

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

دورة "استرجاع المعلومات" باللغة العربية - صيف ٢٠٢١

Information Retrieval – Summer 2021



### 3. Evaluation

Tamer Elsayed  
Qatar University

# Announcements



- Today's Lab is open to all!
  - At 6:45pm
  - Use same lecture link
- Lecture 1 is on YouTube!

# Today's Roadmap

- Evaluation in IR: why and how?
  - Set-based measures
  - Rank-based measures
  - Building test collections
- 
- *Evaluation at Large Search Engines*



# IR as an Experimental Science!

- Formulate a research question: the hypothesis
- Design an experiment to answer the question
- Perform the experiment
  - Compare with a baseline “control”
- Does the experiment answer the question?
  - Are the results significant? Or is it just luck?
- Report the results!
- Repeat...

# Questions About the Black Box

- Example “question”: Does expanding the query with synonyms improve retrieval performance?
- Corresponding experiment?
  - Expand queries with synonyms and compare against baseline unexpanded queries

# Questions That Involve Users

- Example “question”: Is letting users weight search terms better?
- Corresponding experiment?
  - Build two different interfaces, one with term weighting functionality, and one without; run a user study

# Types of Evaluation Strategies

## ○ System-centered studies

- Given documents, queries, and relevance judgments
- Try several variations of the system
- Measure which system returns the “best” hit list
- Laboratory experiment

## ○ User-centered studies

- Given several users, and at least two retrieval systems
- Have each user try the same task on both systems
- Measure which system works the “best”

# Why is Evaluation Important?

- The ability to **measure differences** underlies experimental science
  - How well do our systems work?
  - Is A better than B?
  - Is it really?
  - Under what conditions?
- Evaluation **drives what to research**
  - Identify techniques that work and don't work

# Evaluation Criteria

## ○ Effectiveness

- How “good” are the documents that are returned?
- System only, human + system

## ○ Efficiency

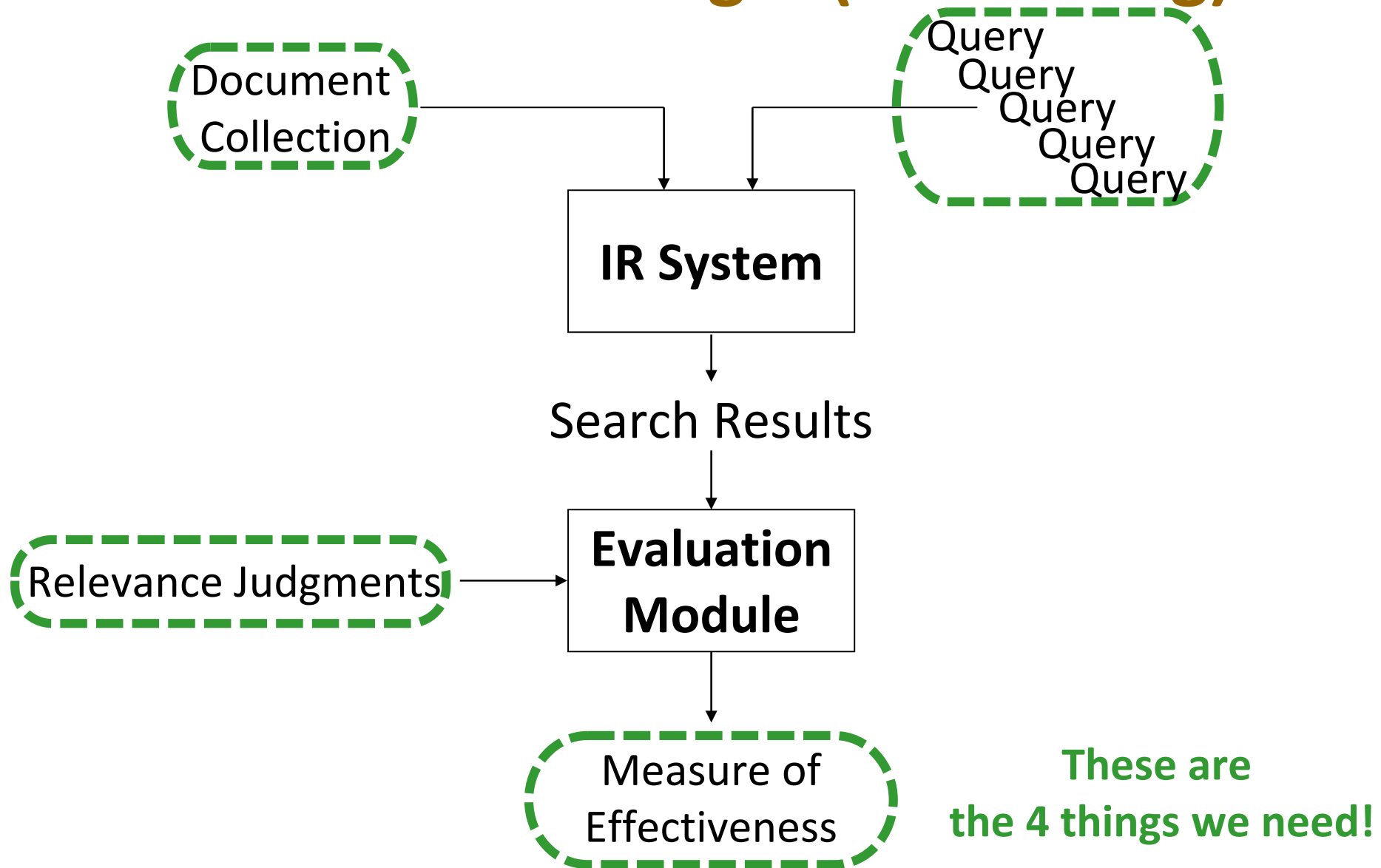
- Retrieval time (query latency/response time), indexing time, index size

## ○ Usability

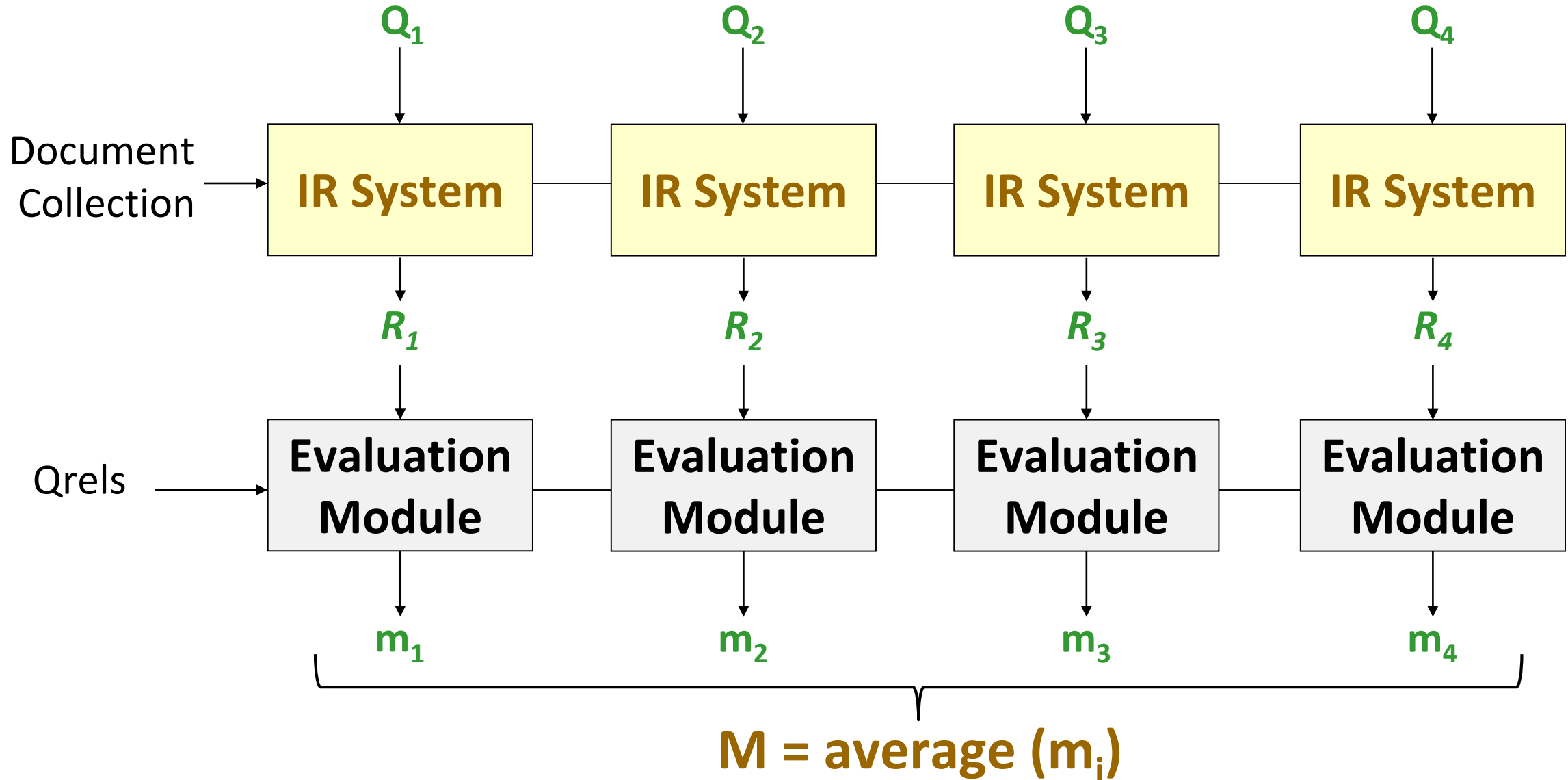
- Learnability, flexibility
- Novice vs. expert users



# Cranfield Paradigm (Lab setting)



# Cranfield Paradigm (Lab setting)



# Reusable Test Collections

## 1. Collection of documents

- Should be “representative”
- Things to consider: size, sources, genre, topics, ...

## 2. Sample of information needs

- Should be “randomized” and “representative”
- Usually formalized **topic** statements

## 3. Known relevance judgments

- Assessed by humans, for each topic-document pair (topic, not query!)
- Binary judgments make evaluation easier, but could be graded

# Good Effectiveness Measures

- Should capture some aspect of **what the user wants**
  - That is, the measure should be **meaningful**
- Should be **easily replicated** by other researchers
- Should be **easily comparable**
  - Optimally, expressed as a single number





10



## A test collection has 3 components (choose 3):

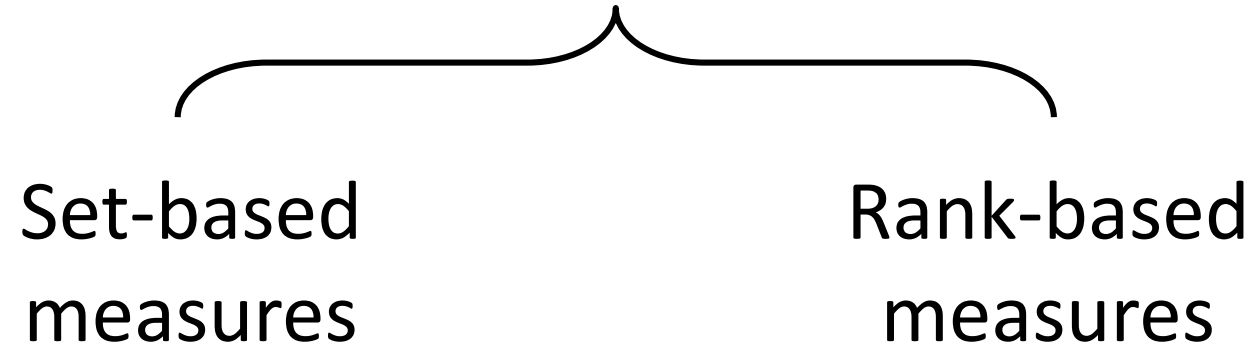
- Set of information needs (topics)
- Collection of documents
- Set of queries for one information need
- Set of evaluation measures
- Relevance judgments
- Set of IR systems

## Relevance judgments indicate ...

- how good the IR system is.
- which documents are relevant to which topics.
- which topics are good for evaluating the IR systems.



# Effectiveness Evaluation Measures







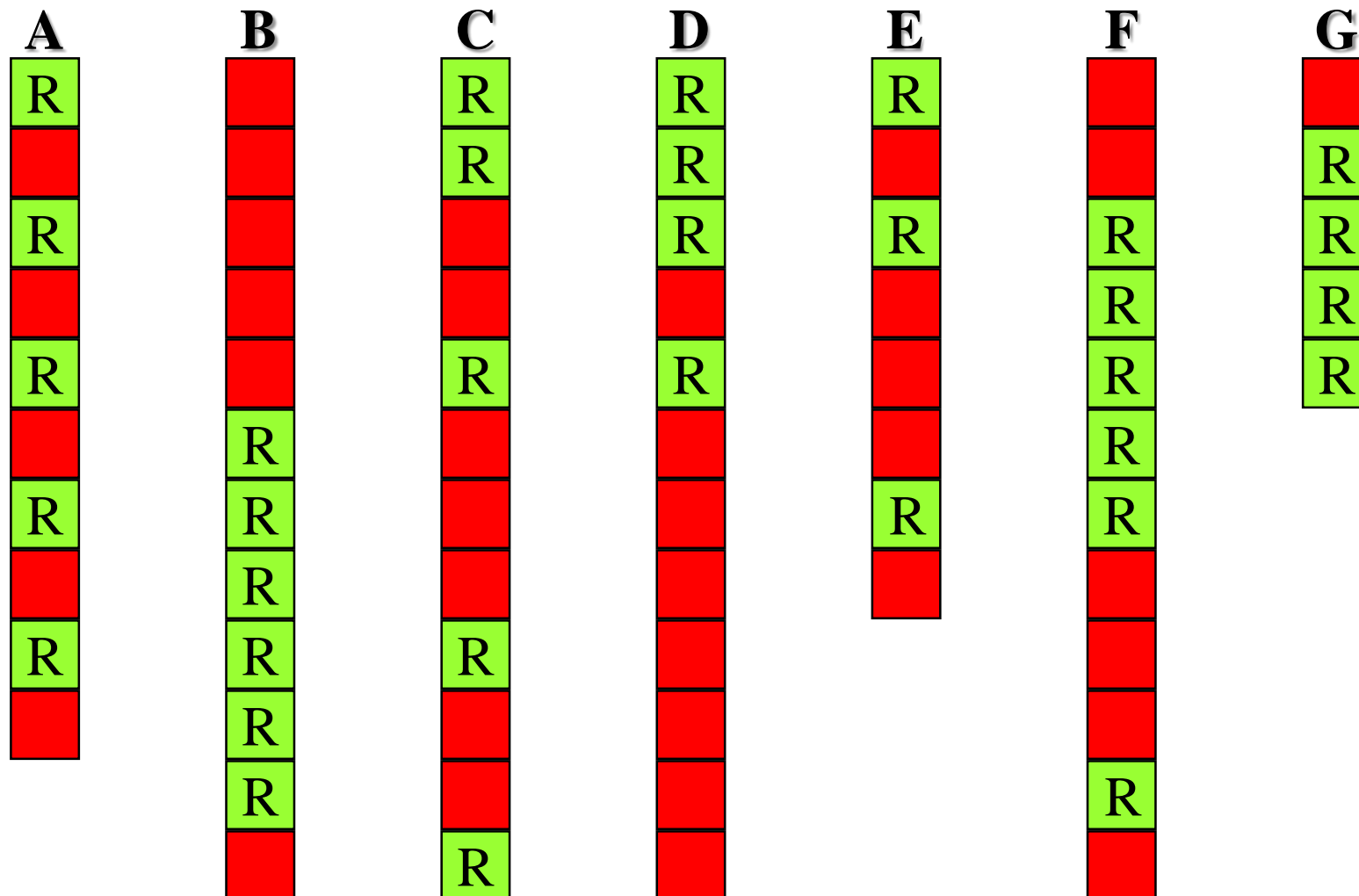
## SET-BASED MEASURES

# Set-Based Measures

- Assuming IR system returns set of retrieved results without ranking.
- No certain number of results per query
- Suitable for Boolean Search

# Which is the Best Set?

For **query Q**, collection has 9 relevant documents and systems A-G *retrieved* following results:

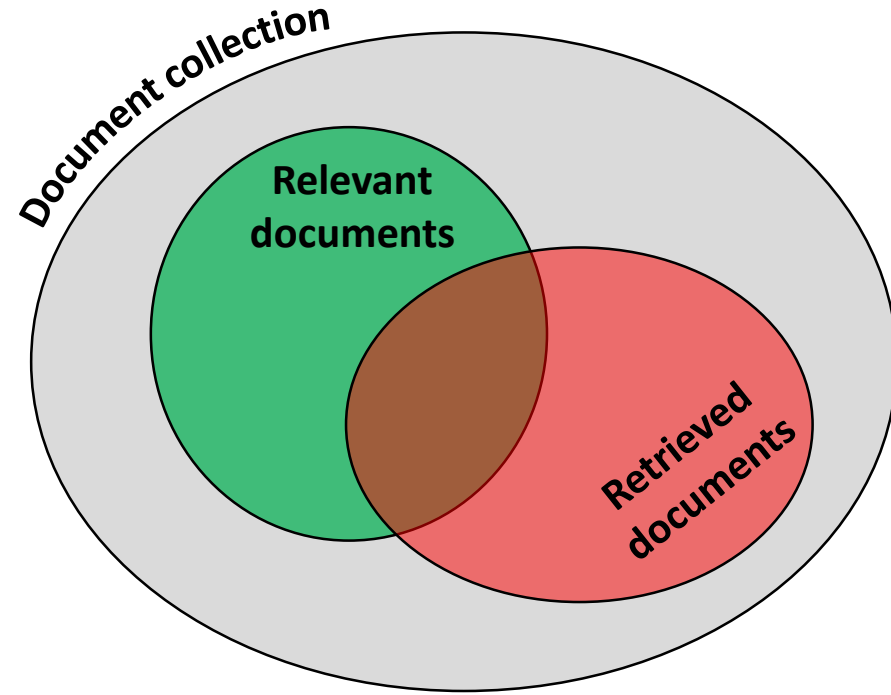


# Precision and Recall

## ○ Precision:

*What fraction of these retrieved docs are relevant?*

$$P = \frac{rel \cap ret}{retrieved} = \frac{TP}{TP + FP}$$



## ○ Recall:

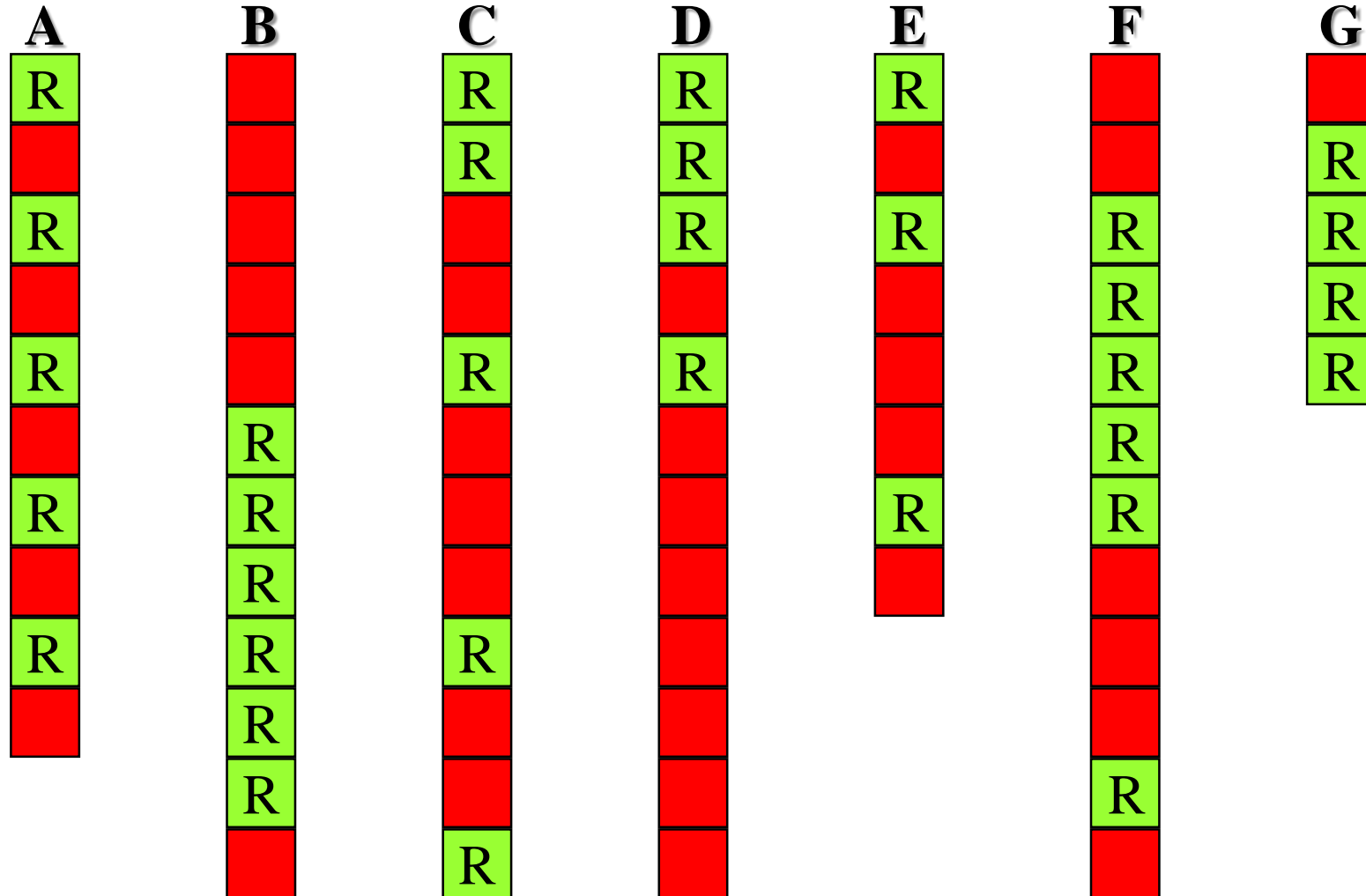
*What fraction of the relevant docs were retrieved?*

$$R = \frac{rel \cap ret}{relevant} = \frac{TP}{TP + FN}$$

relevant irrelevant	FP	TN
	TP	FN
	retrieved	not retrieved

# Precision & Recall?

For **query Q**, collection has 9 relevant documents and systems A-G *retrieved* following results:





# Precision & Recall?

For **query Q**, collection has 9 relevant documents and systems A-G *retrieved* following results:

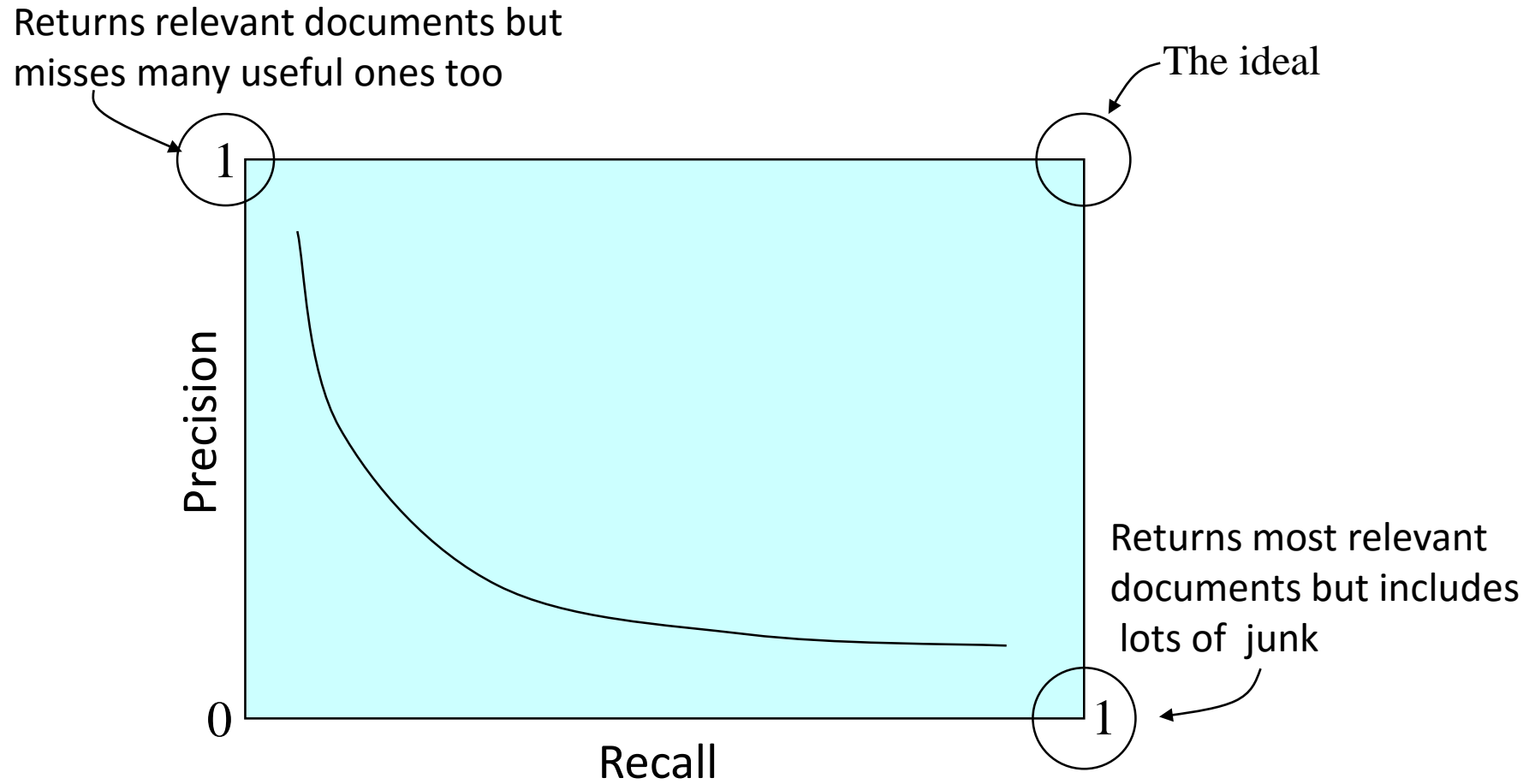
E	G
R	
	R
R	R
	R
	R
R	

# Trade-off between P & R

- Precision: The ability to retrieve top-ranked documents that are mostly relevant.
- Recall: The ability to retrieve ***all*** of the relevant items in the corpus.
- Retrieve more docs:
  - Higher chance to find all relevant docs → R ↑↑
  - Higher chance to find more irrelevant docs → P ↓↓



# Trade-off between P & R



# One Measure? F-Measure

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

- Harmonic mean of recall and precision

- emphasizes the importance of small values, whereas the arithmetic mean is affected more by outliers that are unusually large.

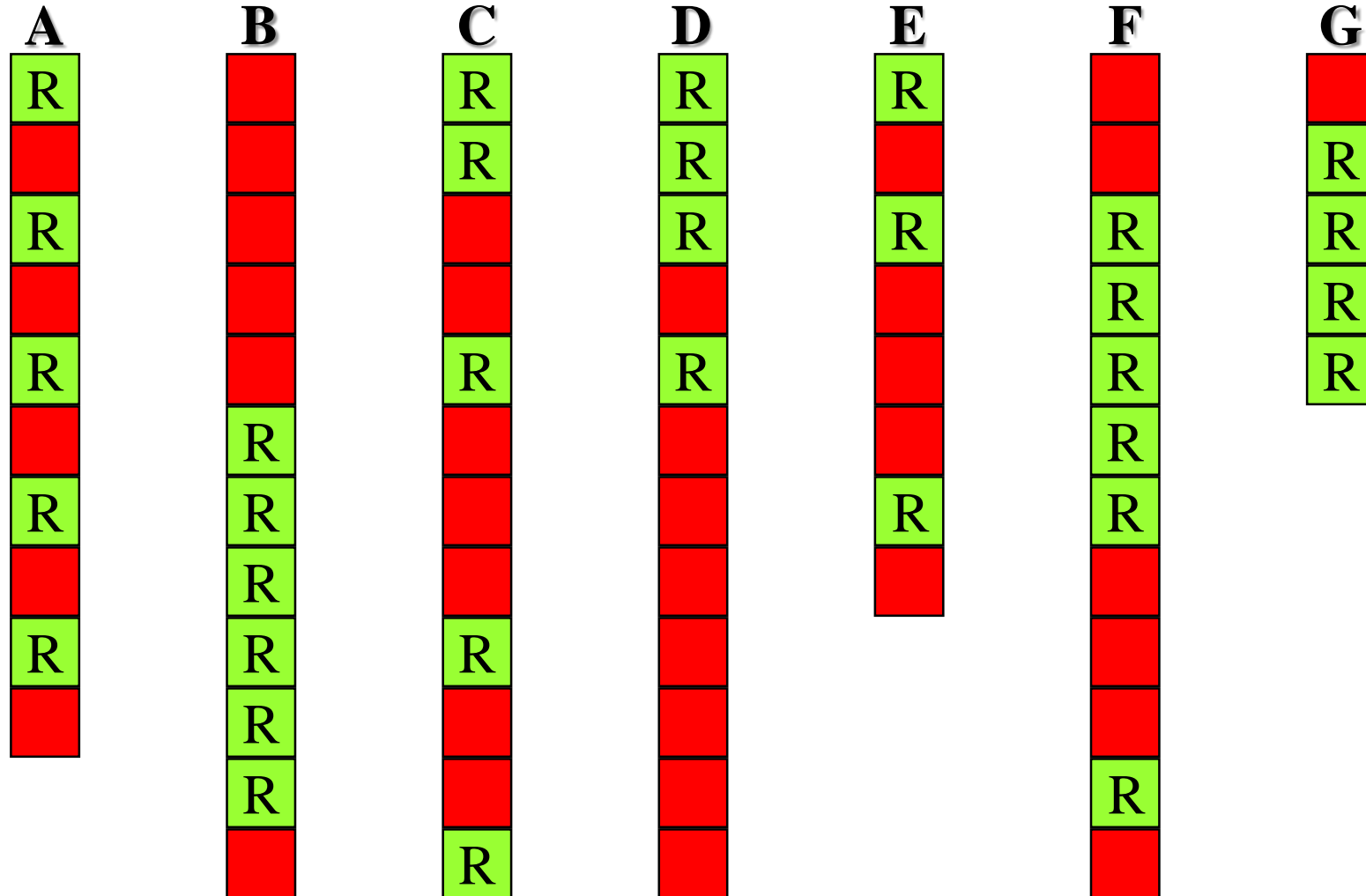
$$F_\beta = \frac{(\beta^2 + 1)P \cdot R}{\beta^2 P + R}$$

- Beta controls relative importance of precision and recall

- Beta = 1, precision and recall equally important →  $F_1$
- Beta = 5, recall is 5 times more important than precision

# F1?

For **query Q**, collection has 9 relevant documents and systems A-G *retrieved* following results:



# Today's Roadmap

- Evaluation in IR: why and how?
  - Set-based measures
  - Rank-based measures
  - Building test collections
- 
- *Evaluation at Large Search Engines*





## RANK-BASED MEASURES

# Rank-based IR measures

- Consider systems A & B
  - Both retrieved 10 docs, only 5 are relevant
  - P, R & F are the same for both systems
  - Should their performance considered equal?
- Ranked IR requires taking “ranks” into consideration!

	A	B
1		R
2		R
3	R	R
4		R
5	R	R
6		
7	R	
8		
9	R	
10	R	

**How to do that?**

# Which is the Best Ranked List?

For **query Q**, collection has 9 relevant documents and systems A-G produced following results:

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>
1	R		R	R	R		
2			R	R			R
3	R			R	R	R	R
4						R	R
5	R		R	R		R	R
6		R				R	
7	R	R			R	R	
8		R					
9	R	R	R				
10		R					
11		R				R	
12			R				

# Precision @ K

- $k$  (a fixed number of documents)
- Have a cut-off on the ranked list at rank  $k$ , then calculate precision!
- Perhaps appropriate for most of web search: all people want are good matches on the first one or two results pages.



# Precision @ 5?

For **query Q**, collection has 9 relevant documents and systems A-G produced following results:

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>
1	R		R	R	R		
2			R	R			R
3	R			R	R	R	R
4						R	R
5	R		R	R		R	R
6		R				R	
7	R	R			R	R	
8		R					
9	R	R	R				
10		R					
11		R				R	
12			R				

# When a user can stop?

- P@k assumes every user will stop inspecting results at rank k.

***Is that realistic?***

- IR objective: “satisfy user’s information need”
- Assumption: a user will stop once his/her information need is satisfied.

***Where?***

# When a user can stop?

- User will keep looking for relevant docs in the ranked list, read them, then stop once he/she feels satisfied → user will stop at a relevant document

***Which relevant document  
the user will stop at?***

- What about calculating the averages over all potential stops?
  - Every time you find relevant doc at rank  $x$ , calculate  $P@x$ , then take the average at the end.

# Average Precision (AP)

$Q_1$   
(has 5 rel. docs)

1	R	$1/1=1.00$
2	R	$2/2=1.00$
3		
4		
5	R	$3/5=0.60$
6		
7		
8		
9	R	$4/9=0.44$
10		

---

$$\text{AP} = 3.04 / 5$$
$$= \mathbf{0.608}$$

$Q_2$   
(has 3 rel. docs)

1		
2		
3	R	$1/3=0.33$
4		
5		
6		
7	R	$2/7=0.29$
8		
...		$\frac{3}{\infty}=0$

---

$$\text{AP} = 0.62 / 3$$
$$= \mathbf{0.207}$$

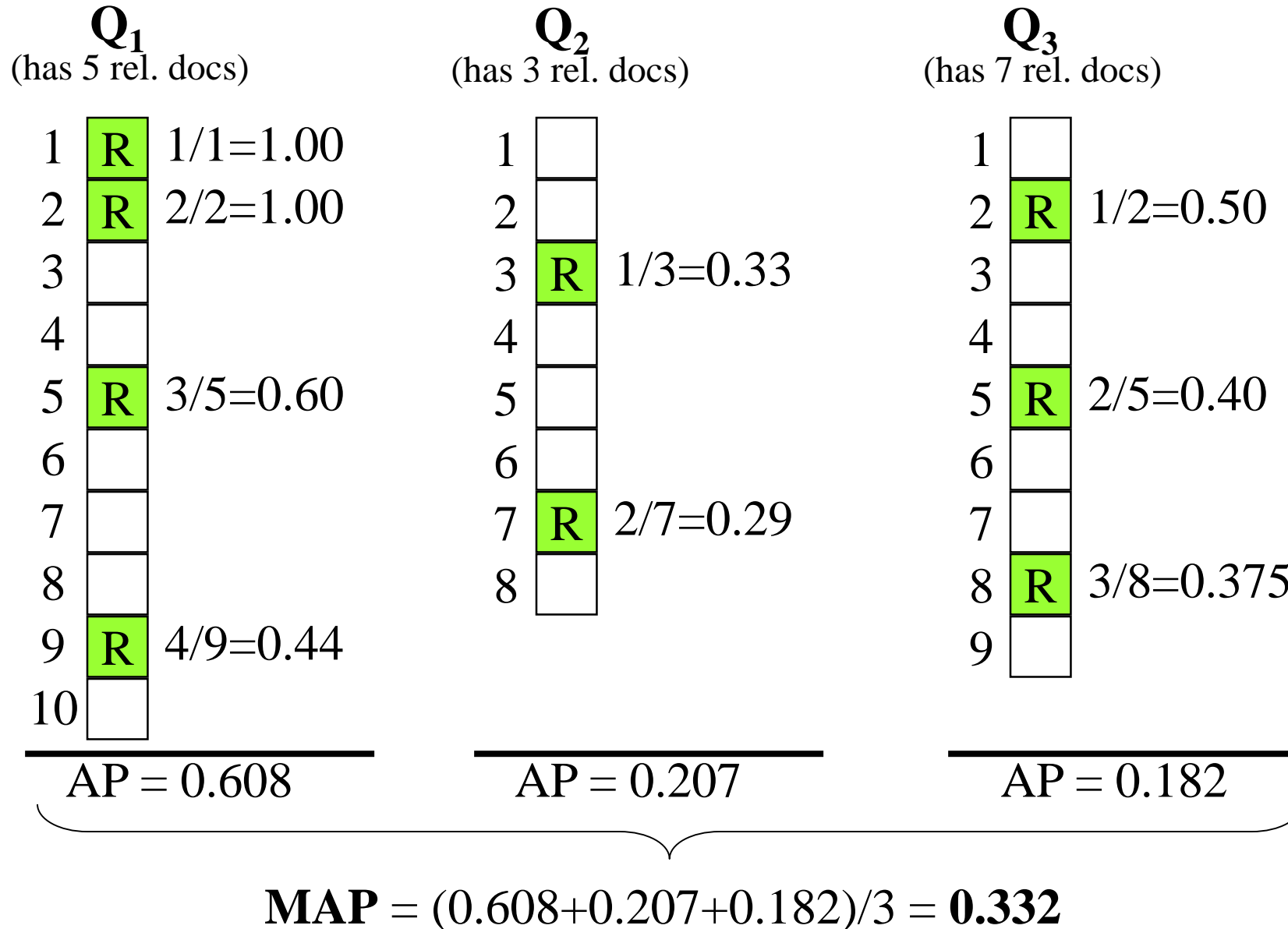
$Q_3$   
(has 7 rel. docs)

1		
2	R	$1/2=0.50$
3		
4		
5	R	$2/5=0.40$
6		
7		
8	R	$3/8=0.375$
9		
...		

---

$$\text{AP} = 1.275 / 7$$
$$= \mathbf{0.182}$$

# Mean Average Precision (MAP)



# AP/MAP

- A mix between precision and recall.
- Highly focus on finding relevant documents as early as possible.
- MAP is the most commonly-used evaluation metric for most IR search tasks.
- When we have only 1 relevant doc for every topic
  - ➔ MAP = MRR (mean reciprocal rank)
- Uses binary relevance:  $rel = 0$  or  $1$

# Binary vs. Graded Relevance

- Some docs are more relevant to a topic than other relevant ones!
  - We need non-binary relevance.
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant.
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined.
- Discounted Cumulative Gain (DCG)
  - Uses graded relevance as a measure of the usefulness
  - The most popular for evaluating web search


# Binary vs. Graded Relevance

- Some docs are more relevant to a topic than other relevant ones!
  - We need non-binary relevance.
- 1. Higher relevance docs should have higher *value* (**gain**) in evaluation.
- 2. But this value will **decay** (be **discounted**) if it appears lower in the ranked list, since it is less likely to be examined.



# Discounted Cumulative Gain (DCG)

- Gain is accumulated starting at the top of the ranking and may be reduced, or **discounted**, at lower ranks
- Users care more about high-ranked documents, so we **discount** results by  $1/\log_2(rank)$ 
  - the discount at rank 4 is  $1/2$ , and at rank 8 is  $1/3$
- $DCG_k$  is the total gain accumulated at a particular rank  $k$  (sum of DG up to rank  $k$ ):

$$DCG_k = rel_1 + \sum_{i=2}^k \frac{rel_i}{\log_2(i)}$$


0, 1, 2, 3, ...  
(graded)

# DCG Example

k	G
1	3
2	2
3	3
4	0
5	0
6	1
7	2
8	2
9	3
10	0

# DCG Example

k	G	DG
1	3	3
2	2	2
3	3	1.89
4	0	0
5	0	0
6	1	0.39
7	2	0.71
8	2	0.67
9	3	0.95
10	0	0

# DCG Example

k	G	DG	DCG@10
1	3	3	
2	2	2	
3	3	1.89	
4	0	0	
5	0	0	
6	1	0.39	
7	2	0.71	
8	2	0.67	
9	3	0.95	
10	0	0	9.61

***DCG can be any positive real number!***

***Why is that a problem?***

# Normalized DCG (nDCG)

- DCG values are often *normalized* by comparing the DCG at each rank with the DCG value for the *perfect ranking*
  - makes averaging easier for queries with different numbers of relevant documents
- $\text{NDCG}@k = \text{DCG}@k / \text{iDCG}@k$  (divide actual by ideal)
  - $\text{nDCG} \leq 1$  at any rank position
  - *ideal ranking* has **nDCG** of 1.0

# nDCG Example

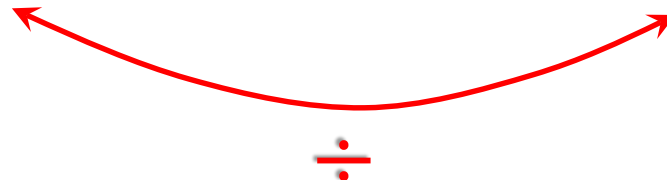
k	G	DG	DCG@10	iG
1	3	3		3
2	2	2		3
3	3	1.89		3
4	0	0		2
5	0	0		2
6	1	0.39		2
7	2	0.71		1
8	2	0.67		0
9	3	0.95		0
10	0	0	9.61	0

# nDCG Example

k	G	DG	DCG@10	iG	iDG
1	3	3		3	3
2	2	2		3	3
3	3	1.89		3	1.89
4	0	0		2	1.00
5	0	0		2	0.86
6	1	0.39		2	0.77
7	2	0.71		1	0.36
8	2	0.67		0	0.00
9	3	0.95		0	0.00
10	0	0	9.61	0	0.00

# nDCG Example

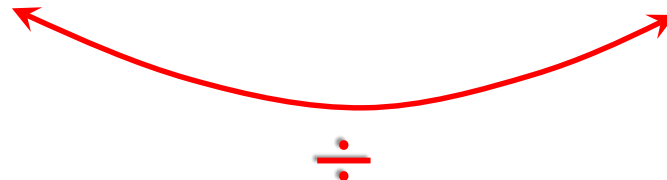
k	G	DG	DCG@10	iG	iDG	iDCG@10
1	3	3		3	3	
2	2	2		3	3	
3	3	1.89		3	1.89	
4	0	0		2	1.00	
5	0	0		2	0.86	
6	1	0.39		2	0.77	
7	2	0.71		1	0.36	
8	2	0.67		0	0.00	
9	3	0.95		0	0.00	
10	0	0	9.61	0	0.00	10.88





# nDCG Example

k	G	DG	DCG@10	iG	iDG	iDCG@10	nDCG@k
1	3	3		3	3		
2	2	2		3	3		
3	3	1.89		3	1.89		
4	0	0		2	1.00		
5	0	0		2	0.86		
6	1	0.39		2	0.77		
7	2	0.71		1	0.36		
8	2	0.67		0	0.00		
9	3	0.95		0	0.00		
10	0	0	9.61	0	0.00	10.88	0.88







# BUILDING TEST COLLECTIONS

# Reusable Test Collections

- Document Collection
- Topics (sample of information needs)
- Relevance judgments (qrels)

# How can we get it?

- For web search, companies apply their own studies to assess the performance of their search engine.
- Web-search performance is monitored by:
  - Traffic
  - User clicks and session logs
  - Labelling results for selected users' queries
- Academia (or lab settings):
  - Someone goes out and builds them (expensive)
  - As a byproduct of large scale evaluations (collaborative effort)
- IR **Evaluation Campaigns** are created for this reason

# IR Evaluation Campaigns

- IR test collections are provided for scientific communities to develop better IR methods.
- Collections and queries are provided, relevance judgements are built during the campaign.
- TREC = Text REtrieval Conference <http://trec.nist.gov/>
  - Main IR evaluation campaign, sponsored by NIST (US gov).
  - Series of annual evaluations, started in 1992.
- Other evaluation campaigns
  - CLEF: European version (since 2000)
  - NTCIR: Asian version (since 1999)
  - FIRE: Indian version (since 2008)

# TREC Tracks and Tasks

- TREC (and other campaigns) are formed of a set of **tracks**, each track is about (one or more) search **tasks**.
  - Each track/task is about searching a set of documents of given genre and domain.
- Examples
  - TREC Web track
  - TREC Medical track
  - TREC Legal track → CLEF-IP track → NTCIR patent mining track
  - TREC Microblog track
    - Adhoc search task
    - Filtering task

# TREC Collection

- A set of hundreds of thousands or millions of docs
  - 1B in case of web search (TREC ClueWeb09)
- The typical format of a document:

```
<DOC>
<DOCNO> 1234 </DOCNO>
<TEXT>
    This is the document.
    Multilines of plain text.
</TEXT>
</DOC>
```



# TREC Topic

- **Topic: a statement of information need**
- Multiple topics (~50) developed (mostly) **at NIST** for a collection.
- Developed by experts and associated with additional details.
  - Title: the query text
  - Description: description of what is meant by the query.
  - Narrative: what should be considered relevant.

`<num>189</num>`

`<title>Health and Computer Terminals</title>`

`<desc>Is it hazardous to the health of individuals to work with computer terminals on a daily basis?</desc>`

`<narr>Relevant documents would contain any information that expands on any physical disorder/problems that may be associated with the daily working with computer terminals. Such things as carpel tunnel, cataracts, and fatigue have been said to be associated, but how widespread are these or other problems and what is being done to alleviate any health problems</narr>`

# Relevance Judgments

- For each topic, set of relevant docs is required to be known for an effective evaluation!
- **Exhaustive assessment** is usually impractical
  - TREC usually has 50 topics
  - Collection usually has >1 million documents
- **Random sampling** won't work
  - If relevant docs are rare, none may be found!
- **IR systems can help** focus the sample (**Pooling**)
  - Each system finds some relevant documents
  - Different systems find different relevant documents
  - Together, enough systems will find most of them

# Pooled Assessment Methodology

1. Systems submit top **1000** documents per topic
  2. Top **100** documents from each are *manually* judged
    - Single pool, duplicates removed, arbitrary order
    - Judged by the person who developed the topic
  3. Treat unevaluated documents as **not** relevant
  4. Compute MAP (or others) down to **1000** documents
- To make pooling work:
- Good number of participating systems
  - Systems must do reasonably well
  - Systems must be different (not all “do the same thing”)





*How can we “rank” search results?*



# EVALUATION AT LARGE SEARCH ENGINES

# Evaluation at Large Search Engines

- Recall is difficult to measure on the web – why?
- Search engines often use
  - precision at top  $k$ , e.g.,  $k = 10$
  - measures that reward you more for getting rank 1 right than for getting rank 10 right (nDCG).
  - 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 (online testing)

- 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.