MSCA ITN/ETN No. 860721

DoSSIER
Domain Specific Systems for
Information Extraction and Retrieval

Webinar on Basics of IR and Evaluations

January 21, 2021
Second part: Evaluations

Dr. **Marco Viviani**

University of Milano-Bicocca
(Milan, Italy)

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Introduction

- **Why** is an IR system subjected to an evaluation process?
  - How do we know if our results are any good?
  - It is difficult to establish whether the system has failed or not
    - What does it mean to “have failed”? → Next slides

- For this reason, an **evaluation methodology** must be developed and must be applied to working systems

Evaluating User «Happiness»

- How do we evaluate the retrieved results?
- “Measure” **user happiness**
- Happiness is hard to quantify
- Must be broken down into **quantifiable factors**

Some «Happiness Factors»

- **Relevance** → Next slides

- **Time** and **space**
  - How fast does it index
    - Number of documents/hour
    - (Average document size)
  - How fast does it search
  - …

EFFICIENCY

- **Coverage** of a topic
  - In Web Search Engines, usually dependent on coverage of crawlers

- **Expressiveness** of query language
  - Ability to express complex information needs

- **Usability** (Layout)

- …

Relevance as a Measure

- **Relevance is everything**!
  - How appropriate are the documents retrieved in satisfying the user's information needs
- **Subjective**, but one assumes it is **measurable**
  - Measurable to some extent
    - E.g., how often do people agree a document is relevant to a query
      - More often than expected
  - How well does it answer the question?
    - Complete answer? Partial?
    - Background Information?
    - Hints for further exploration?

SETTING UP AN EXPERIMENTAL EVALUATION

The Cranfield Paradigm

- **Our focus**: evaluating **retrieval EFFECTIVENESS**
  - The ability of a system to retrieve relevant documents while at the same time holding back non-relevant ones

- **Cranfield Paradigm** (by Cyril W. Cleverdon)
  - Dates back to 1960s
  - Makes use of experimental collections
  - Ensures comparability and repeatability of the experiments

Evaluations with Experimental Collections – Components

- The **components** of a standard evaluation experiment of an IR system (search engine), seen as a **black box**, are:
  - A benchmark document collection
  - A benchmark suite of information needs (or topics)
  - Relevance judgments (binary or graded), also called relevance assessments (or ground-truth, or qrels)

  - EVALUATION METRICS → Second part of this presentation

Evaluations with Experimental Collections – Relevance Judgements

- **Human experts' judgements**, for each information need (topic) and for each document
  - at least for subset of docs that some system returned for that information need

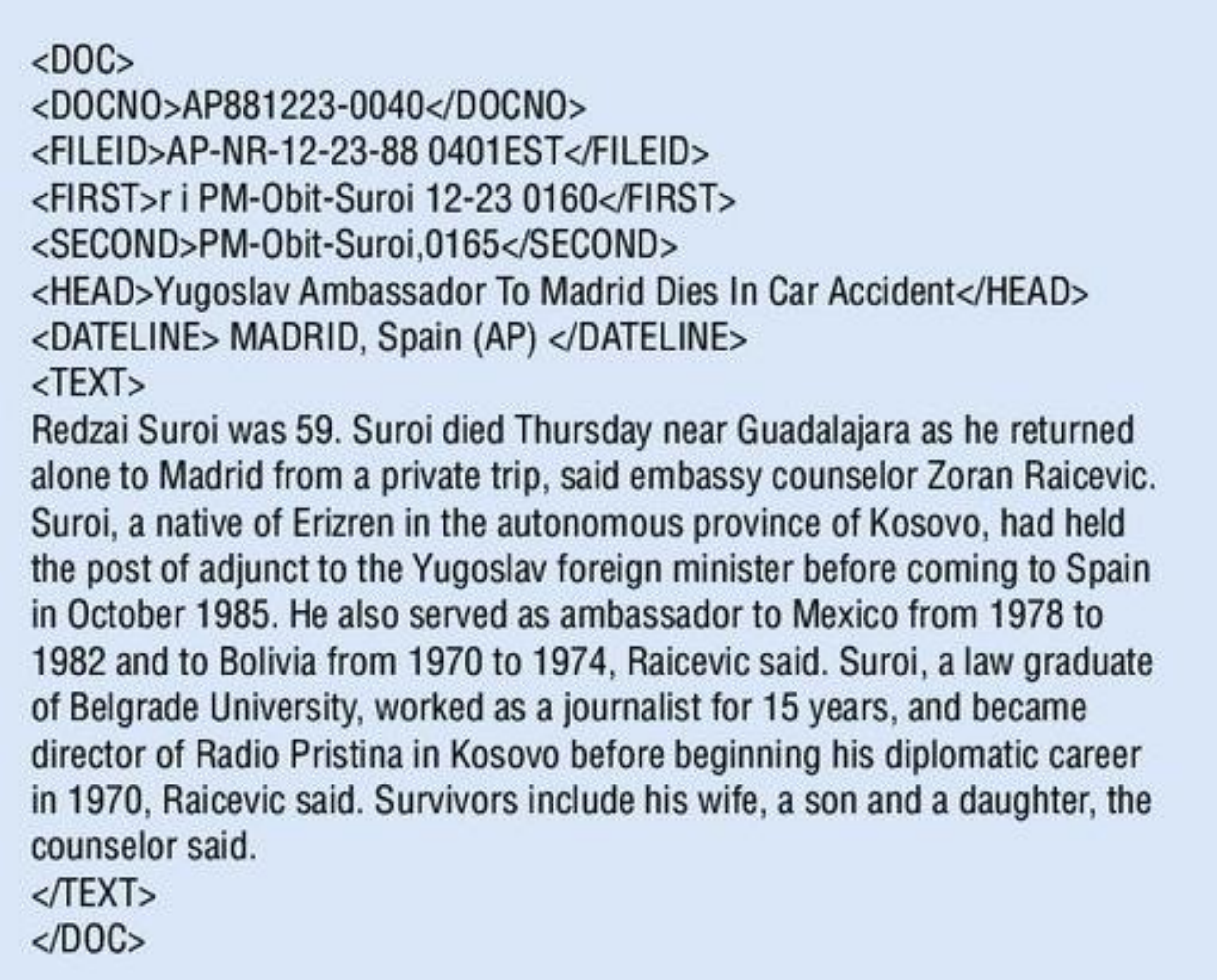
- **Binary judgements**
  - 1: “relevant”
  - 0: “non-relevant”

- **Multi-graded judgements**
  - 3: “highly relevant”
  - 2: “fairly relevant”
  - 1: “partially relevant”
  - 0: “not relevant”

Evaluations with Experimental Collections – Relevance Judgements

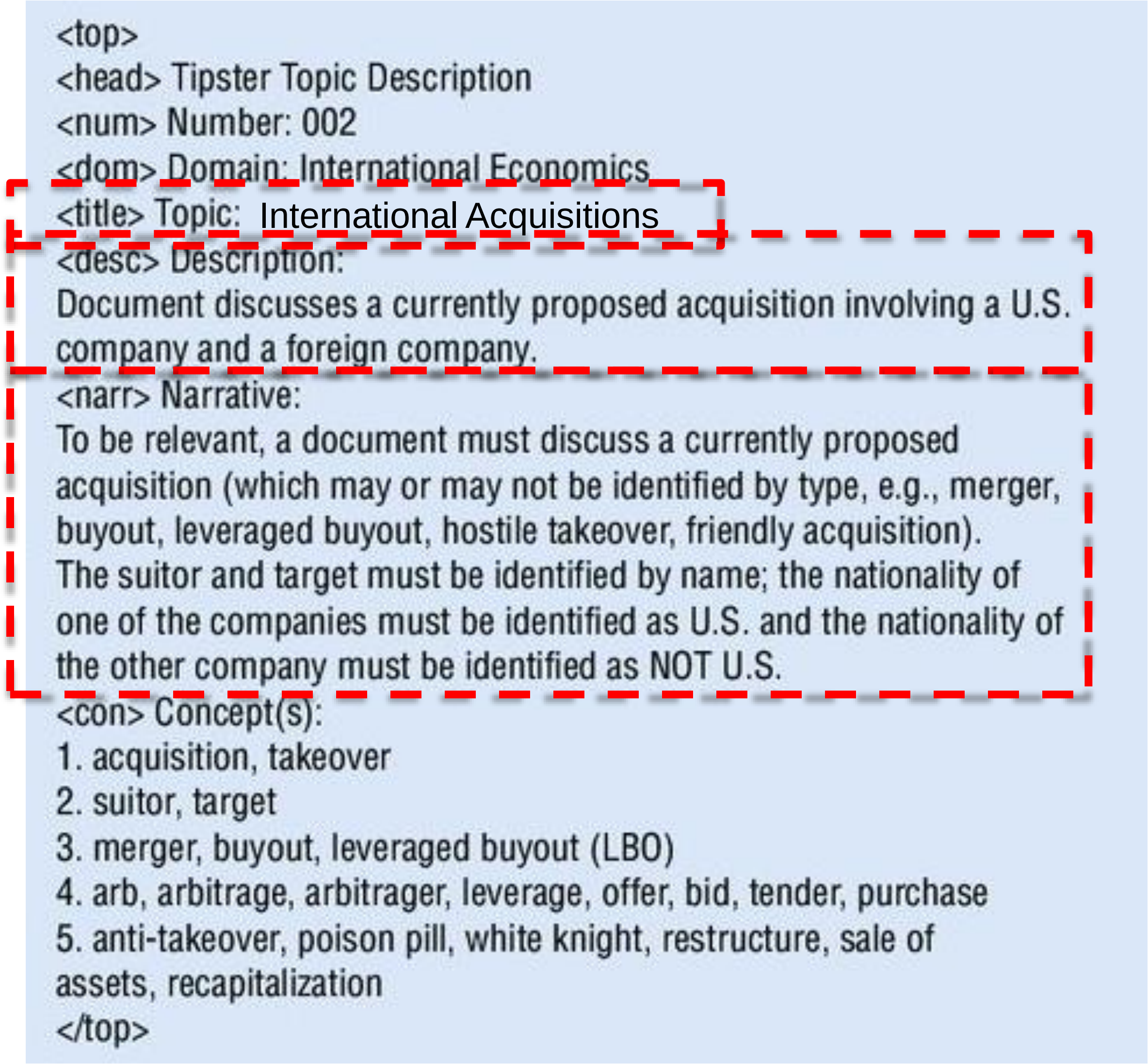
- Experts assign **relevance judgements** to documents in the collection <u>without using the system</u> to be evaluated

- The relevance of a document is **independent** of the relevance of other documents



```
<DOC>
<DOCNO>AP881223-0040</DOCNO>
<FILEID>AP-NR-12-23-88 0401EST</FILEID>
<FIRST>r i PM-Obit-Suroi 12-23 0160</FIRST>
<SECOND>PM-Obit-Suroi,0165</SECOND>
<HEAD>Yugoslav Ambassador To Madrid Dies In Car Accident</HEAD>
<DATELINE> MADRID, Spain (AP) </DATELINE>
<TEXT>
Redzai Suroi was 59. Suroi died Thursday near Guadalajara as he returned
alone to Madrid from a private trip, said embassy counselor Zoran Raicevic.
Suroi, a native of Erizren in the autonomous province of Kosovo, had held
the post of adjunct to the Yugoslav foreign minister before coming to Spain
in October 1985. He also served as ambassador to Mexico from 1978 to
1982 and to Bolivia from 1970 to 1974, Raicevic said. Suroi, a law graduate
of Belgrade University, worked as a journalist for 15 years, and became
director of Radio Pristina in Kosovo before beginning his diplomatic career
in 1970, Raicevic said. Survivors include his wife, a son and a daughter, the
counselor said.
</TEXT>
</DOC>
```

Evaluations with Experimental Collections – Topics

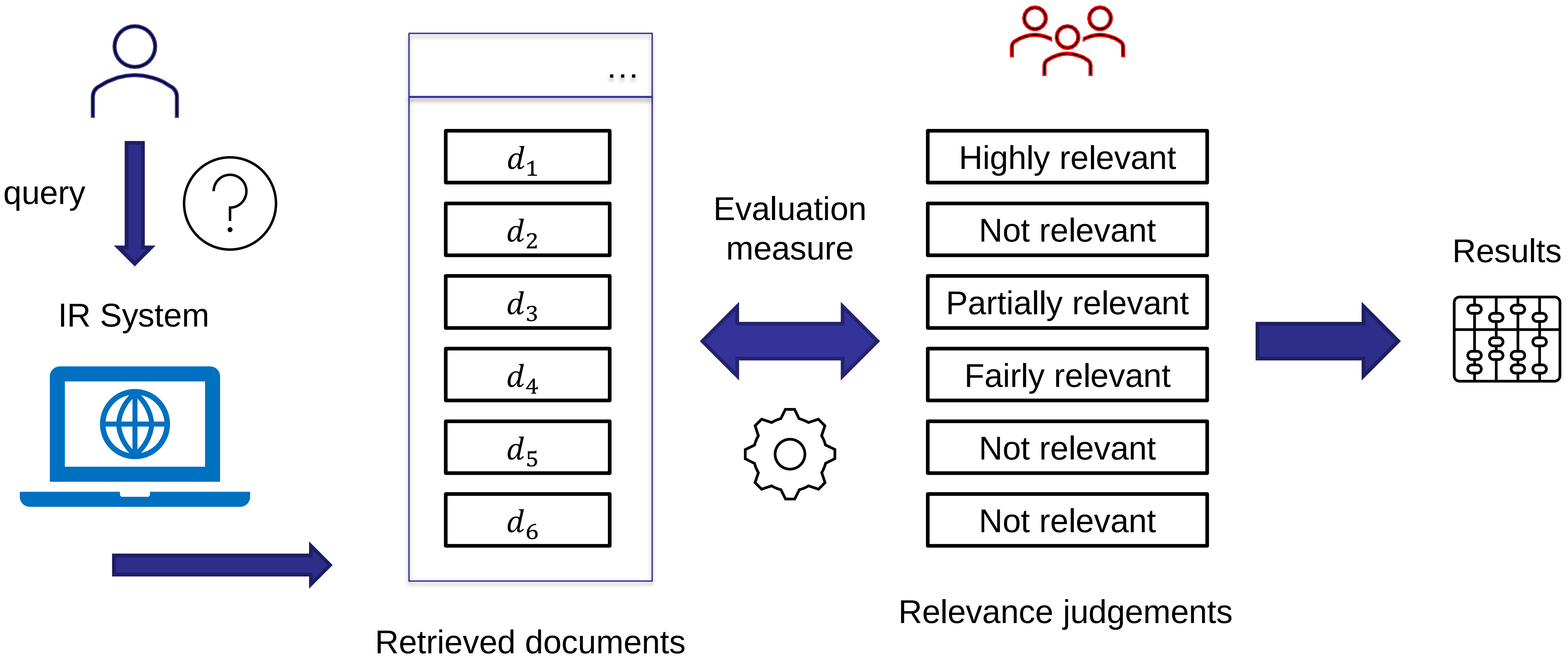


```
<top>
<head> Tipster Topic Description
<num> Number: 002
<dom> Domain: International Economics
<title> Topic:  International Acquisitions
<desc> Description:
Document discusses a currently proposed acquisition involving a U.S.
company and a foreign company.
<narr> Narrative:
To be relevant, a document must discuss a currently proposed
acquisition (which may or may not be identified by type, e.g., merger,
buyout, leveraged buyout, hostile takeover, friendly acquisition).
The suitor and target must be identified by name; the nationality of
one of the companies must be identified as U.S. and the nationality of
the other company must be identified as NOT U.S.
<con> Concept(s):
1. acquisition, takeover
2. suitor, target
3. merger, buyout, leveraged buyout (LBO)
4. arb, arbitrage, arbitrager, leverage, offer, bid, tender, purchase
5. anti-takeover, poison pill, white knight, restructure, sale of
assets, recapitalization
</top>
```

- **Topics consists of**:
  - Title: a brief statement expressing the information need
    - It resembles the typical search engine query
  - Description: more detailed formulation of the information need
  - Narrative: instructions for assessors on when to consider a document relevant

- Typical experimental collections make use of at least **50 topics**

Evaluations with Experimental Collections in a Nutshell



query

IR System

...

d_1

d_2

d_3

d_4

d_5

d_6

Retrieved documents

Evaluation measure

Highly relevant

Not relevant

Partially relevant

Fairly relevant

Not relevant

Not relevant

Relevance judgements

Results

Large-scale Evaluation Initiatives

- **TREC** (Text REtrieval Conference)
  - USA, since 1992
  - https://trec.nist.gov/

- **NTCIR** (NII Testbeds and Community for Information access Research),
  - Japan, since 1999
  - http://research.nii.ac.jp/ntcir/index-en.html

- **CLEF** (Conference and Labs of the Evaluation Forum)
  - Europe, since 2000
  - http://www.clef-initiative.eu/

- **FIRE** (Forum for Information Retrieval Evaluation)
  - India, since 2008
  - http://fire.irsi.res.in/

EVALUATION METRICS

Aim of an Effective Search Engine

Retrieving as many relevant documents as possible

⬆

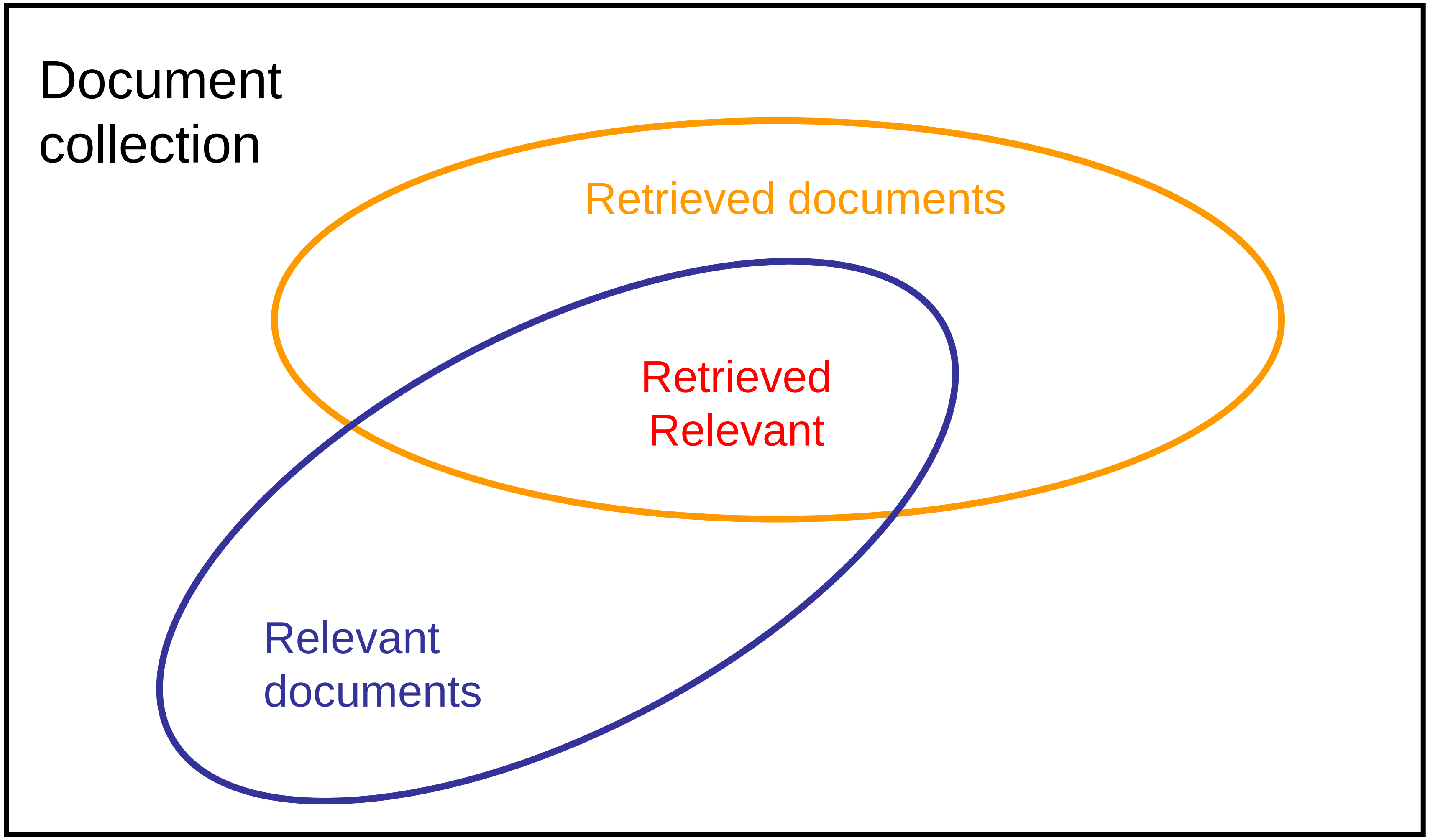
While minimizing the number of non-relevant documents retrieved

⬇

A Taxonomy of Evaluation Measures

	Set-based Measures	Rank-based Measures
Binary Relevance	Precision (P) Recall (R) F-measure (F)	Precision at Document Cut-off (P@k) Recall at Document Cut-off (R@k) R-Precision (Rprec) Average Precision (AP)
Multi-graded Relevance	Not widely agreed generalizations of Precision and Recall	Discounted Cumulated Gain (DCG) …

Relevant vs. Retrieved



Set-based Measures: Precision, Recall, and F-measure

- Together, Precision and Recall measure **retrieval effectiveness**, meant as the ability of a system to retrieve relevant documents while at the same time holding back non-relevant ones
  - Maximizing Precision and Recall corresponds to optimal retrieval in the sense of the Probability Ranking Principle, i.e., ordering documents by their decreasing probability of being relevant

- **F-measure** is the harmonic mean of Precision and Recall, summarizing them into a single score

Definitions of Precision, Recall, and F-measure (1)

- **Precision**:

$$P = \frac{|relevant\ and\ retrieved|}{|retrieved|}$$

- **Recall**:

$$R = \frac{|relevant\ and\ retrieved|}{|relevant|}$$

- **F-measure**:

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2 \cdot \frac{P \cdot R}{P + R}$$

Definitions of Precision, Recall, and F-measure (2)

- By considering another terminology (taken from classification)

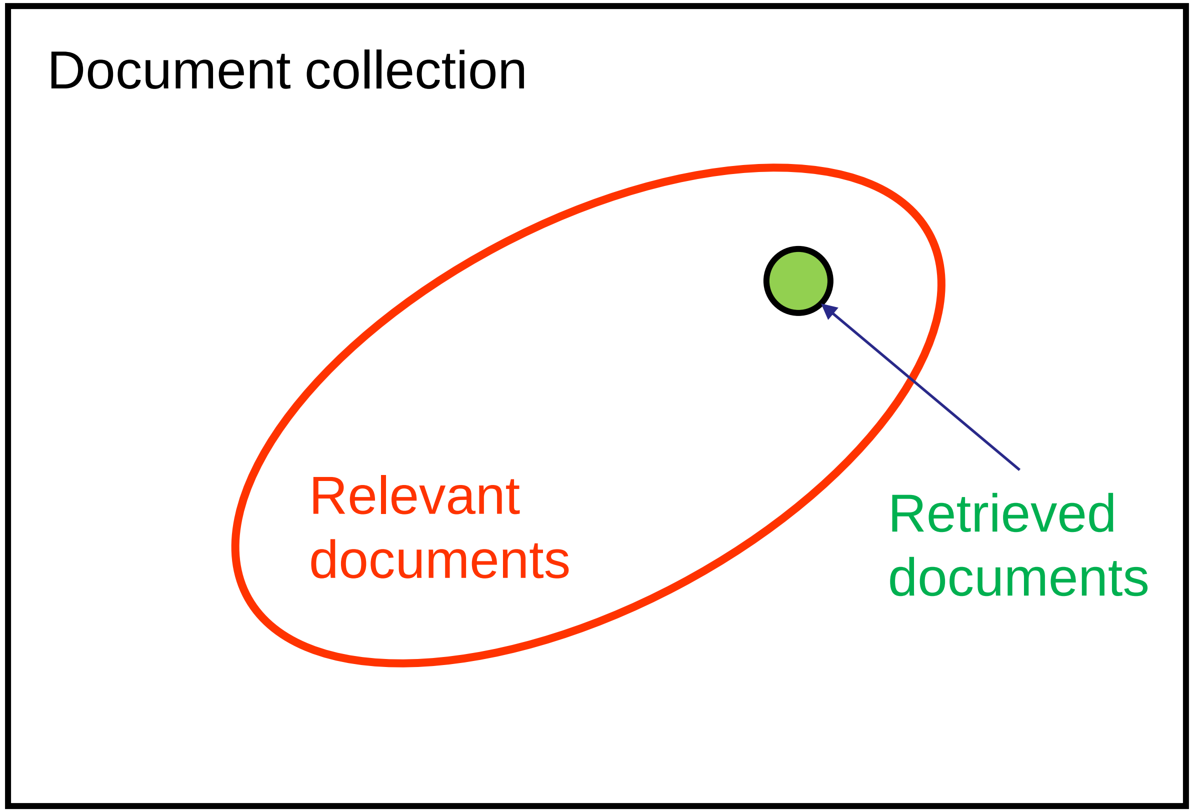
	Relevant	Non-relevant
Retrieved	tp	fp
Not Retrieved	fn	tn

$$P = \frac{tp}{tp + fp} \quad R = \frac{tp}{tp + fn}$$

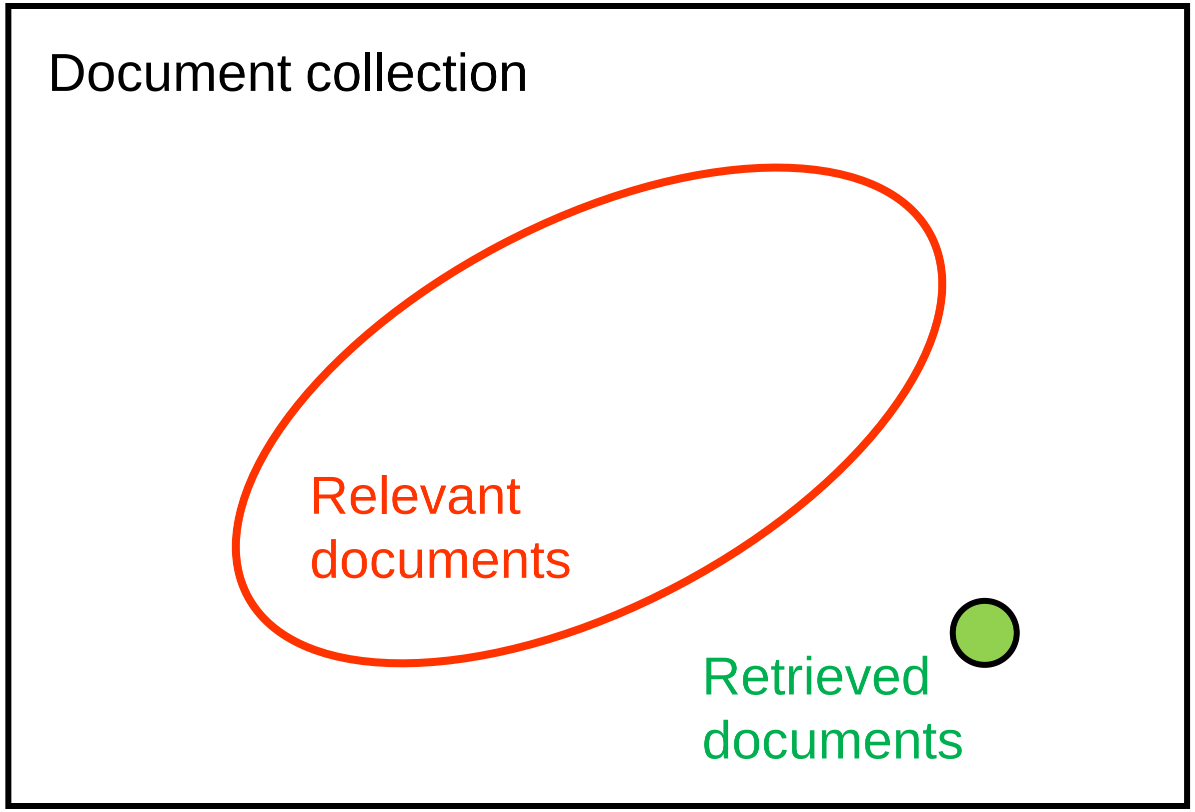
$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2 \cdot \frac{P \cdot R}{P + R}$$

Example: Precision vs. Recall (1)

- Very high precision (1), very low recall

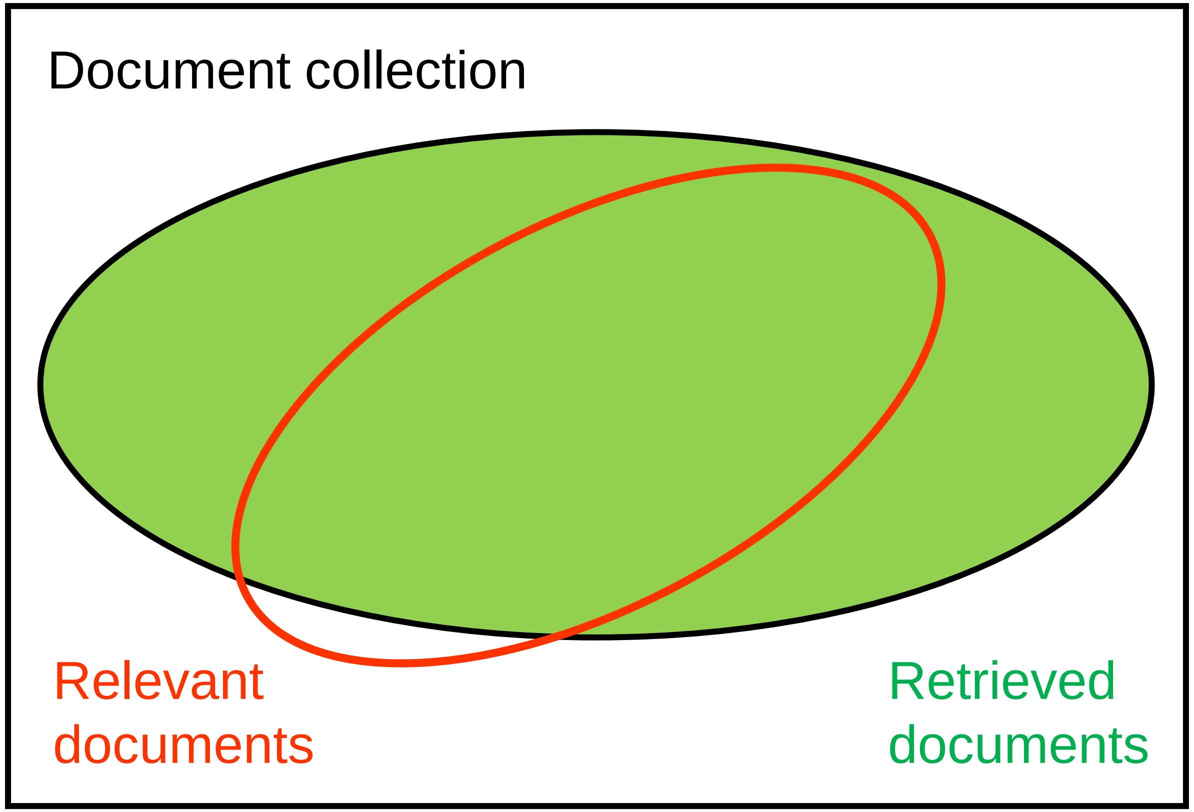


- Very low precision and recall (0)

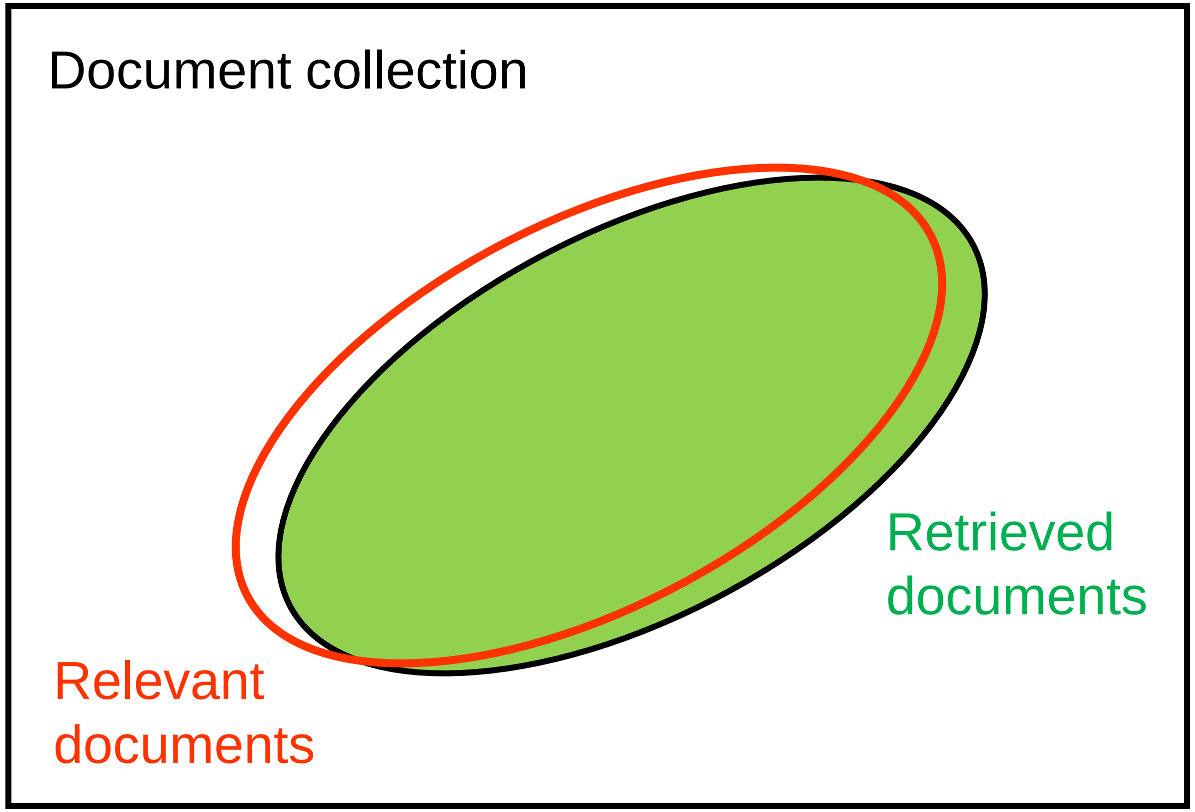


Example: Precision vs. Recall (2)

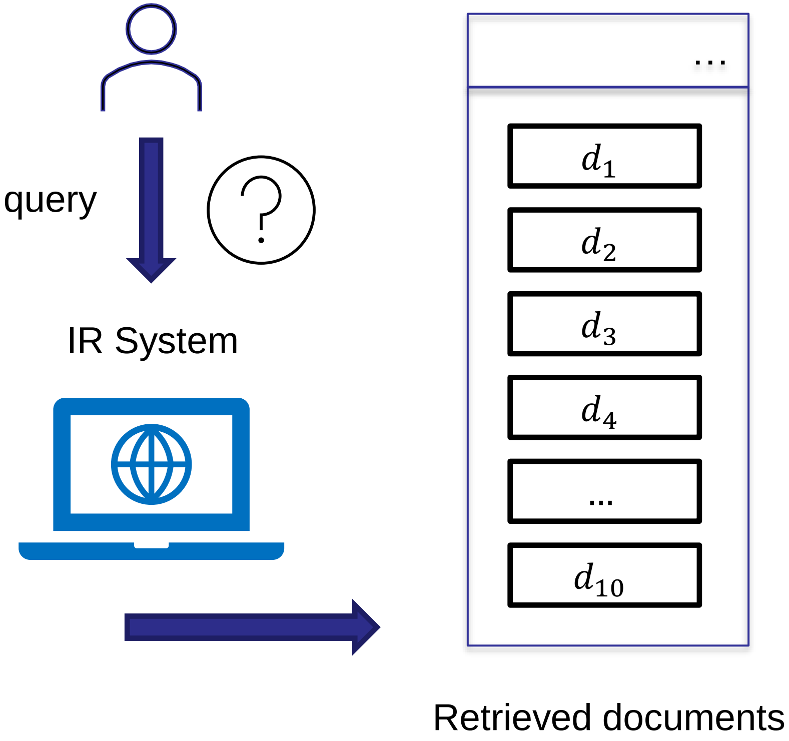
- Low precision, high recall

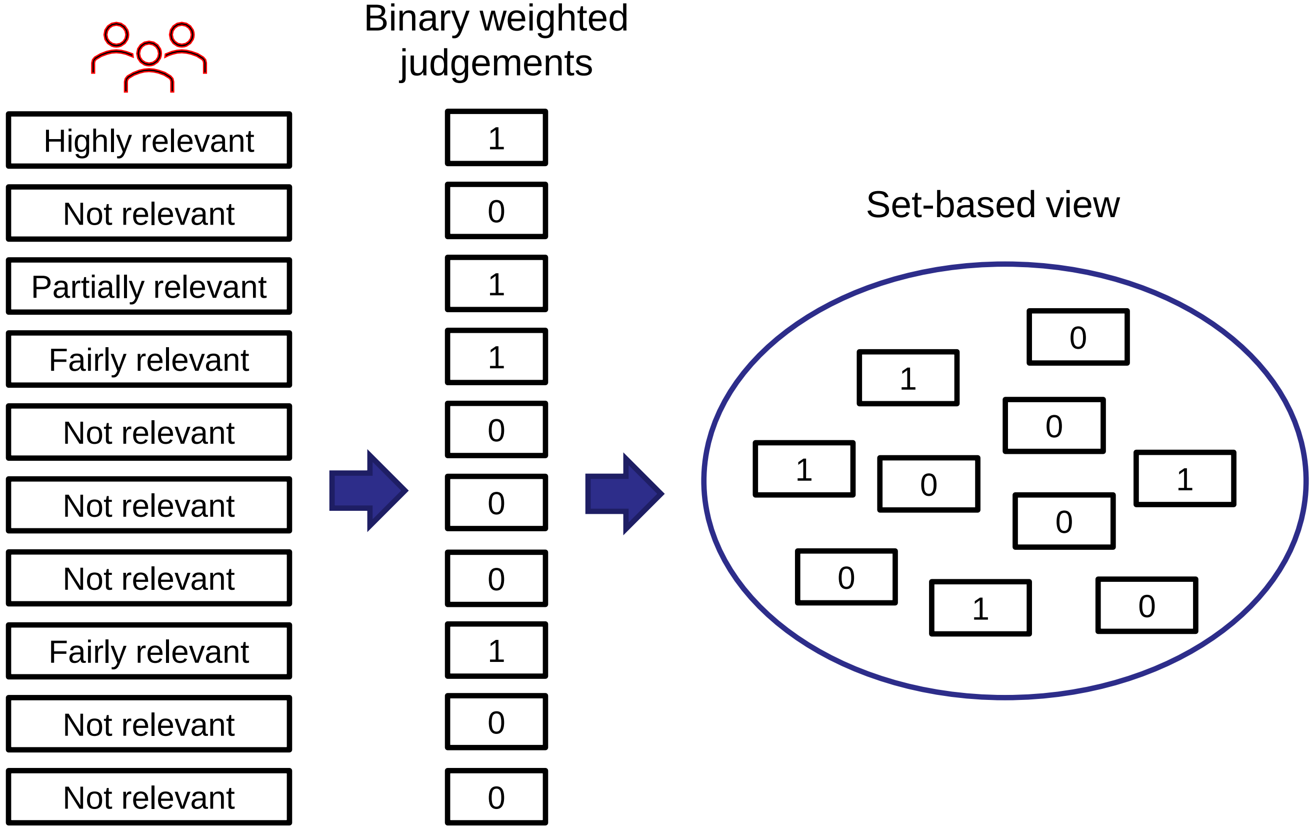


- High precision and recall



Precision, Recall, and F-measure in Action





Relevance judgements	Binary weighted judgements
Highly relevant	1
Not relevant	0
Partially relevant	1
Fairly relevant	1
Not relevant	0
Not relevant	0
Not relevant	0
Fairly relevant	1
Not relevant	0
Not relevant	0

$$P = \frac{|rel. and\ retr.|}{|retr.|} = \frac{4}{10} = 0.40$$

$$R = \frac{|rel. and\ retr.|}{|rel.|} = \frac{4}{8} = 0.50$$

Assuming 8 total relevant documents

$$F = 2 \cdot \frac{\frac{4}{10} \cdot \frac{4}{8}}{\frac{4}{10} + \frac{4}{8}} = \frac{4}{9} = 0.44$$

Issues in computing Precision and Recall

- Precision/Recall are related to each other
  - Combined measures are in some cases more appropriate → F-measure

- A **retrieval batch mode** has been assumed
  - While the interaction with the user can alter the effectiveness of the retrieval
    - Therefore, it would be necessary to quantify the information deriving from the interaction with the user
    - → Interactive Information Retrieval (IIR)

- "True" **Recall** values **can not always be calculated**

Rank-based Measures (1)

- In systems with **millions of documents** (e.g., Web search engines) it can be very difficult to calculate the Recall with respect to a query

- In this case, measures that consider only a **subpart** of the result list are used

- An example is **Precision@k**, i.e., the precision is computed on the sublist constituted by the first k results
  - E.g., Precision@10 is the precision calculated on the first ten results presented to the user

Rank-based Measures (2)

- Measures that consider **the position of the document in the list** of results of a query are:
  - Precision@k ($P@k$): measures the proportion of relevant documents among the first k documents retrieved
  - Recall@k ($R@k$): measures the proportion of relevant documents found in the first k positions of the ranked list of results on the total number of relevant documents in the collection
  - Mean Average Precision (MAP):
    - AP is the average Precision@k calculated on all k positions where a relevant document is found
    - MAP is the average AP on the query set

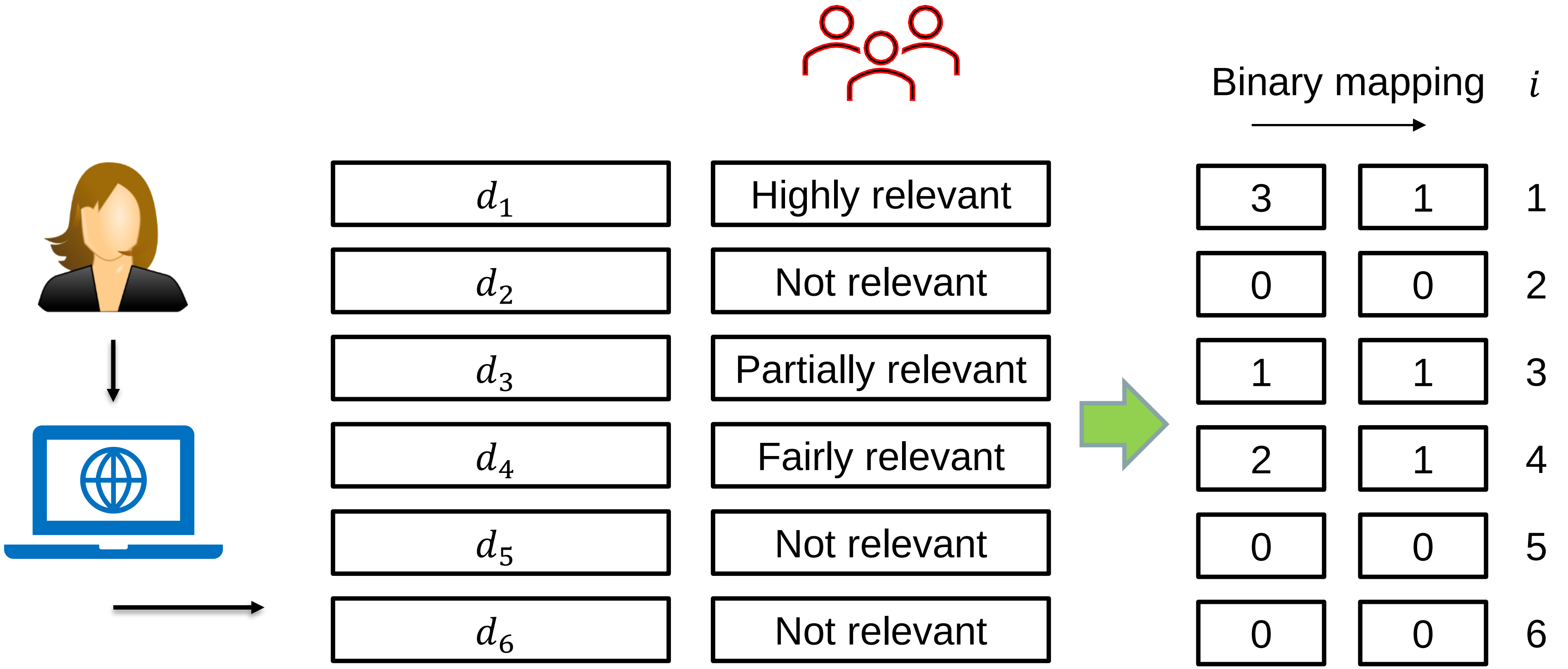
Rank-based Measures: Precision@k

- The user visits exactly ***k* rank positions** and then stops:

$$P@k = \frac{1}{k}\sum_{i=1}^{k} r_i$$

$r_i \in \{0,1\}$ is the relevance judgement of the document at **rank position** i

Example: Precision@k



			Binary mapping	i
d_1	Highly relevant	3	1	1
d_2	Not relevant	0	0	2
d_3	Partially relevant	1	1	3
d_4	Fairly relevant	2	1	4
d_5	Not relevant	0	0	5
d_6	Not relevant	0	0	6

$$P@5 = \frac{1}{5}(1 + 0 + 1 + 1 + 0) = \frac{3}{5} = 0.6$$

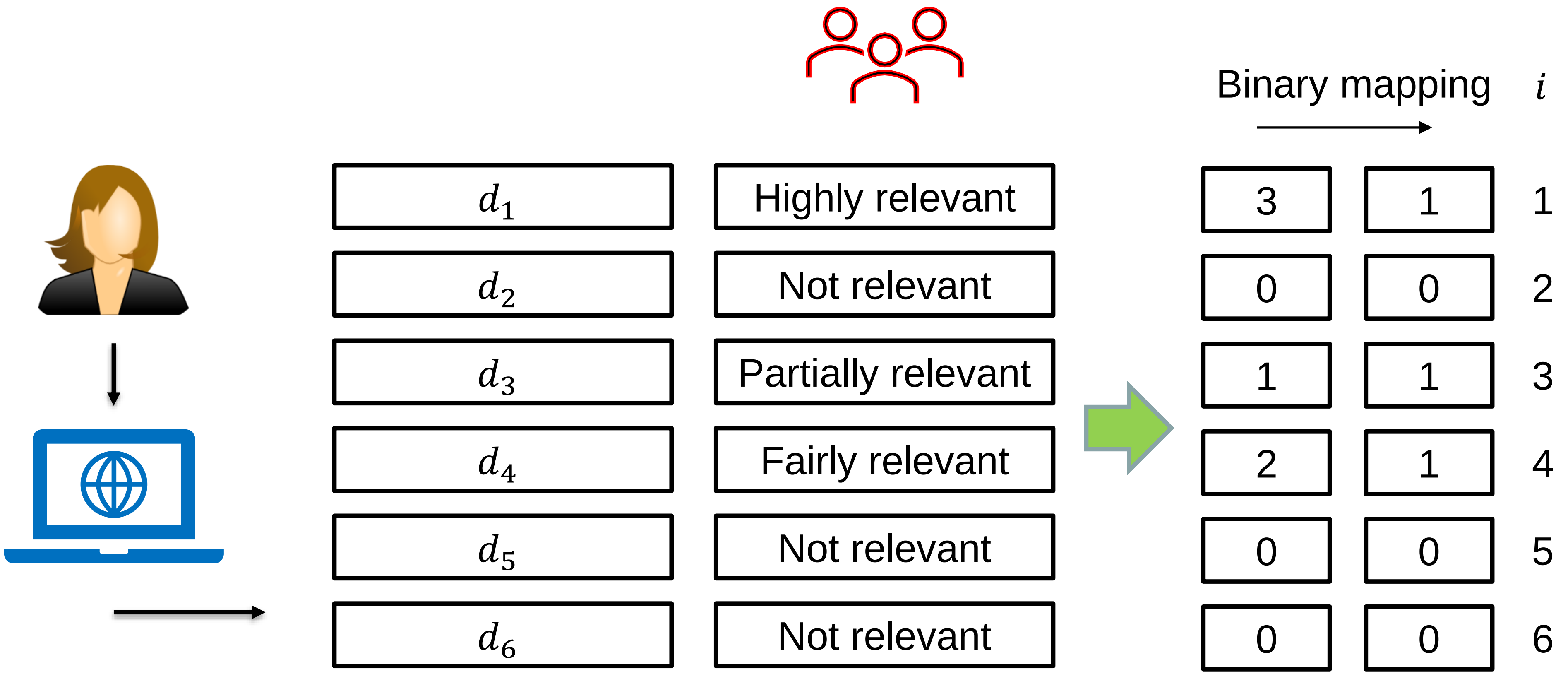
Rank-based Measures: Average Precision

- **Average Precision** (AP): The user randomly selects a relevant document and examines every document down to including that one in the result list:

$$AP = \frac{1}{RB} \sum_{k \in R} P@k$$

$\boldsymbol{RB}$ is the recall-base, i.e., the total number of relevant documents, R is the set of the ranks of the relevant documents

Example: Average Precision



			Binary mapping	i
d_1	Highly relevant	3	1	1
d_2	Not relevant	0	0	2
d_3	Partially relevant	1	1	3
d_4	Fairly relevant	2	1	4
d_5	Not relevant	0	0	5
d_6	Not relevant	0	0	6

$$P@5 = \frac{1}{5}(1+0+1+1+0) = \frac{3}{5} = 0.6$$

$$AP = \frac{1}{10}(P@1 + P@3 + P@4) = \frac{1}{10}\left(1 + \frac{2}{3} + \frac{3}{4}\right) = 0.24$$

Assuming 10 relevant documents in total

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval
MSCA ITN/ETN No. 860721

Rank-based Measures: Mean Average Precision

- The **Mean Average Precision** (MAP) is the mean of AP over a set of topics (each topic is represented by means of a query)
  - Differently from the other measures, this mean has its own name since it is the most widely used single number to summarize the whole performance of an Information Retrieval System

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

Rank-based Measures: Discounted Cumulative Gain (1)

- A widely employed measure to evaluate a search engine's **ability to place in top positions** of the result list those documents that are **highly relevant**

- It is **assumed** that the relevance is not quantified by a binary value but that is expressed in terms of **multi-graded numerical values**

- DCG employs those multi-graded relevance judgements associated to the retrieved documents to evaluate the usefulness, or **gain**, of a document based on its position in the result list
  - Highly relevant documents appearing lower in a search result list should be penalized → Next slides

Rank-based Measures: Discounted Cumulative Gain (2)

- A vector G can be constructed in which position k indicates the **relevance judgement** of the document in position k in the result list

$$G[k] = r_k$$

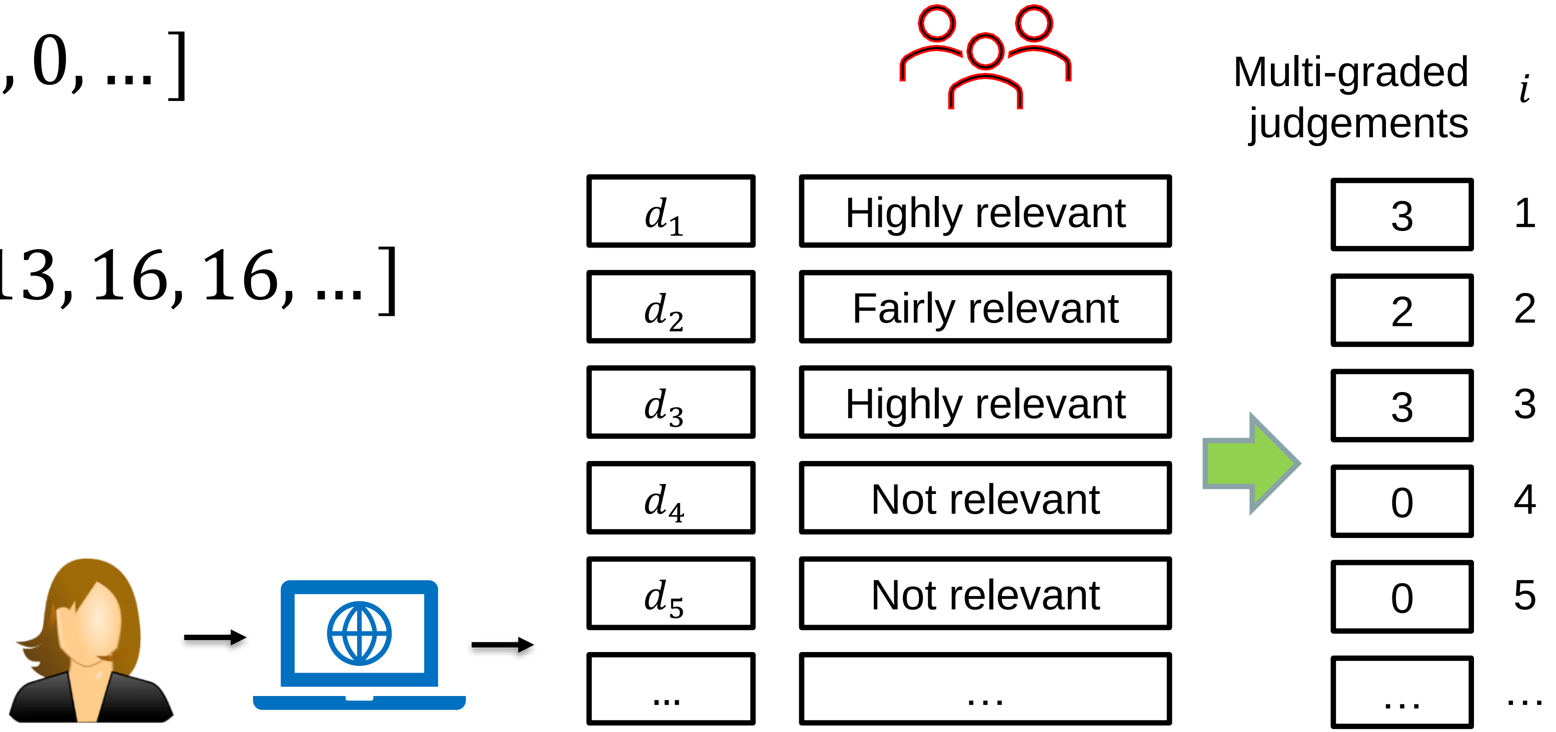
- At this point, it is possible to construct the CG (**Cumulative Gain**) vector with non-decreasing values in the following way:

$$CG[k] = \sum_{i=1}^{k} r_i$$

Cumulative Gain is the simple **sum of relevance judgments**

Example: Cumulative Gain

- It is assumed to have a set of documents judged according to a scale with 4 values (0-3) of relevance.
- $G = [3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots]$
- $CG = [3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \dots]$



Retrieved documents		Multi-graded judgements	i
d_1	Highly relevant	3	1
d_2	Fairly relevant	2	2
d_3	Highly relevant	3	3
d_4	Not relevant	0	4
d_5	Not relevant	0	5
…	…	…	…

Rank-based Measures: Discounted Cumulative Gain (3)

- The **greater the rank**, the **less useful** the document for the user

- It is therefore necessary a **function** that progressively reduces the value of documents as the rank increases
  - For example, dividing the value of the document by the logarithm of its rank
    - With the choice of the base of the logarithm, one can choose what weight to associate with the position of the documents

- **Characteristics**:
  - DCG naturally handles multi-graded relevance
  - DCG does not depend on the Recall Base (RB)
  - DCG is not bounded in [0, 1]

Rank-based Measures: Discounted Cumulative Gain (4)

- At this point we can define the DCG (**Discounted Cumulative Gain**) vector as follows:

$$DCG[k] = \sum_{i=1}^{k} \frac{r_i}{\max(1, \log_b i)}$$

where the base of the **logarithm** indicates the patience of the user in scanning the result list ($b = 2$ is an impatient user, $b = 10$ is a patient user)

- The ability of a search engine to classify very relevant documents in the first positions of the result list is thus represented by the values in the CG and DCG vectors

Example: Discounted Cumulative Gain

- It is assumed to have a set of documents judged according to a scale with 4 values (0-3) of relevance.

- $G = [3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \ldots]$

- $CG = [3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \ldots]$

- $DCG = [3, 5, \boxed{6.89,} 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, \ldots]$ $DCG[k] = \sum_{i=1}^{k} \frac{r_i}{\max(1, \log_b i)}$

- $DCG(3) = \frac{3}{\max(1, \log_2(1))} + \frac{2}{\max(1, \log_2(2))} + \frac{3}{\max(1, \log_2(3))} = \boxed{6.89}$

Rank-based Measures: nDCG

- The **normalized Discounted Cumulative Gain** is defined as the DCG at rank k, i.e., $DCG(k)$, normalized by the value of DCG at the rank k calculated on the ideal list of results, i.e., $iDCG(k)$

- The **ideal list** is the one that presents the most relevant document in the first position, the second most relevant in second position, etc.

$$nDCG(k) = \frac{DCG(k)}{iDCG(k)}$$

- nDCG is a widely used measure in the evaluation of search engines on the Web

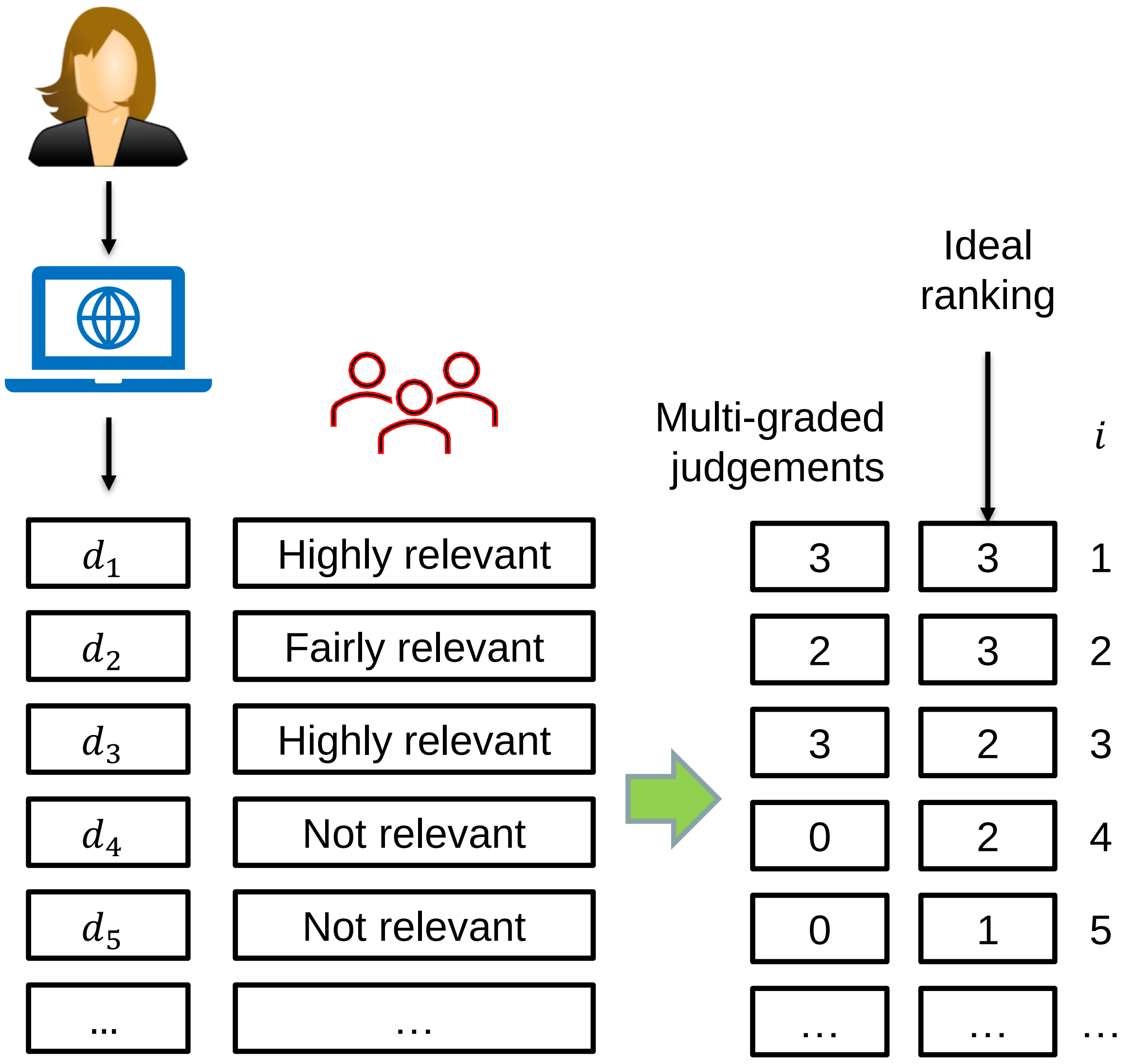
Example: nDCG

- $G = [3, 2, 3, 0, 0, \dots]$
- $iG = [3, 3, 2, 2, 1, \dots]$

- $CG = [3, 5, 8, 8, 8, \dots]$
- $iCG = [3, 6, 8, 10, 11, \dots]$

- $DCG = [3, 5, 6.89, 6.89, 6.89, \dots]$
- $iDCG = [3, 6, 10.82, 12.26, 12.88, \dots]$

- $nDCG(5) = \frac{DCG(5)}{iDCG(5)} = \frac{6.89}{12.88} = 0.535$



Retrieved documents		Multi-graded judgements	Ideal ranking	i
d_1	Highly relevant	3	3	1
d_2	Fairly relevant	2	3	2
d_3	Highly relevant	3	2	3
d_4	Not relevant	0	2	4
d_5	Not relevant	0	1	5
...	…	…	…	…

EVALUATIONS IN
INTERACTIVE INFORMATION RETRIEVAL

Users' Interactions with the System

- The **incorporation of users** into IR system evaluation and the study of users' information search behaviors and interactions have been identified as **important concerns for IR researchers**

- The study of IR systems has a prescribed and dominant evaluation method that can be traced back to the **Cranfield Paradigm**

- Studies of users and their interactions with information systems find their place in **Interactive Information Retrieval** (IIR), where users are typically studied along with their interactions with systems and information

The IIR Evaluation Model

- The model builds on **three basic components**:
  - The involvement of potential users as test participants
  - The application of dynamic and individual information needs (real, and simulated information needs)
  - The employment of multidimensional and dynamic relevance judgements

- The aim of the IIR evaluation model is to facilitate IIR evaluation as close as possible to actual information searching and IR processes, though still in a **relatively controlled evaluation environment**
  - → **User Studies**

An ongoing Research (1)

- Difficult to evaluate **system actions** beyond ranking and **user actions** beyond clicking
  - Expected Search Length and RF (relevance feedback) measures
  - TREC interactive track
  - TREC session track
  - NDCG and variations
  - User behavior models and simulations
  - User studies and crowdsourcing

Bruce Croft. *The Importance of Interaction in Information Retrieval*. SIGIR 2019 https://www.sigir.org/sigir2019/slides/10.1145-3331184.3331185.pdf

An ongoing Research (2)

- Chengxiang Zhai. **Interactive Information Retrieval: Models, Algorithms, and Evaluation**
  - Tutorial @ SIGIR 2020 - https://sigir.org/sigir2020/tutorials/
    - Challenges in IIR Evaluation
    - Simulation-Based Evaluation
    - Formal Models for User Simulation
    - Other Strategies of IIR Evaluation

ANCILLARY INFORMATION

Evaluation Initiatives in Personalized Search (UNIMIB)

- **WEPIR:** Workshop on Evaluation of Personalization in Information Retrieval (ongoing) – SIGIR/CHIIR
  - Nicholas J. Belkin, Rutgers University, USA
  - Gareth J. F. Jones, ADAPT Centre, Dublin City University, Ireland
  - Noriko Kando, National Institute of Informatics, Tokyo
  - Gabriella Pasi, University of Milano-Bicocca, Italy

- **PIR-CLEF:** Evaluation of Personalized Information Retrieval (to be resumed) – CLEF
  - Gabriella Pasi, University of Milano-Bicocca (DISCo), Milan, Italy
  - Gareth J. F. Jones, ADAPT Centre, Dublin City University, Ireland

- **CLEF eHealth Evaluation Lab:** Evaluation Challenge in the Medical and Biomedical Domain (ongoing) – CLEF
  - Many people are/have been involved in CLEF eHealth over the years

Reference Material

- This presentation is partly based on the **reworking of the following material**:

- **Nicola Ferro**. ***Foundations of IR Evaluation***. 12th European Summer School in Information Retrieval (ESSIR 2019) 15-19 July 2019, Milan, Italy
  https://github.com/ir-laboratory/essir2019/blob/master/ferro-essir2019.pdf

- **Hinrich Schütze, Christopher D. Manning**, and **Prabhakar Raghavan**. ***Introduction to information retrieval***. Vol. 39. Cambridge: Cambridge University Press, 2008.

- **Diane Kelly**. ***Methods for evaluating Interactive Information Retrieval systems with users***. Foundations and trends in Information Retrieval 3.1—2 (2009): 1-224.

- **Colleen Cool**, and **Nicholas J. Belkin**. ***Interactive information retrieval: history and background*** (2011): 1-14. https://pdfs.semanticscholar.org/6d34/db18dfea72d00c2b6a7050efc22961228290.pdf

- **Pia Borlund**. ***Interactive Information Retrieval: An introduction***. Journal of Information Science Theory and Practice 1.3 (2013): 12-32.

- **Bruce W. Croft**. ***The Importance of Interaction for Information Retrieval***. SIGIR. Vol. 19. 2019.