Search Engine Evaluation

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

Tecnologías de Gestión de Información No Estructurada

Prof. Dr. David E. Losada

USC Universidade de Santiago de Compostela

CAMPUS VIDA

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

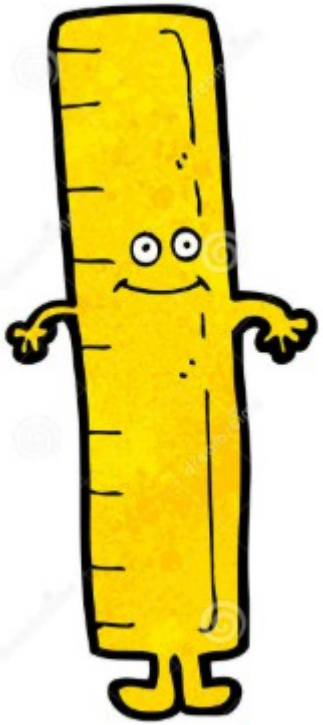
SE evaluation

to figure out **which retrieval method works the best** (advancing our knowledge)

to assess the actual **utility** of an overall TR system

- measures must reflect the utility to the **actual users** in the real application (as opposed to measures on each individual retrieval result)
- Typically done via **user studies**—where human users interact with the corpus via the system

what to measure?



effectiveness or accuracy: how accurate are the search results?

measures the system's capability of ranking relevant documents on top of non-relevant ones

efficiency: how quickly can a user get the results? how large are the computing resources that are needed to answer a query?

measures the space and time overhead of the system

usability: how useful is the system for real user tasks? interfaces and other elements are also important

typically via user studies

Cranfield Evaluation Methodology (1960s)

to build **reusable test collections** and define measures using these collections:

1) the assembled test **collection of documents** is similar to a real document collection in a search application

2) a sample set of **queries or topics** that simulate the user's information need

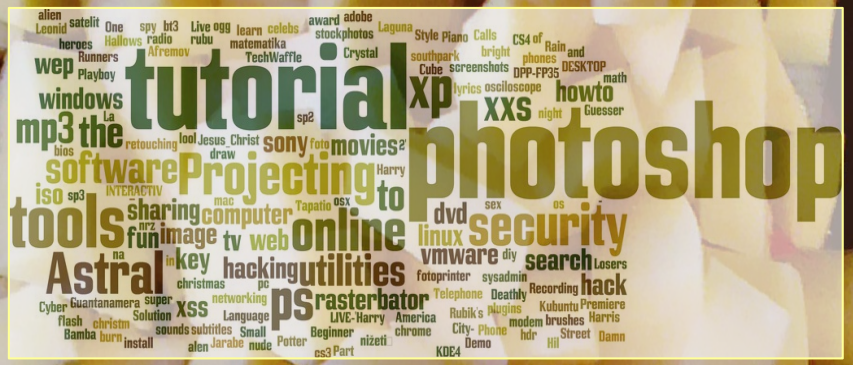
3) **relevance judgments** (which documents should be returned for which queries).

ideally, made by users who formulated the queries (because those are the people that know exactly what the documents would be used for)

information retrieval evaluation:
3 main ingredients

information retrieval evaluation:
3 main ingredients

queries

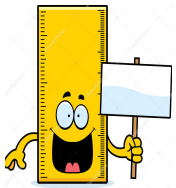


information retrieval evaluation:
3 main ingredients

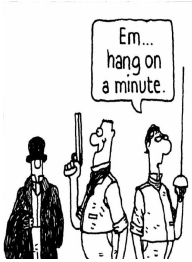
relevance judgements

!

Cranfield Evaluation Methodology



need **measures** to quantify how well a system's result matches the ideal ranked list that would be constructed based on users' relevance judgements



fair comparison: the evaluation is exactly the same for each algorithm (same criteria, same corpus, and same relevance judgements)

the test set can be **reused** many times: once such a test collection is built, it can be used again and again to test different algorithms or ideas

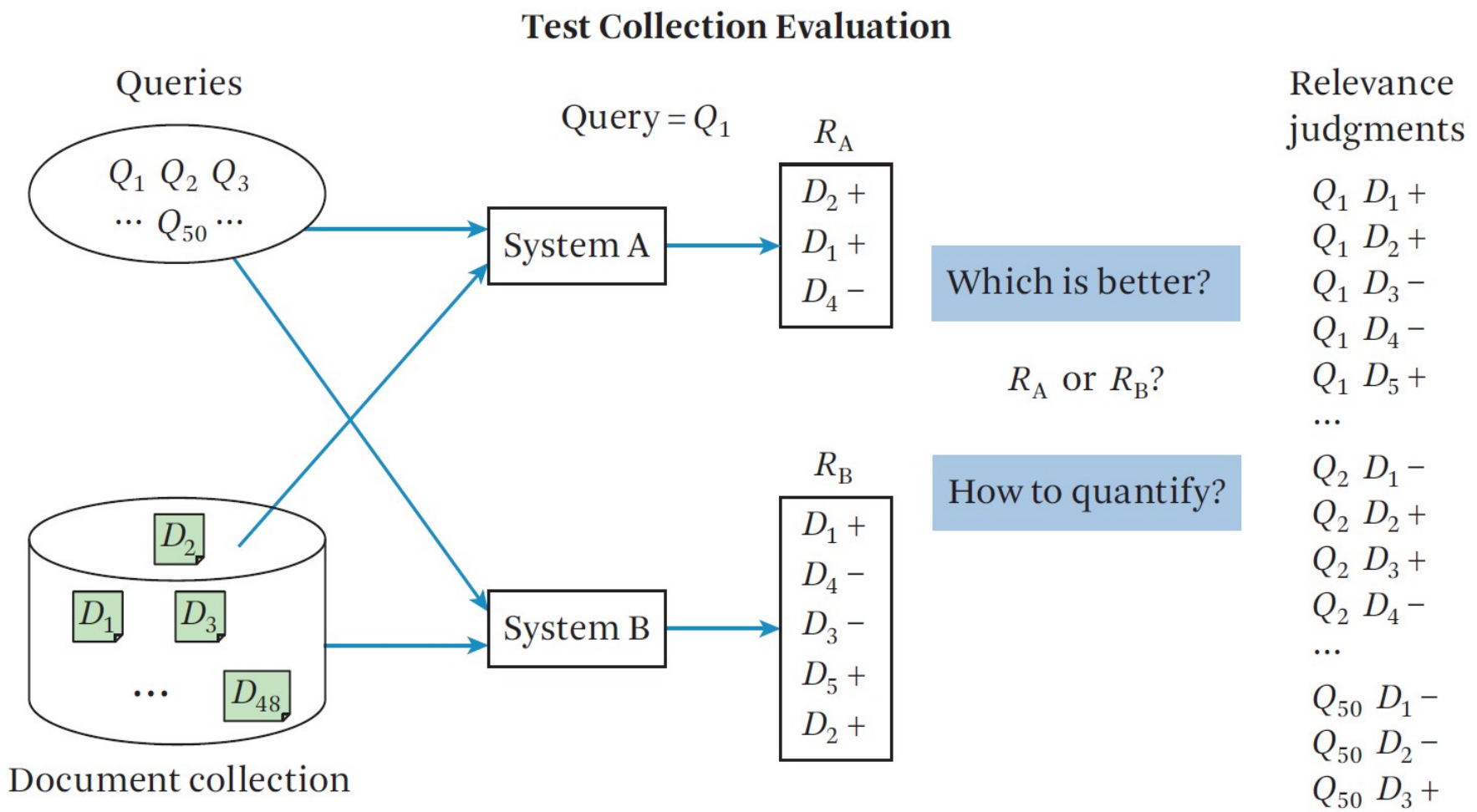


Figure 9.1 Illustration of Cranfield evaluation methodology.

system A is precise but system B returned more relevant docs

which one is better? depends on the user's task and on the individual users!

(e.g. prior art in patent search vs web search)

set retrieval evaluation

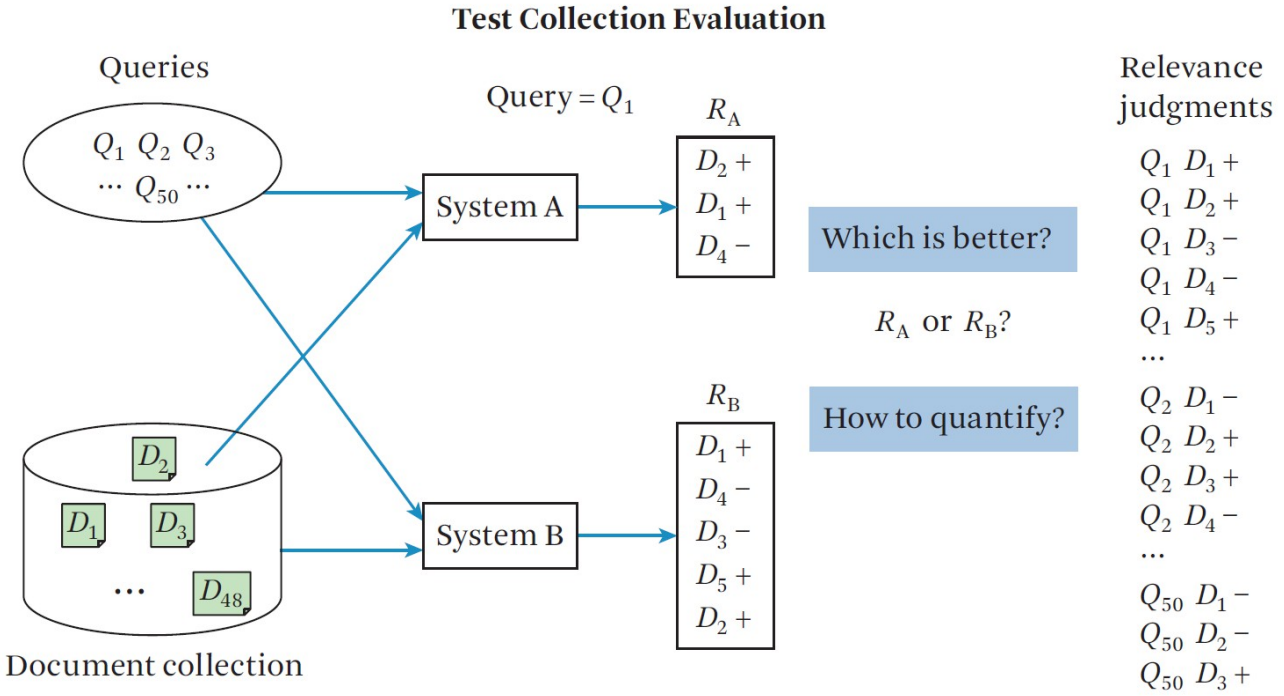


Figure 9.1 Illustration of Cranfield evaluation methodology.

precision: simply computes to what extent all the retrieved results are relevant

system A has a precision of 2/3

system B has a precision of 3/5

set retrieval evaluation

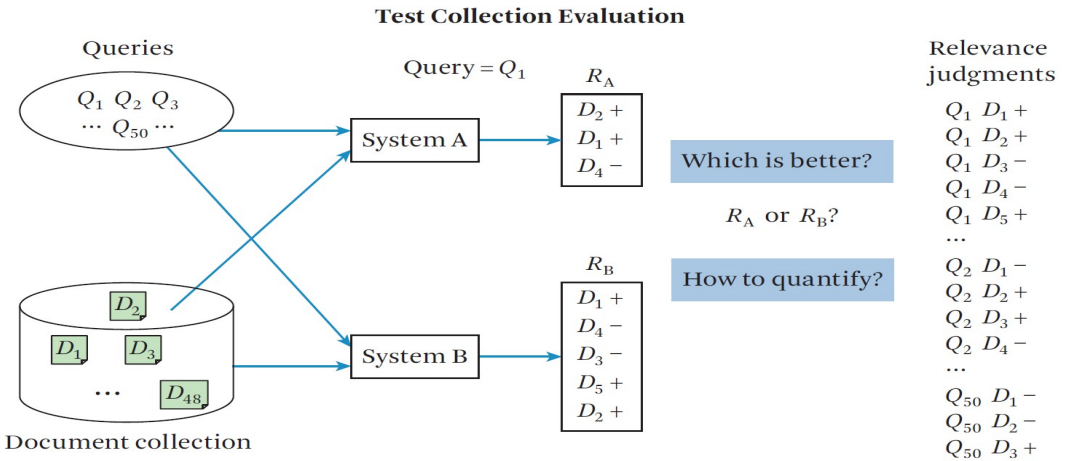


Figure 9.1 Illustration of Cranfield evaluation methodology.

recall: measures the completeness of coverage of relevant documents in your retrieval result (compares the number of total relevant documents to the number that is actually retrieved)

Suppose we have a total of ten relevant documents in the corpus for Q1

system A has a recall of 2/10, system B has a precision of 3/10

set retrieval evaluation

		Action: Retrieved	Action: Not retrieved
Doc	Relevant	a	b
	Not relevant	c	d

$$\text{Precision} = \frac{a}{a+c}$$

Ideal results: precision = recall = 1.0

$$\text{Recall} = \frac{a}{a+b}$$

In reality, high recall tends to be associated with low precision

Set can be defined by a cutoff (e.g., precision @ 10 docs)

we often are interested in the precision up to ten documents for web search

F Measure: Combining Precision and Recall

there tends to be a tradeoff between precision and recall, so it is natural to combine them…

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

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

$$\text{where } P = \text{precision}, R = \text{recall}, \beta = \text{parameter (often set to 1)}$$

β controls the emphasis between precision and recall

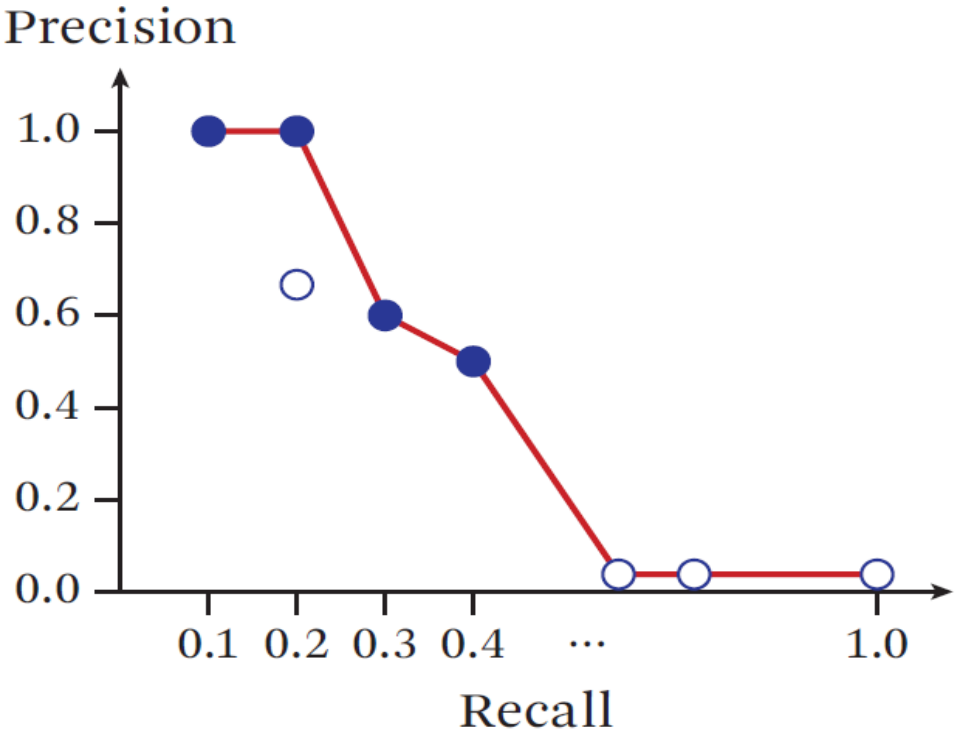
when β = 1 : harmonic mean of precision and recall

evaluation of a ranked list

Evaluating Ranking: Precision–Recall (PR) Curve

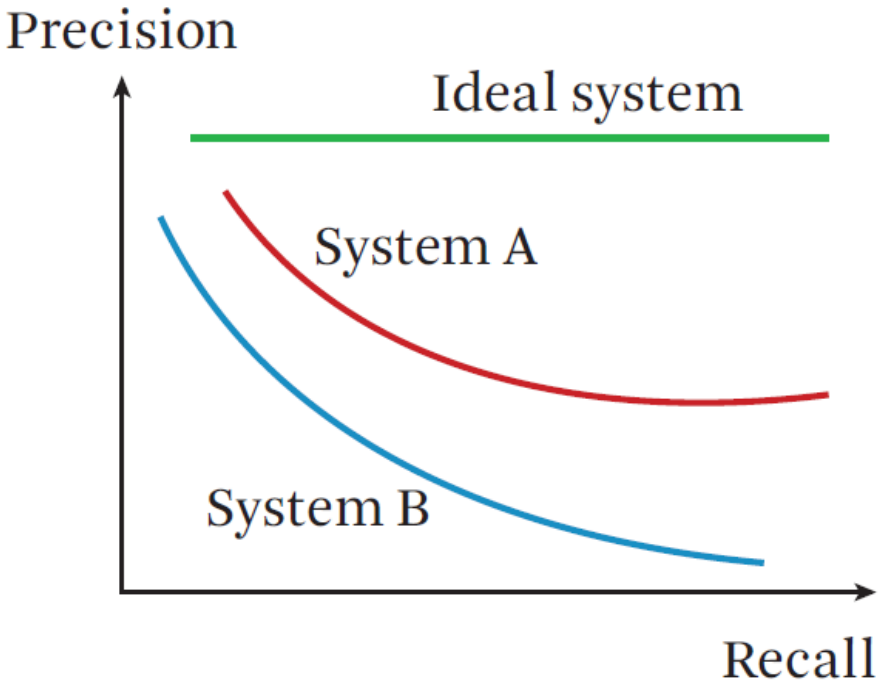
Total number of relevant documents in collection: 10

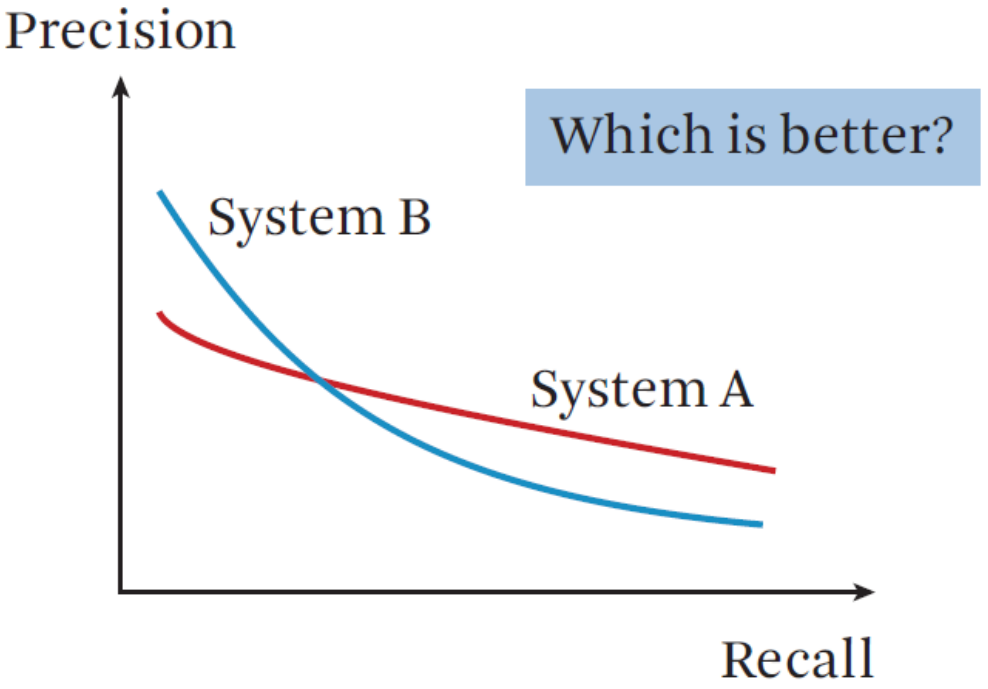
	Precision	Recall
D_1 +	1/1	1/10
D_2 +	2/2	2/10
D_3 –	2/3	2/10
D_4 –		
D_5 +	3/5	3/10
D_6 –		
D_7 –		
D_8 +	4/8	4/10
D_9 –		
D_{10} –	?	10/10



Computing a precision-recall curve.

evaluation of a ranked list





Comparison of two PR curves. (Courtesy of Marti Hearst)

evaluation of a ranked list: average precision (AP)

mean of the **precisions at the positions where we retrieved** the first **relevant document**, the second, and so on.

$$\text{avp}(L) = \frac{1}{|Rel|} \sum_{i=1}^{n} p(i),$$

if the ranking missed many relevant documents so in all of these cases we assume that they have zero precision

evaluation of a ranked list: average precision (AP)

$$\text{avp}(L) = \frac{1}{|Rel|} \sum_{i=1}^{n} p(i),$$

Unlike P@K, AP is sensitive to the ranking of each individual relevant document

i	Rel	$p(i)$
1	+	$\frac{1}{1} = 1.0$
2	+	$\frac{2}{2} = 1.0$
3	–	0.0
4	–	0.0
5	+	$\frac{3}{5} = 0.6$
6	–	0.0
7	–	0.0
8	+	$\frac{4}{8} = 0.5$
⋮	–	0.0
sum		3.1
avp		$\frac{3.1}{10} = 0.31$

evaluating ranked lists from multiple queries

Mean Average Precision (MAP):

average of the average precision over all the queries

$$MAP(\mathcal{L}) = \frac{1}{m} \sum_{i=1}^{m} \text{avp}(\mathcal{L}_i).$$

known-item search

known item search: one relevant document in the entire collection

AP boils down to the reciprocal rank: 1/r, where r is the position of the (single) relevant document

mean reciprocal rank (MRR): average of all the reciprocal ranks over a set of topics

evaluation with multi-level judgements (NDCG)

	Gain	Cumulative gain	Discounted cumulative gain
D_1	3	3	3
D_2	2	3 + 2	$3 + 2/\log 2$
D_3	1	3 + 2 + 1	$3 + 2/\log 2 + 1/\log 3$
D_4	1	3 + 2 + 1 + 1	...
D_5	3	...	
D_6	1		
D_7	1		
D_8	2		
D_9	1		
D_{10}	1		

$$\text{Normalized DCG} = \frac{\text{DCG@10}}{\text{IdealDCG@10}}$$

$$\text{DCG@10} = 3 + 2/\log 2 + 1/\log 3 + \cdots + 1/\log 10$$

$$\text{IdealDCG@10} = 3 + 3/\log 2 + 3/\log 3 + \cdots + 3/\log 9 + 2/\log 10$$

Relevance level: $r = 1$ (non-relevant), 2 (marginally relevant), 3 (very relevant)

Assume: there are 9 documents rated "3" in total in the collection

practical issues in evaluation

documents and **queries** must be **representative**

use **many queries** and **many documents** in order to avoid biased conclusions

complete relevance **judgements** vs minimizing human effort

correlate the **evaluation measures** with the perceived **utility of users**

statistical significance tests

mathematically quantify whether the evaluation scores of two systems are **indeed different**

gives us an idea as to how likely a difference in evaluation scores is due to **random chance**

consistently better

vs

random fluctuation

Experiment I			Experiment II		
Query	System A	System B	Query	System A	System B
1	0.20	0.40	1	0.02	0.76
2	0.21	0.41	2	0.39	0.07
3	0.22	0.42	3	0.16	0.37
4	0.19	0.39	5	0.58	0.21
5	0.17	0.37	6	0.04	0.02
6	0.20	0.40	6	0.09	0.91
7	0.21	0.41	7	0.12	0.46
Average	0.20	0.40	Average	0.20	0.40

Statistical significance: two sets of experiments with an identical MAP. (Courtesy of Douglas W. Oard and Philip Resnik)

statistical significance tests

the idea behind these tests is to assess the variance in average precision scores (or any other score) across the queries. If there's a big variance, that means that the results could fluctuate according to different queries, which makes the result unreliable.

Query	System A	System B	Sign Test	Wilcoxon
1	0.02	0.76	+	+0.74
2	0.39	0.07	−	−0.32
3	0.16	0.37	+	+0.21
4	0.58	0.21	−	−0.37
5	0.04	0.02	−	−0.02
6	0.09	0.91	+	+0.82
7	0.12	0.46	+	+0.34
Average	0.20	0.40	$p = 1.0$	$p = 0.9375$

Statistical significance tests. (Courtesy Douglas W. Oard and Philip Resnik)

building (incomplete) relevance judgments: pooling

if we can't afford judging all the documents in the collection, which subset should we judge?

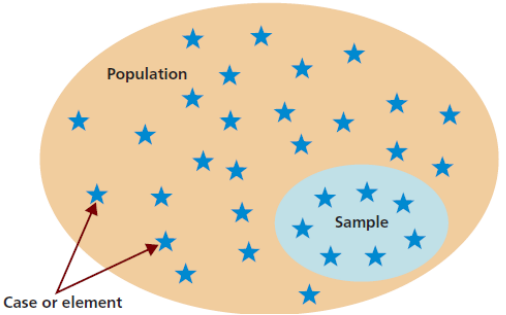
1) choose a **diverse set of ranking methods**; these are different types of retrieval systems.

these methods help us nominate likely relevant documents.

2) each system returns the **top k documents** according to its ranking function

3) **combine** all these top k sets to form a **pool** of documents for human assessors to judge

4) the remaining **unjudged** documents are assumed to be **non-relevant**



relevance assessments are incomplete

relevance assessments are incomplete

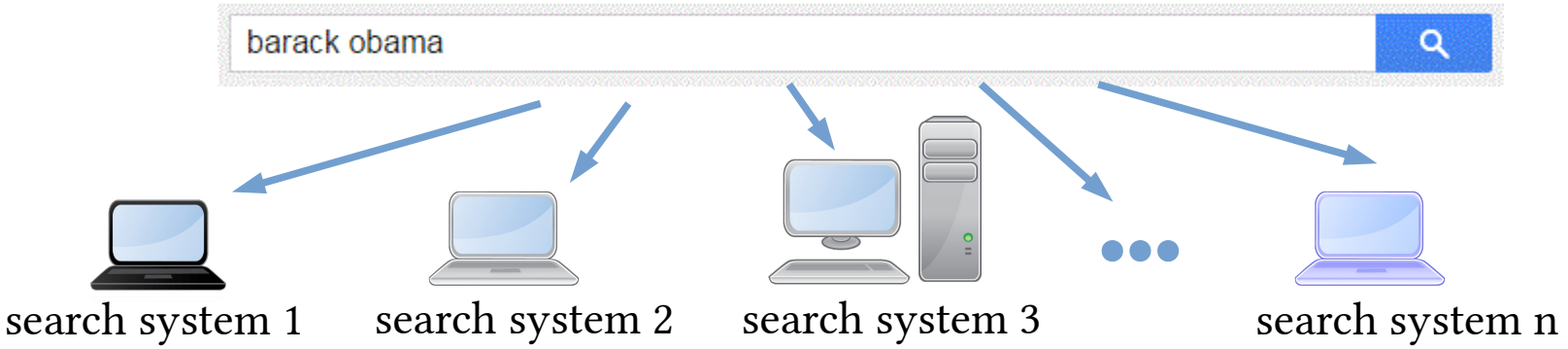
search system 1

search system 2

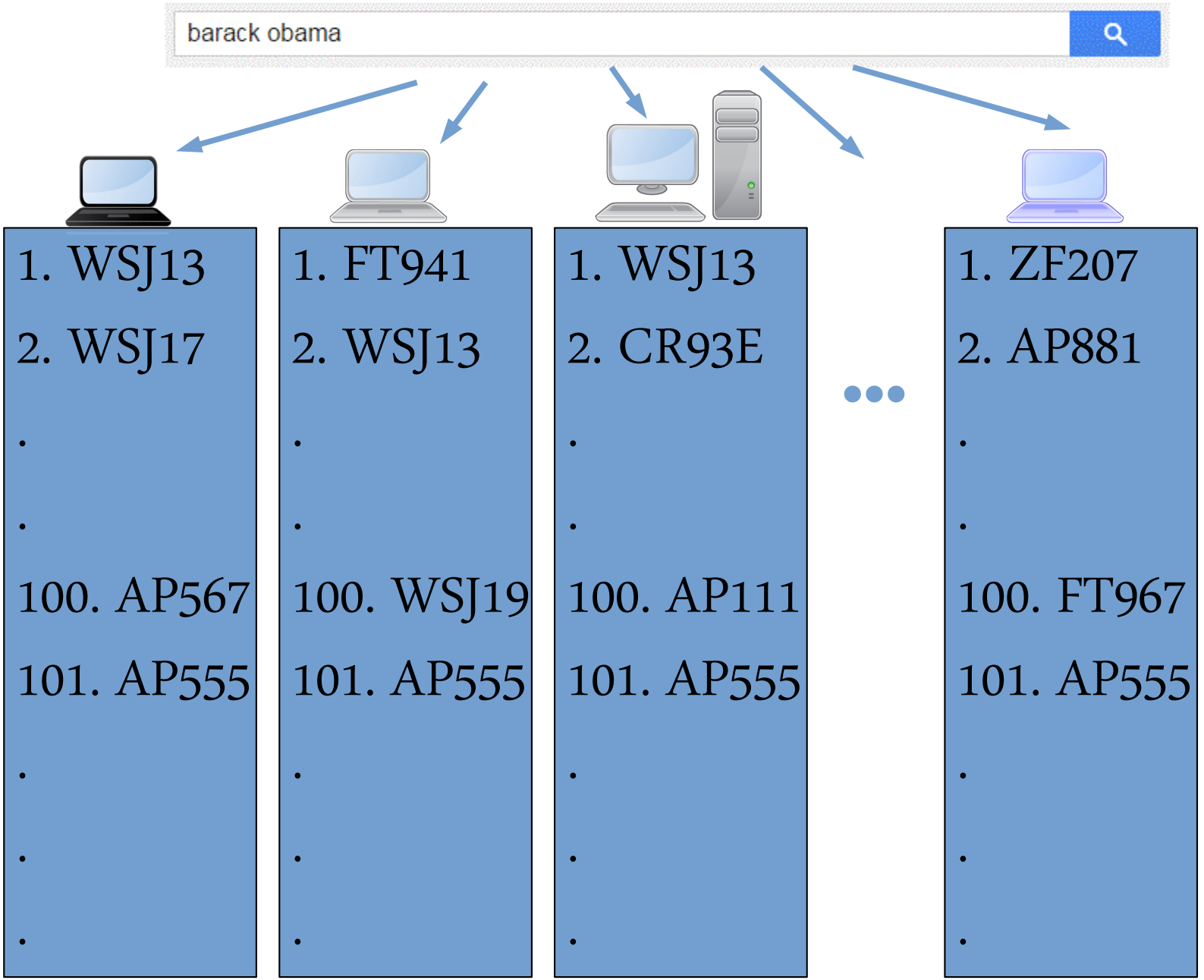
search system 3

search system n

relevance assessments are incomplete

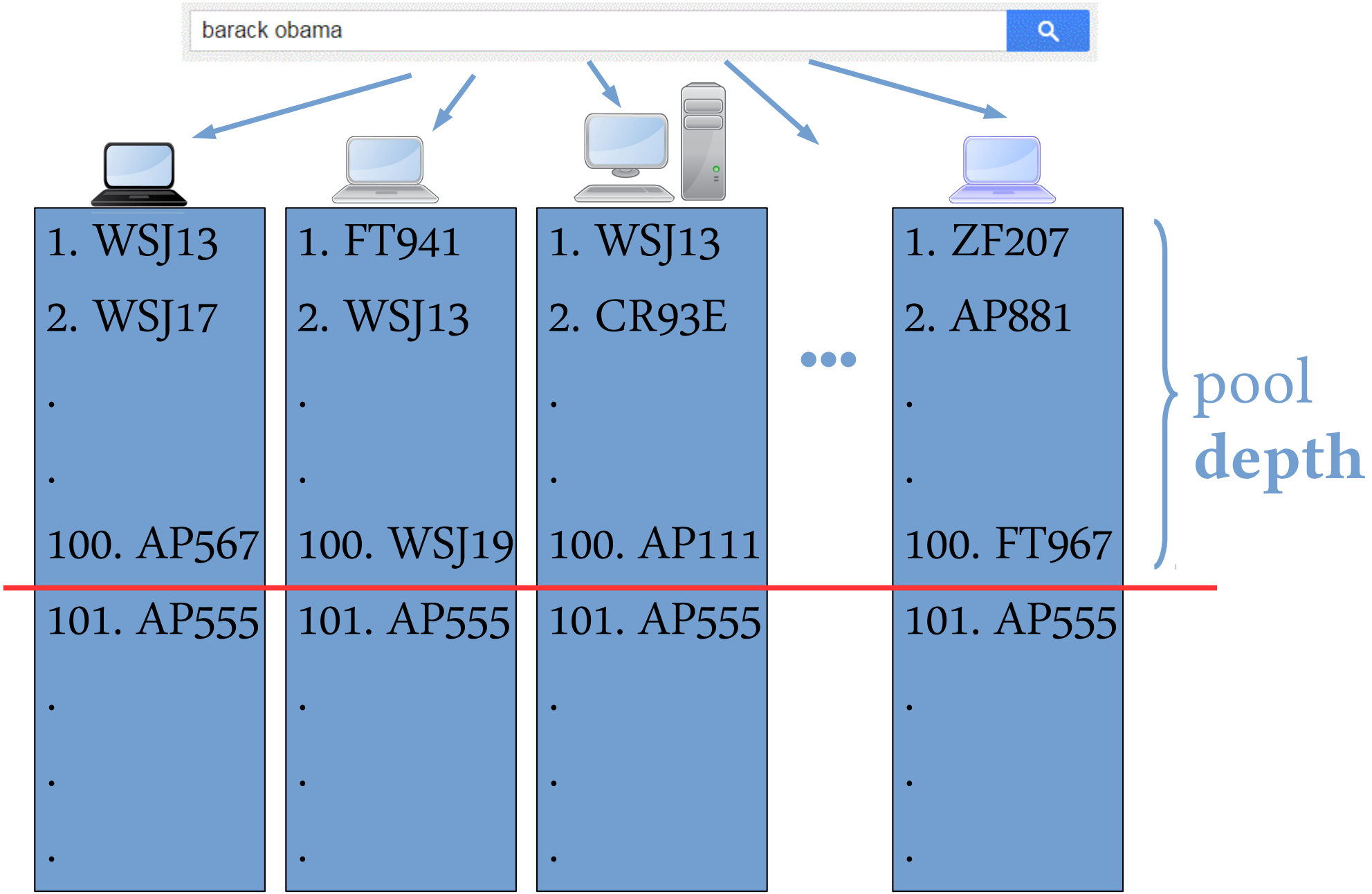


relevance assessments are **incomplete**

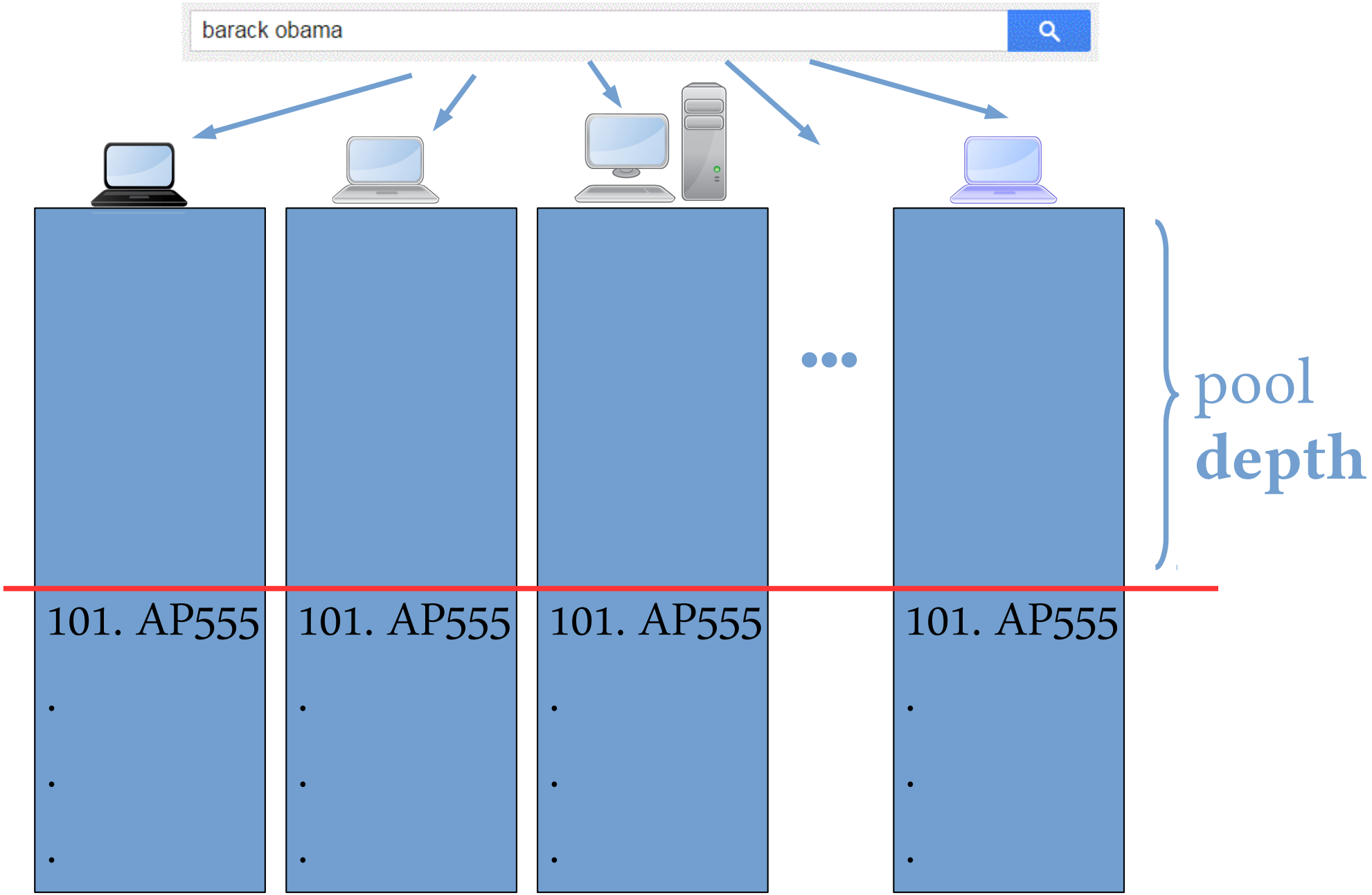


rankings of docs by **estimated relevance** (runs)

relevance assessments are **incomplete**

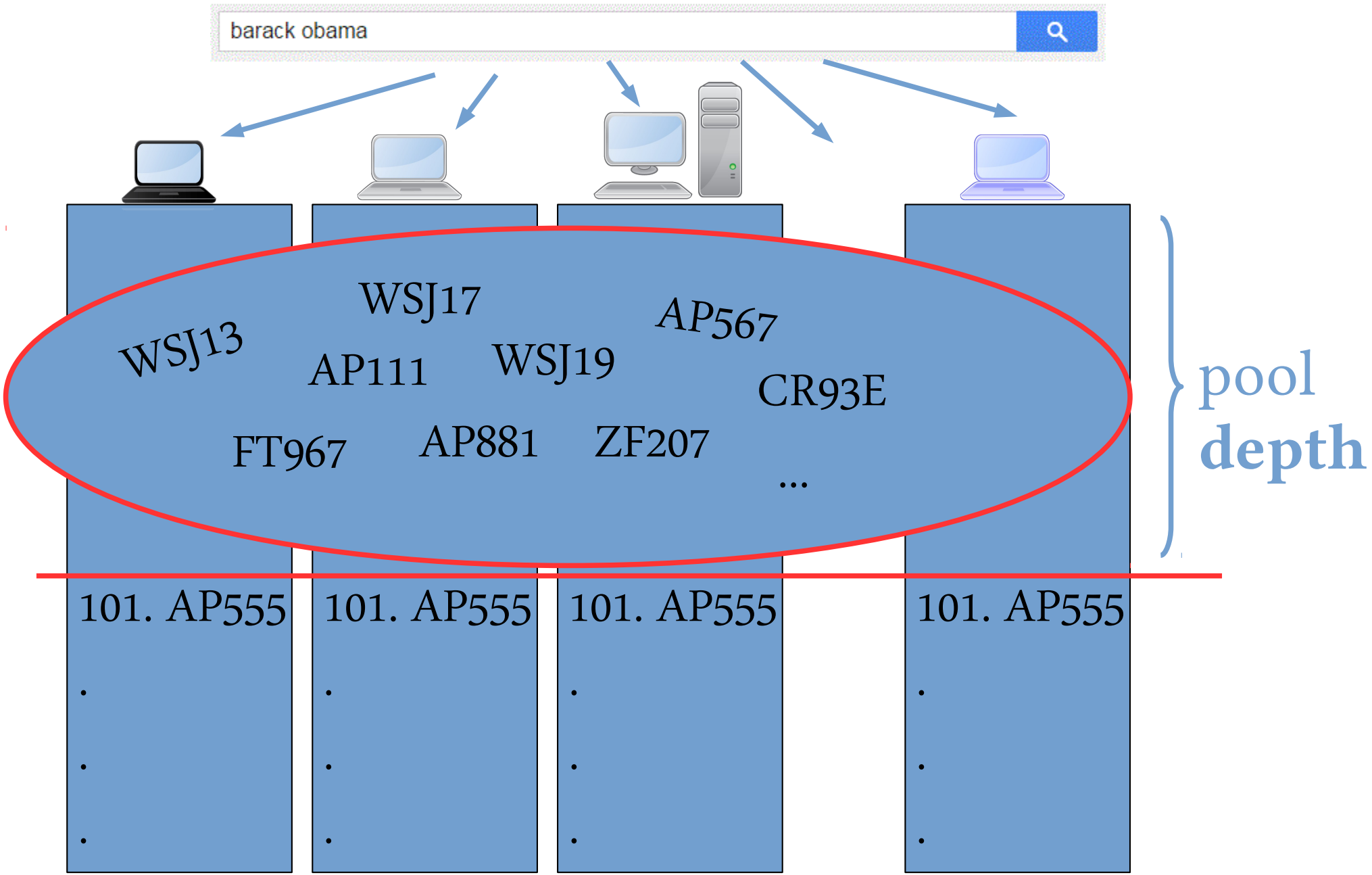


relevance assessments are **incomplete**



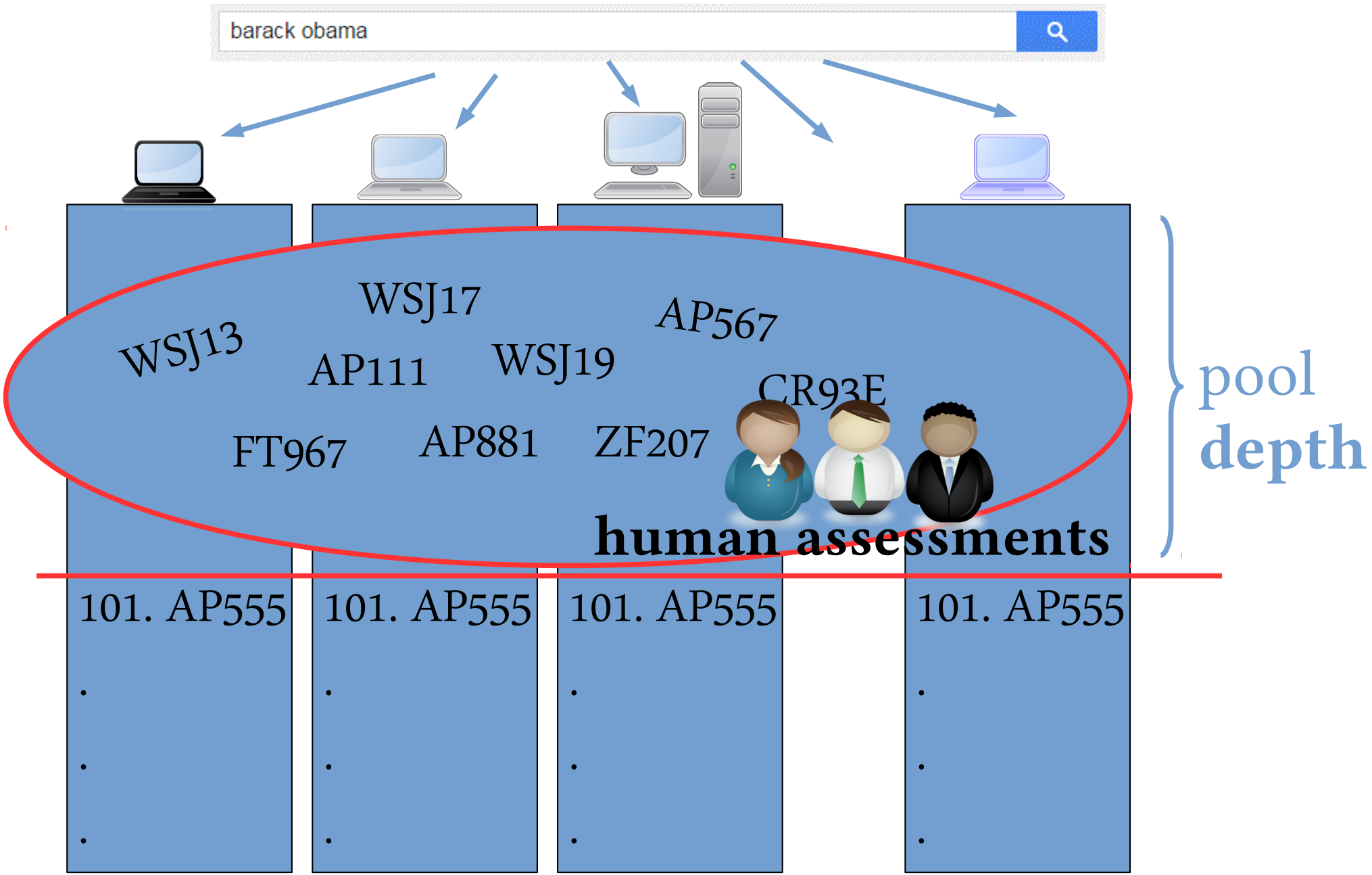
rankings of docs by **estimated relevance** (runs)

relevance assessments are **incomplete**



rankings of docs by **estimated relevance** (runs)

relevance assessments are **incomplete**



rankings of docs by **estimated relevance** (runs)