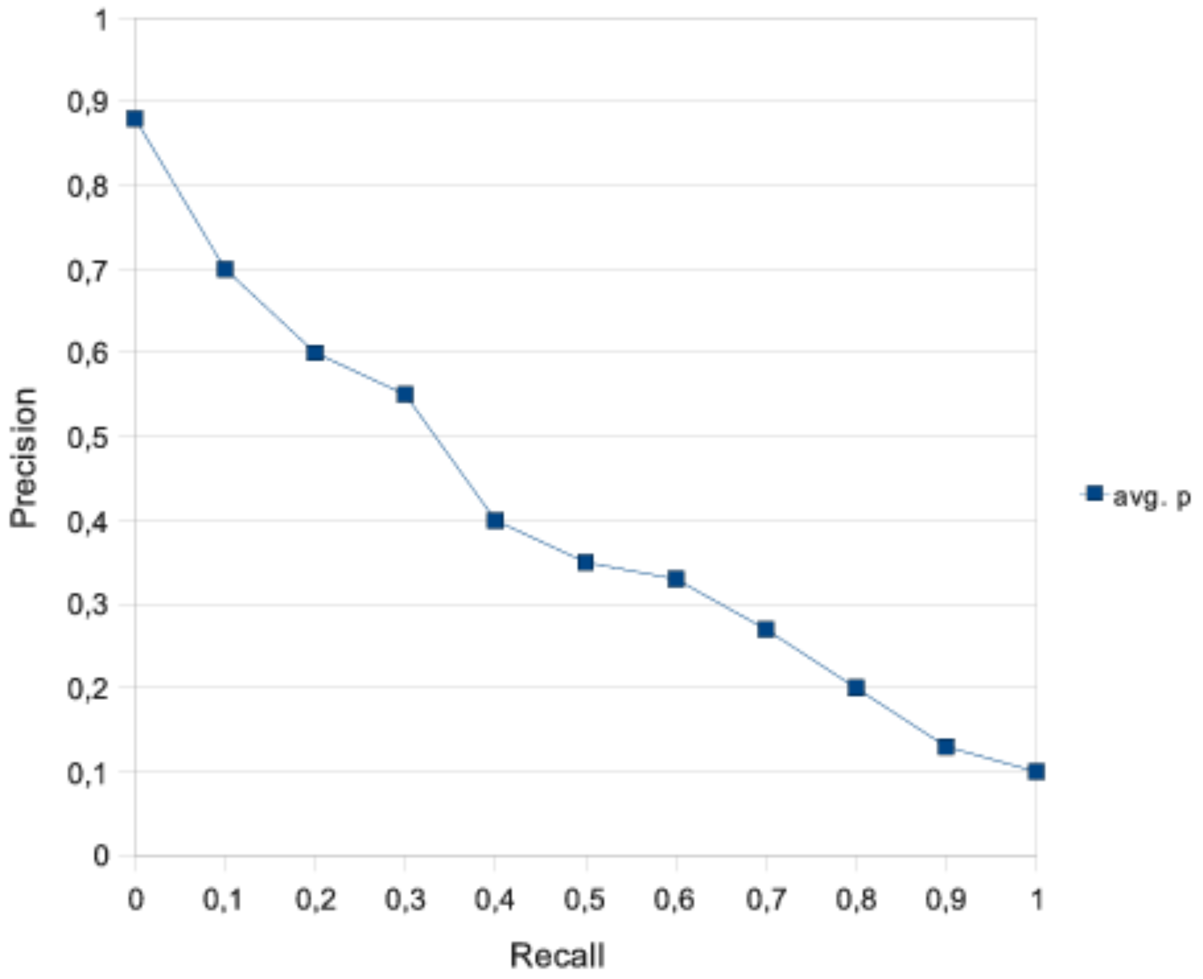


1. d_{17}
2. d_{3}
3. d_{4}
4. d_{10}
5. d_{14}
6. d_{6}
7. d_{45}
8. d_{9}
9. d_{8}
10. d_{21}
11. d_{22}
12. d_{78}
13. d_{1}
14. d_{33}
15. d_{11}
16. d_{2}
17. d_{29}
18. d_{18}
19. d_{51}
20. d_{5}

11-Point Precision Recall Graph

- Fixed set of recall values
  - 0 to 1, steps 0.1
  - Interpolated precision

    p_interp(r) = max_{r' >= r} p(r')



Mean Average Precision

- Mean Average Precision:

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{j=1}^{m_i}P\left(k_{ij}\right)$$

- One integrated value for the quality of a ranking
  - m_i number of relevant documents for query q_i
  - k_{ij} position of the j-th relevant document for query q_i
  - $P(k_{ij})$ precision @ k_{ij} for query q_i (set to 0 if document is not in the result list)

MAP – an example

- Average precision (AP) for one query

Document	Position	Precision
d_{17}	1	1.000
d_4	3	0.667
d_{10}	4	0.750
d_{45}	7	0.571
d_{21}	10	0.500
d_{78}	12	0.500
d_{11}	15	0.467
d_{51}	19	0.421
Average Precision		***0.609***

1.	d_{17}
2.	d_3
3.	d_4
4.	d_{10}
5.	d_{14}
6.	d_6
7.	d_{45}
8.	d_9
9.	d_8
10.	d_{21}
11.	d_{22}
12.	d_{78}
13.	d_1
14.	d_{33}
15.	d_{11}
16.	d_2
17.	d_{29}
18.	d_{18}
19.	d_{51}
20.	d_5

- MAP: Mean over AP for several queries

MAP – an example

universität koblenz
weiter:denken

- Assume two documents are missing in the result set

Document	Position	Precision
d_{17}	1	1.000
d_4	3	0.667
d_{10}	4	0.750
d_{45}	7	0.571
d_{21}	10	0.500
d_{78}	12	0.500
d_{11}	15	0.467
d_{51}	19	0.421
d_{73}	-	0
d_{39}	-	0
Average Precision		***0.488***

1. d_{17}
2. d_3
3. d_4
4. d_{10}
5. d_{14}
6. d_6
7. d_{45}
8. d_9
9. d_8
10. d_{21}
11. d_{22}
12. d_{78}
13. d_1
14. d_{33}
15. d_{11}
16. d_2
17. d_{29}
18. d_{18}
19. d_{51}
20. d_5

Further Evaluation Approaches

universität koblenz
weiter:denken

Indirect Measures

- User behaviour when seeking information
  - Time
  - Number of interactions
  - Viewed documents
  - Query modifications
  - Methods:
    - Clickstream mining
    - Lab tests, observation
- User surveys
  - Ask for satisfaction
  - A/B testing

Evaluation at large search engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top k, e.g., k = 10
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right.
  - NDCG (Normalized Cumulative Discounted Gain)
- Search engines also use non-relevance-based measures
  - Clickthrough on first result
    - Not very reliable if you look at a single clickthrough … but pretty reliable in the aggregate
  - Studies of user behavior in the lab
  - A/B testing

A/B testing

universität koblenz
weiter:denken

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness
- Probably the evaluation methodology that large search engines trust most
- In principle less powerful than doing a multivariate regression analysis, but easier to understand

Summary

Summary

universität koblenz
weiter:denken

- At the end of this lecture, you are expected to
  - understand how to evaluate an IR system
  - understand the difference between evaluation measures that ignore the ranking and those that consider the ranking