

Information Retrieval and Text Mining for Biology

Julien Gobeill, PhD

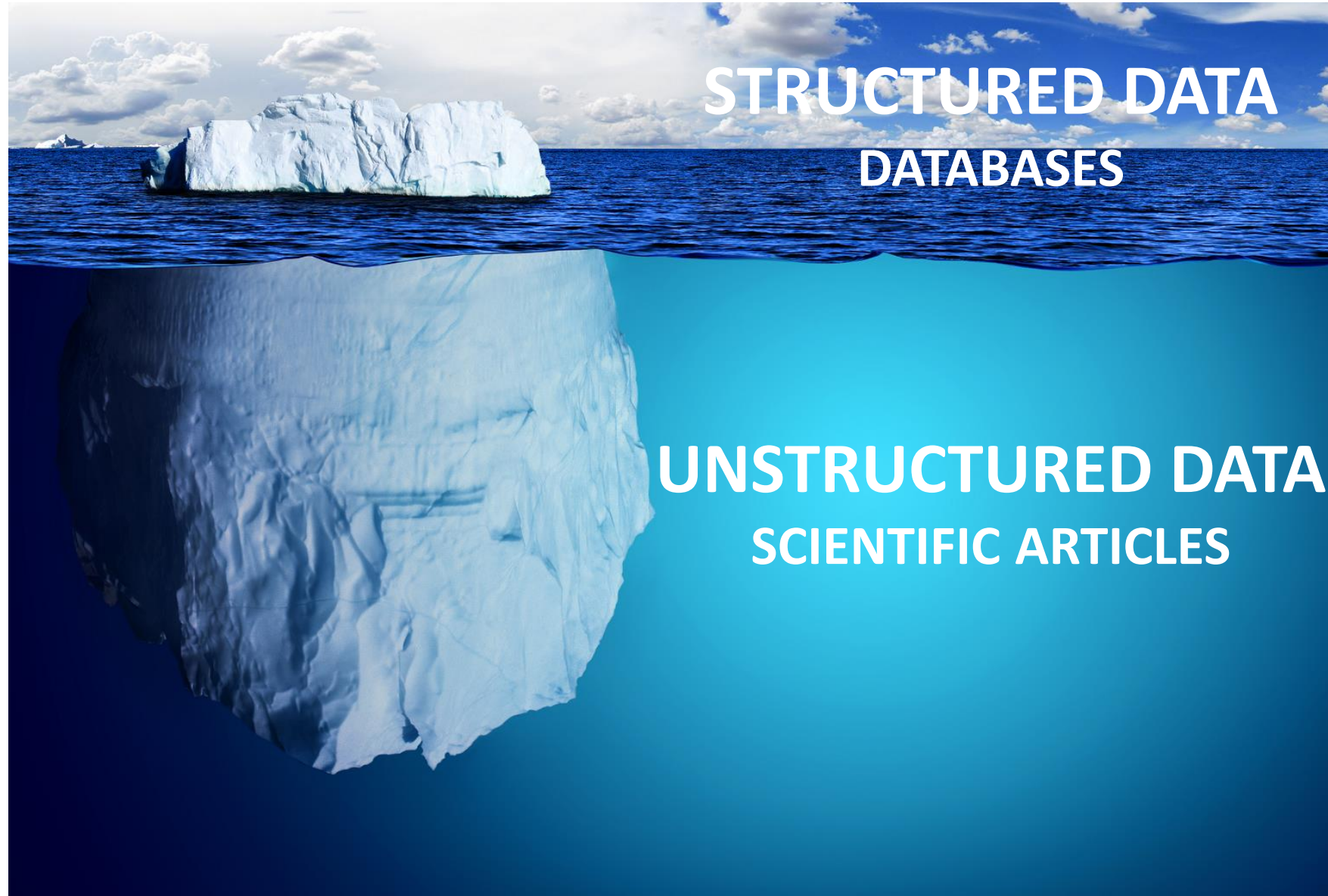
SIBtex group



Swiss Institute of
Bioinformatics

Introduction

Structured versus unstructured data

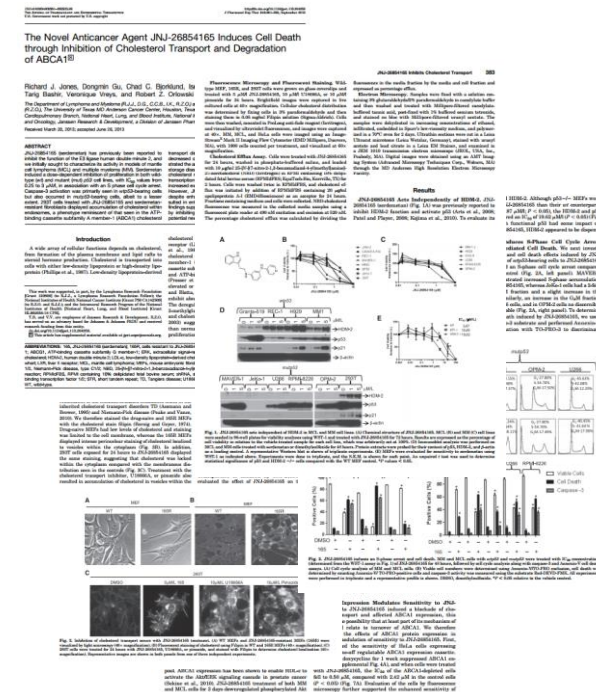


P04637 - P53_HUMAN

Protein | Cellular tumor antigen p53

GO - Molecular function

- ATP binding Source: UniProtKB
- chaperone binding Source: UniProtKB
- chromatin binding Source: UniProtKB



Information Retrieval ?



Information Retrieval

finding relevant documents in a collection of documents

Text Mining

extracting or deriving high-quality information from a text

Text Mining ?



IR and TM for Biology

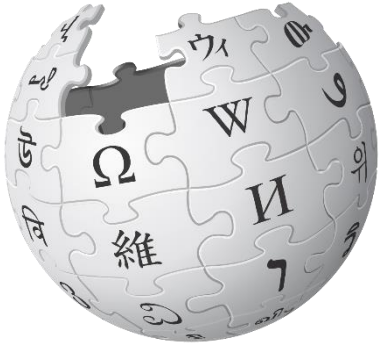
- Semi-automatic mode : to help biologists to access the information
 - Standard or knowledge-based **search engines**
 - **Triage** : tools that filter relevant articles for curation
 - **Named-Entities Recognition** : tools that highlight or normalize entities
 - Extraction of relations (PPI)
- Fully automatic mode :
 - Automatically populate databases

Challenges

- TREC for IR, BioCreative for biology
- Goal : to promote research by giving a common evaluation platform
- This includes :
 - Definition of user-centric relevant tasks
 - Preparation of data and infrastructure
 - Release of training and test data
 - **Evaluation**
 - Workshop / discussion / demo / articles

Information Retrieval Evaluation

Information Retrieval (IR)



Information retrieval is the activity of obtaining **information resources relevant** to an **information need** from a **collection** of **information resources**.

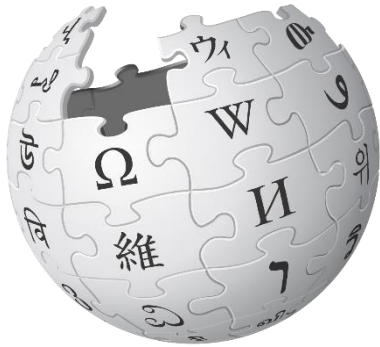
Information resources -> Documents / images / videos...

Collection

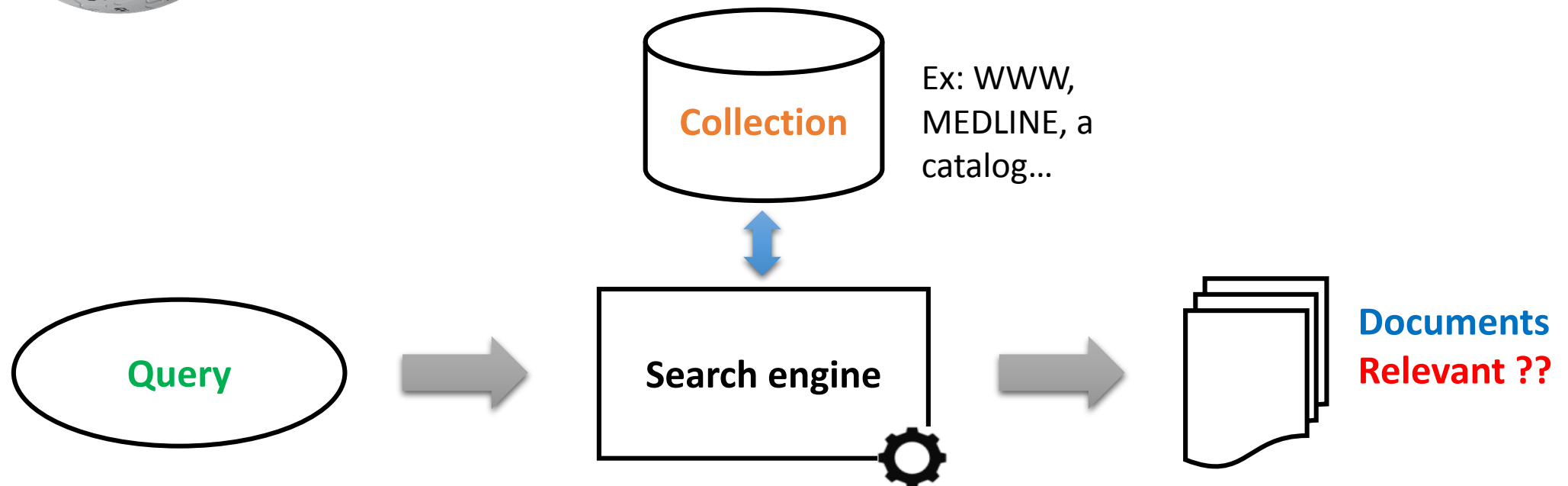
Relevant ??

Information need = query ??

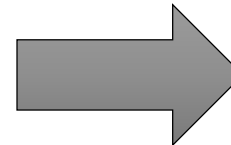
Information Retrieval



Information retrieval is the activity of obtaining **information resources relevant** to an **information need** from a **collection** of **information resources**.



Information need, query, relevance



Web Actualités Vidéos Images Maps Plus Outils de recherche

Environ 570 000 résultats (0,16 secondes)

Ajax fils de Télamon — Wikipédia
[fr.wikipedia.org/wiki/Ajax_fils_de_Télamon](https://fr.wikipedia.org/wiki/Ajax_fils_de_T%C3%A9lamon) ▼
Dans la mythologie grecque, **Ajax** (en grec ancien Αἴας Τηλαμῶνιος / Aías Tēlamōnios), fils de Télamon (roi de Salamine) et de Péribée, est un héros de la ...
Légende - Culte héroïque - Sources - Notes

Ajax — Wikipédia
fr.wikipedia.org/wiki/Ajax ▼
Aller à l'historique des modifications | Modifier le code | Opération Ajax, un c...
appuyé par les États-Unis et le Royaume-Uni en Iran en 1953. Ajax ...
Ajax fils de Télamon - Ajax (informatique) - (1404) Ajax

L'histoire de l'Ajax Amsterdam - Ajax en France
www.ajaxenfrance.com/page-histoire.html ▼
C'était la première fois que le club était champion national depuis 1919. Le 1...
1931, l'Ajax obtient la plus large victoire de son histoire face à VUC 17-0.

L'Histoire d'Ajax - Colgate
www.colgate.fr/app/PDP/Ajax/FR/About/Story.cwsp ▼
Une chronologie retraçant l'histoire d'Ajax de 1958 à nos jours.



Information need and query



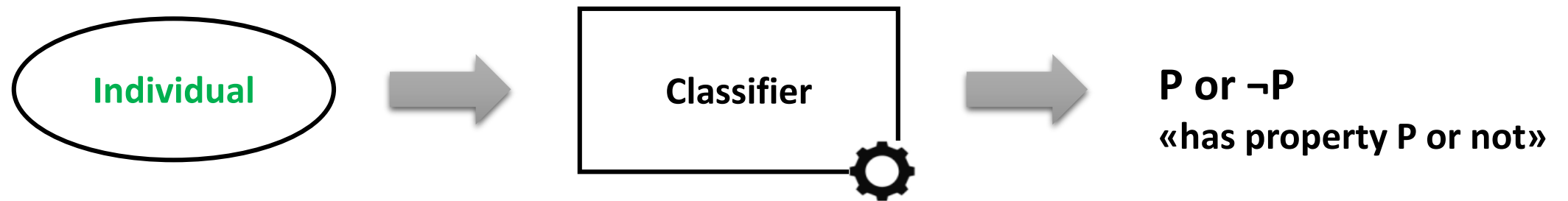
What is your favorite search engine ?

- And why ?



How to evaluate ?

Binary classifier evaluation



First classifier

Task: predicting property P for an individual (for ex: your kitten is a boy).



Results

Task: predict the gender of 40 kittens.

Method A		Real gender	
		♂	♀
Predicted gender	♂	18	8
	♀	2	12

Method B		Real gender	
		♂	♀
Predicted gender	♂	15	7
	♀	5	13

How to evaluate which method is the best ?

Accuracy : Method A 0.75 (30/40) > Method B 0.70 (28/40)

Confusion matrix

Task: to predict a property P

		Real class	
		P	$\neg P$
Predicted class	P	TP	FP
	$\neg P$	FN	TN

TP : True Positives



FP : False Positives

FN : False Negatives

TN : True Negatives

Confusion matrix

Information Retrieval task: spam filter : P normal mail ; \neg P spam

		Real class	
		P (normal)	\neg P (spam)
Predicted class		TP: 50	FP: 60
		FN: 0	TN: 890

- TP • • normal predicted as normal
- FP • • normal predicted as spam
- FN • • spam predicted as spam
- TN • • spam predicted as normal

TP : normal predicted as normal
FP : spam predicted as normal
FN : normal predicted as spam
TN : spam predicted as spam



Metrics : Accuracy

		Real class	
		P	¬P
Predicted class	P	TP	FP
	¬P	FN	TN

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

=
«percentage of predictions that
are correct»

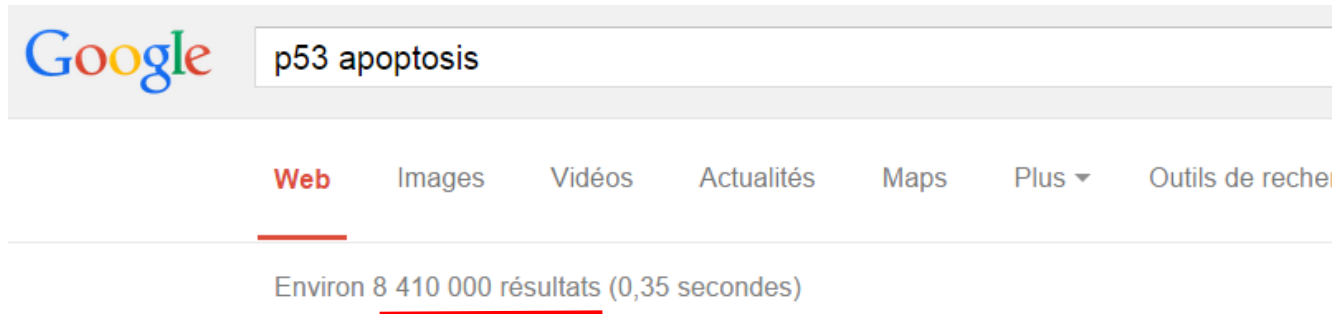
Confusion matrix : 2nd ex

		Filter A		Filter B	
		Real class		Real class	
		P (ham)	$\neg P$ (spam)	P (ham)	$\neg P$ (spam)
Predicted class		50	60	0	0
		0	890	50	950

Which filter has the best accuracy ?

Accuracy : Filter A 0.94 (940/1000) < Filter B 0.95 (950/1000)

Are there unbalanced data in IR ?

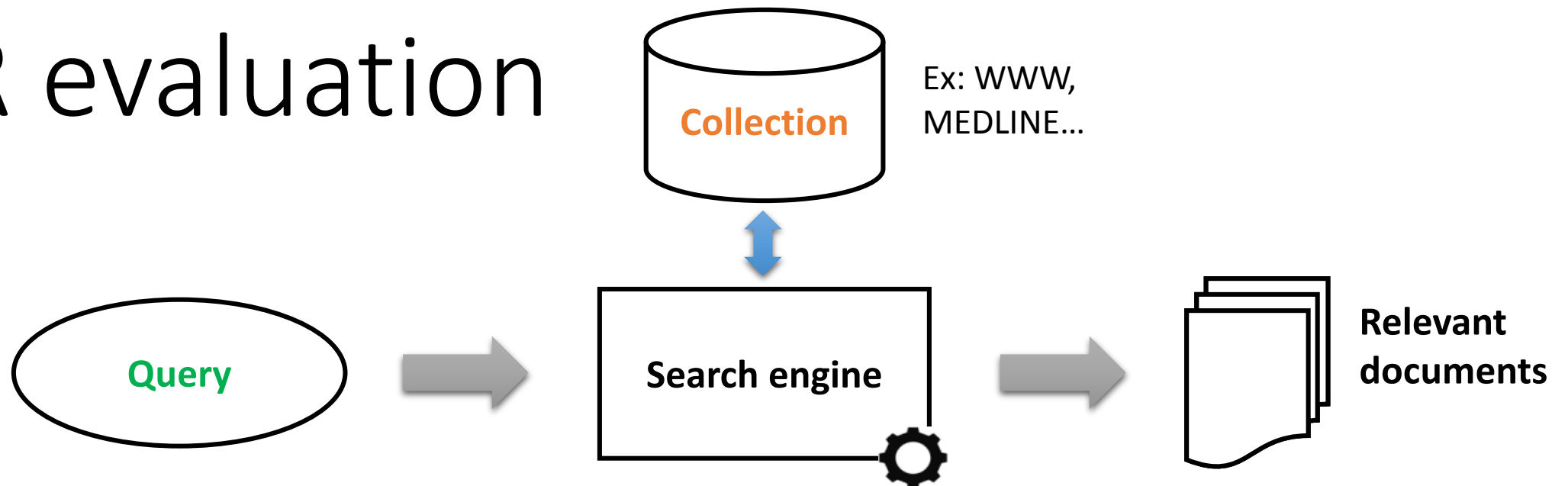


1000 billions of webpages

P: page talks about p53 apoptosis		Real class	
		P	$\neg P$
Predicted class	P	TP	FP
	$\neg P$	FN	TN

- How much is TP + FP ?
→ 8 410 000
- What are the TN ?
→ irrelevant pages
→ 999 991 590 000

IR evaluation



Metrics : Precision and Recall

- We only consider True Positives

		Real class	
		P	¬P
Predicted class	P	TP	FP
	¬P	FN	TN

$$\text{Precision } P = \frac{TP}{TP + FP}$$

$$\text{Recall } R = \frac{TP}{TP + FN}$$

Metrics : Precision and Recall

- Describe Precision and Recall with these words :

documents	relevant	that are
percentage	retrieved	of

		Real class	
		P	¬P
Predicted class	P	TP	FP
	¬P	FN	TN

$$\text{Precision } P = \frac{TP}{TP + FP}$$

=

«percentage of retrieved documents
that are relevant»

$$\text{Recall } R = \frac{TP}{TP + FN}$$

=

«percentage of relevant documents
that are retrieved»

Confusion matrix: 2nd ex

What is Precision P and Recall R for these 3 filters ?

Filter A		Real class	
		P	¬ P (spam)
Predict ed class	P	50	60
	¬ P (spam)	0	890

$$P(A) = 50/110 = 0.46$$

$$R(A) = 50/50 = 1$$

Filter B		Real class	
		P	¬ P (spam)
Predict ed class	P	0	0
	¬ P (spam)	50	950

$$P(B) = \text{undef or } 0$$

$$R(B) = 0/50 = 0$$

Filter C		Real class	
		P	¬ P (spam)
Predict ed class	P	31	12
	¬ P (spam)	19	938

$$P(C) = 31/43 = 0.72$$

$$R(C) = 31/50 = 0.62$$

Which is the best one ?

Metrics : F-measure

		Real class	
		P	¬P
Predicted class	P	TP	FP
	¬P	FN	TN

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

=
Harmonic mean of P and R

Confusion matrix: 2nd ex

What is Precision P and Recall R for these 3 filters ?

Filter A		Real class	
		P	¬ P (spam)
Predict ed class	P	30	70
	¬ P (spam)	20	890

$$P(A) = 30/100 = 0.30$$

$$R(A) = 30/50 = 0.60$$

Filter B		Real class	
		P	¬ P (spam)
Predict ed class	P	50	950
	¬ P (spam)	0	0

$$P(B) = 50/1000 = 0.05$$

$$R(B) = 50/50 = 1.00$$

Arithm. means

$$A : (0.30+0.60)/2 = \mathbf{0.45}$$

$$B : (0.05+1.00)/2 = \mathbf{0.525}$$

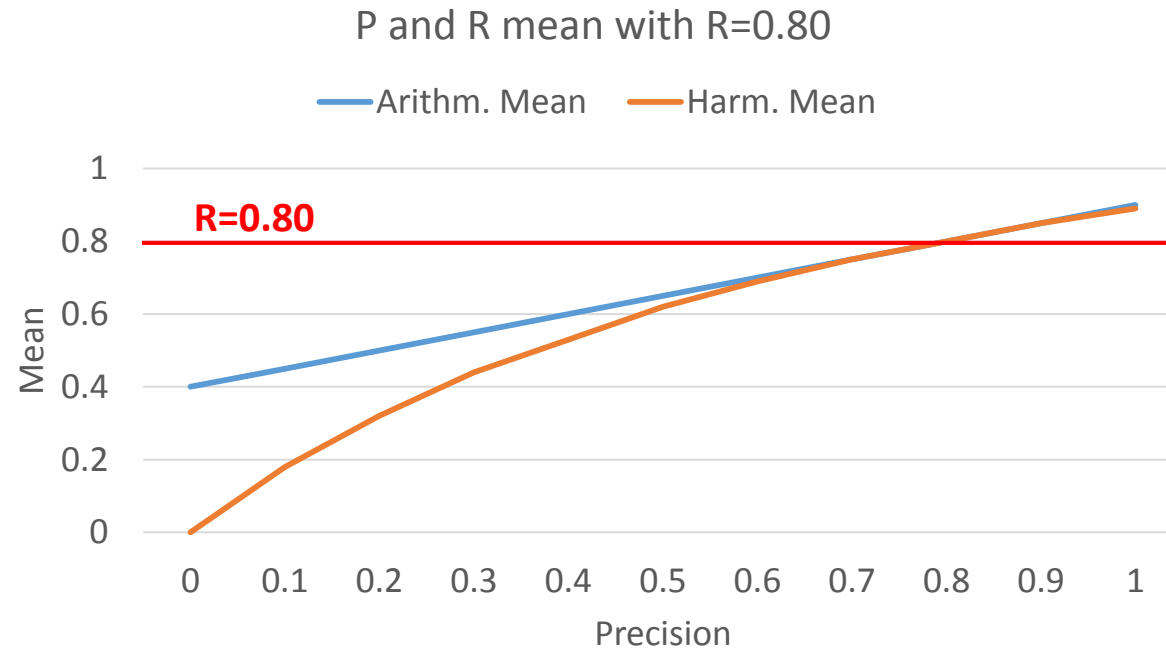
Harm. means

$$A : (2*0.30*0.60)/(0.30+0.60) = \mathbf{0.400}$$

$$B : (2*0.05*1)/(0.05+1) = \mathbf{0.095}$$

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

Harm. vs arithm. mean



➔ Weak values damage the mean.

P	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Arithm. Mean	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
Harm. Mean	0.00	0.18	0.32	0.44	0.53	0.62	0.69	0.75	0.80	0.85	0.89

Precision and recall : statement

- *A search engine retrieves 8 relevant documents, and 10 irrelevant ones. There are 20 relevant documents in the collection. For this search, please give :*
 - *The Precision P ?*
 - *The Recall R ?*
 - *The F -measure F ?*

$$P = 8/18 = 0.44$$

$$R = 8/20 = 0.40$$

$$F = 0.42$$

Ranked results evaluation

Google cancer prostate symptomes

Web Actualités Images Vidéos Maps Plus ▾ Outils de recherche

Environ 84 900 résultats (0,49 secondes)

Cancer de la Prostate - tena.fr ⓘ
Annonce www.tena.fr/CancerProstate-Incontinence ▾
Se rétablir après une opération Le témoignage de Richard, 66 ans
Fuites Urinaires : Causes Echantillon Gratuit
La Gamme Tena Men Espace Exercices

Symptômes du cancer de la prostate - ameli-santé
www.ameli-sante.fr ▾ Cancer de la prostate ▾
15 janv. 2014 - Au départ, le cancer de la prostate évolue sans symptômes. S'il s'agit d'un cancer localisé, il n'y a généralement pas de troubles urinaires.

Le Cancer De La Prostate: Ce Que Tout Homme Doit Savoir
<https://www.health.ny.gov/publications/0429/> ▾
Les Noirs américains ont le taux de cancer de la prostate le plus élevé au monde. ...
Quels Sont Les Symptomes Du Cancer De La Prostate? Au début de la ...

Cancer de la prostate : les symptômes - E-Santé
www.e-sante.fr ▾ Maladies ▾ Cancers ▾ Cancer de la prostate ▾
★★★★★ Note : 5 - 3 votes
Le plus souvent, le cancer de la prostate reste longtemps silencieux. Quand il se manifeste, on recense les signes suivants qui concernent essentiellement des ...



- What about P and R here ?
→ P et R metrics for sets of retrieved documents, not ranked lists

How to compare ranked results

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

Which is the best search engine ?

P et R at rank k

8 relevant docs in collection

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

P and R at rank 3 ?

→ A: $P_3=2/3$; $R_3=2/8$

→ B: $P_3=1/3$; $R_3=1/8$

P and R at rank 5 ?

→ A: $P_5=2/5$; $R_3=2/8$

→ B: $P_5=2/5$; $R_5=2/8$

→ For evaluation of the top results

→ Useful when we need few relevant documents

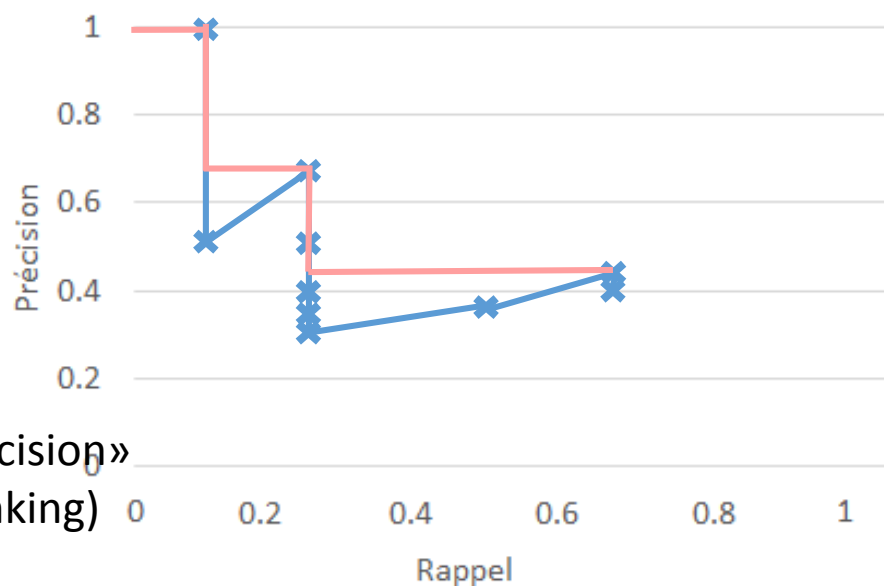
P and R at all ranks

6 relevant docs in collection

Rang	1	2	3	4	5	6	7	8	9	10
Moteur A	R	N	R	N	N	N	N	R	R	N
P at k	1	0.5	0.67	0.5	0.4	0.33	0.29	0.37	0.44	0.4
R at k	0.17	0.17	0.33	0.33	0.33	0.33	0.33	0.5	0.66	0.66

P/R curve, interpolated P

Rank	1	2		4	5	6	7	8		10
P at k	1	0.5	0.67	0.5	0.4	0.33	0.29	0.37	0.44	0.4
R at k	0.17	0.17	0.33	0.33	0.33	0.33	0.33	0.5	0.66	0.66



Interp. P at R=0.00 ?

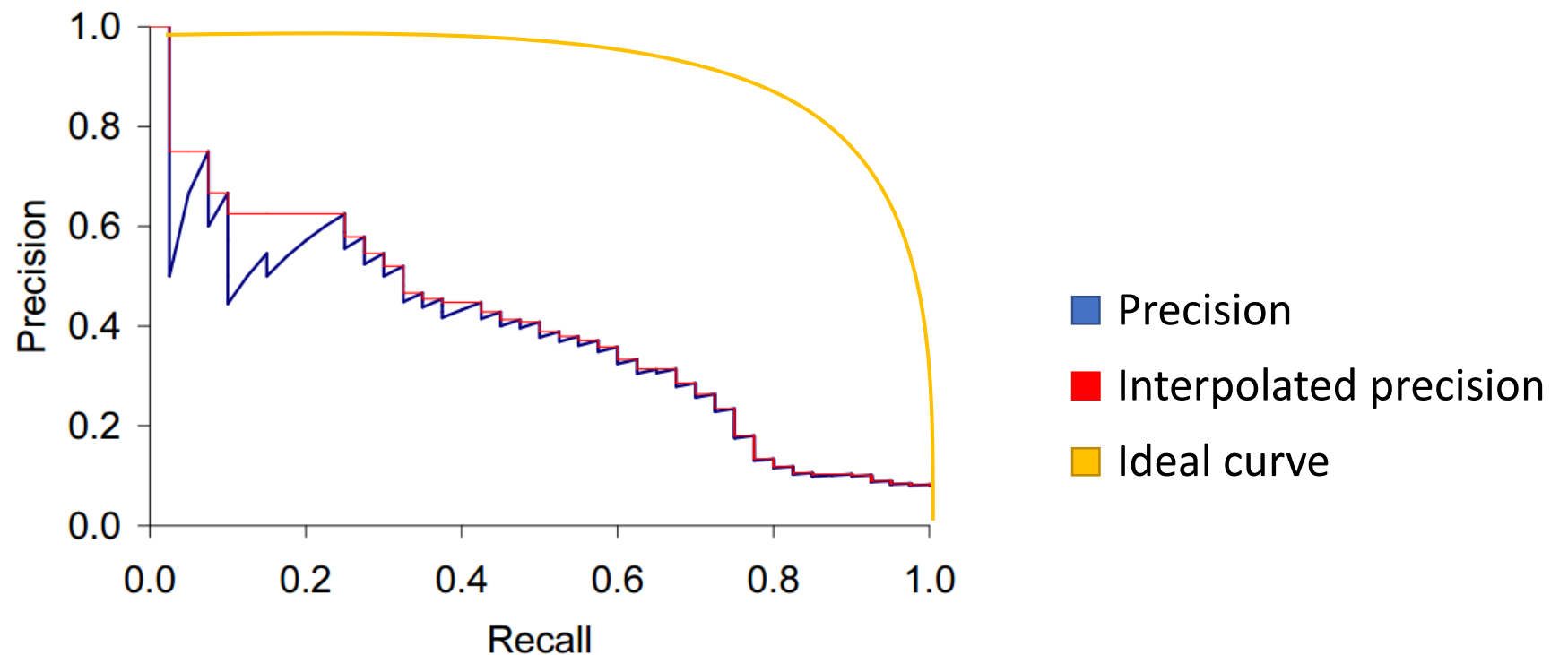
= «highest observed precision»
(useful for top of the ranking)

$$\text{Interp. } P(r) = \max P(r') \text{ with } r' \geq r$$

= «Maximum observed Precision for Recall higher or equal to r »

R at k	0.17	0.33	0.5	0.66
Interp. P	1	0.67	0.44	0.44

P/R curves, ROC curve



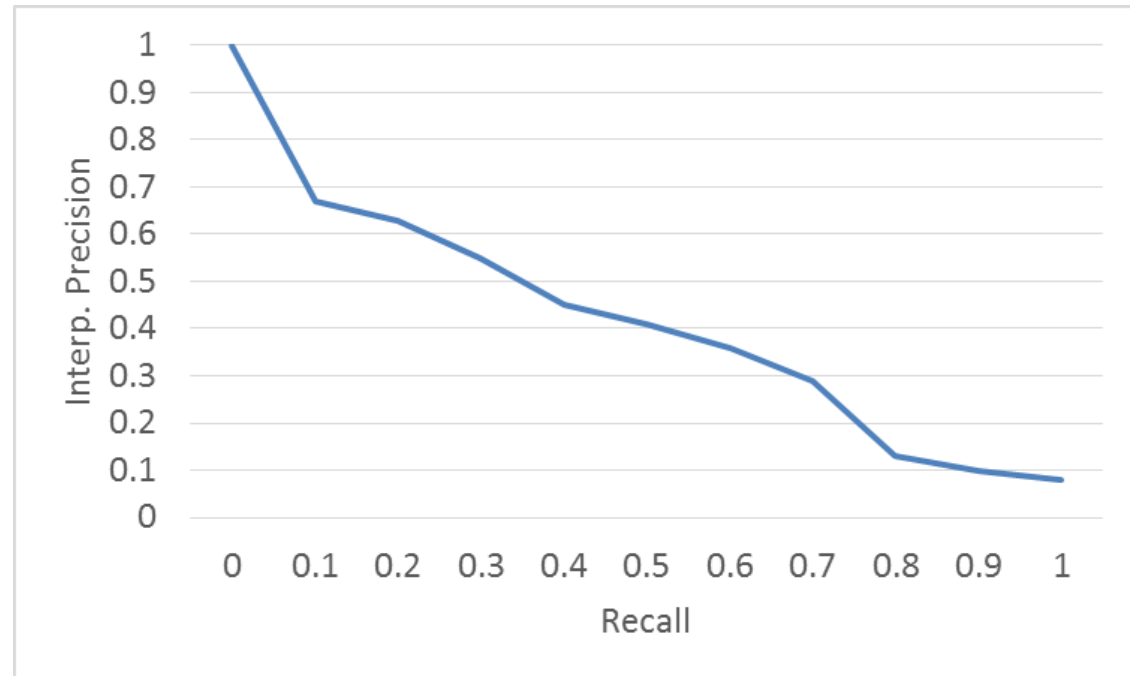
→ ROC curves are kind of P/R curves

How to project these curves into metrics ?

11-point Interp. P

- Interp. P at 11 different recall points
- Best with means of many queries

Recall	Interp. Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08



Mean Average Precision (MAP)

→ TREC metric, discriminative and stable

6 relevant docs in collection

Rank	1	2	3	4	5	6	7	8	9	10
P at k	1	0.5	0.67	0.5	0.4	0.33	0.29	0.37	0.44	0.4
R at k	0.17	0.17	0.33	0.33	0.33	0.33	0.33	0.5	0.66	0.66

Mean of first P values when a new relevant document is retrieved.
(0 for non retrieved relevant docs)

$$\text{Average Prec} = \frac{1 + 0.67 + 0.37 + 0.44 + 0 + 0}{6} = 0.41$$

→ Useful when we want many relevant documents

MAP in practical use

4 relevant docs in collection

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

Which search engine has the best MAP ?

$$MAP(A) = \frac{1 + 0.67 + 0.33 + 0.4}{4} = 0.6$$

$$MAP(B) = \frac{0.5 + 0.4 + 0.5 + 0.57}{4} = 0.49$$

Discounted Cumulative Gain

- Popular for evaluating IR on the large collections.
- Uses a relevance value rel_i (better than binary, for ex from 0 to 3)

$$DCG = \sum_{i=1}^n \frac{2^{rel_i} - 1}{\log_2(1 + i)}$$

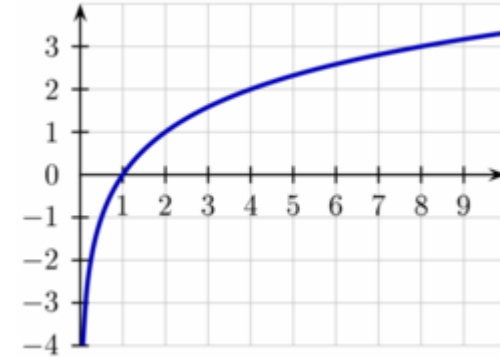
For each retrieved doc at rank i from 1 to n

The more the doc is relevant (high rel_i), the higher is DCG.

The higher is the rank, the less it's useful, so DCG is weaker.

Discounted Cumulative Gain: example

$$DCG = \sum_{i=1}^n \frac{2^{rel_i} - 1}{\log_2(1 + i)}$$



Rank	1	2	3	4	5	6	7	8	9	10	= i
Engine A	3	2	3	0	0	1	2	2	3	0	= rel_i
$2^{rel_i} - 1$	7	3	7	0	0	1	3	3	7	0	
$\log_2(1 + i)$	1	1.6	2	2.3	2.6	2.8	3	3.2	3.3	3.5	
DCG_i	7	1.9	3.5	0	0	0.4	1	0.9	2.1	0	

DCG = 4.4

A framework for evaluation

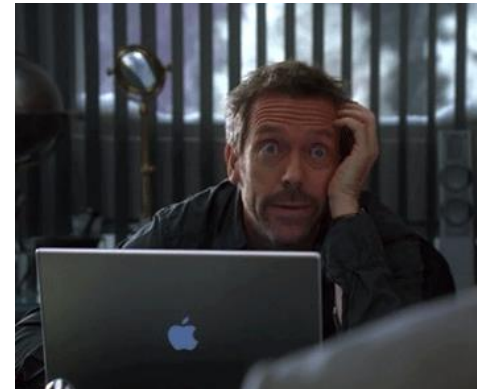
Text REtrieval Conferences (TREC)



- Yearly workshop in the IR domain
 - Several different tasks, many competing teams
 - Workshop at National Institute of Standard and Technologies, Washington DC
- Some tasks :
 - Tera-bytes Track, Legal Track, Spam-filtering, microblogs, Web...
- Impact :
 - 33% of improvements in IR come from TREC
 - For 1\$ invested TREC, ~4\$ economic benefits for public and private RI

One TREC 2014 task

- Clinical Decision Support Track : medical decision support
 - <http://www.trec-cds.org/2014.html>
- Goal: « retrieval of biomedical articles relevant for answering generic clinical questions about medical records »
- Ex. query: « 25-year-old woman with fatigue, hair loss, weight gain, and cold intolerance for 6 months »
- 20 competing groups, up to 4 different engines for each



Collection

- Which documents to return ?
 - ➔ ~~Everybody can return what he wants from the Web ?~~
 - ➔ Unique collection in order to evaluate engines without biases

- In TREC CDS 2014 :
 - 733,000 articles from biomedical journals
 - PubMed Central
 - Full texts in XML format

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 965764, 17 pages
<http://dx.doi.org/10.1155/2014/965764>

Review Article

Musculoskeletal Disorders in Chronic Obstructive Pulmonary Disease

Nele Cielien, Karen Maes, and Ghislaine Gayan-Ramirez

Respiratory Muscle Research Unit, Laboratory of Pneumology and Respiratory Division, Department of Clinical and Experimental Medicine, Katholieke Universiteit Leuven, Leuven, Belgium

Chronic obstructive pulmonary disease (COPD) is a lung disease characterized by airway obstruction and inflammation but also accompanied by several extrapulmonary consequences, such as skeletal muscle weakness and osteoporosis. Skeletal muscle weakness is of major concern, since it leads to poor functional capacity, impaired health status, increased healthcare utilization, and even mortality, independently of lung function. Osteoporosis leads to fractures and is associated with increased mortality, functional decline, loss of quality of life, and need for institutionalization. Therefore, the presence of the combination of these comorbidities will have a negative impact on daily life in patients with COPD. In this review, we will focus on these two comorbidities, their prevalence in COPD, combined risk factors, and pathogenesis. We will try to prove the clustering of these comorbidities and discuss possible preventive or therapeutic strategies.

1. Introduction

The Global Initiative for Chronic Obstructive Lung Disease (GOLD) defines chronic obstructive pulmonary disease (COPD) as "a preventable and treatable disease, characterized by a persistent airflow limitation that is progressive and not fully reversible and associated with an abnormal inflammatory response of the lungs to noxious gases or particles. Exacerbations and comorbidities contribute to the overall severity in individual patients" [1]. Currently, COPD is the fourth leading cause of death by 2030 [2].

COPD is spirometrically diagnosed by the presence of a postbronchodilator $FEV_1/FVC < 0.70$ and is assessed for its severity according to FEV_1 level: mild COPD ($FEV_1 \geq 0.80$ predicted), moderate COPD ($0.50 \leq FEV_1 < 0.80$ predicted), severe COPD ($0.30 \leq FEV_1 < 0.50$ predicted), and very severe COPD ($FEV_1 < 0.30$ predicted) [1]. In 2013, a new classi-

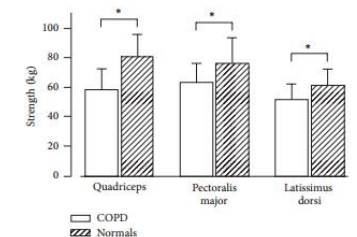


FIGURE 1: Reduced muscle strength of the quadriceps, pectoralis

Queries

- How to build queries ?
 - Queries must be representative of real world queries.
 - TREC CDS 2014 : « The topics for the track are medical case narratives created by expert topic developers that will serve as idealized representations of actual medical records » -> designed by physicians and IR experts
- How many queries for significant results ?
 - Metrics (such as MAP) can highly fluctuate from one query to another.
 - In the state of the art, 30 queries is considered as sufficient.

Relevant documents

- Relevant with respect to what ?
 - A document is relevant or not with respect to a given query.
 - Relevant yes/no ? Lightly/highly relevant ?
- How to obtain judgements ?
 - Human judgements (domain expert)
 - Expensive
 - Reliable, but not fully reliable (subjectivity, Inter Annotator Agreement)
 - TREC CDS 2014 : 750,000 docs x 30 queries= 22M judgements

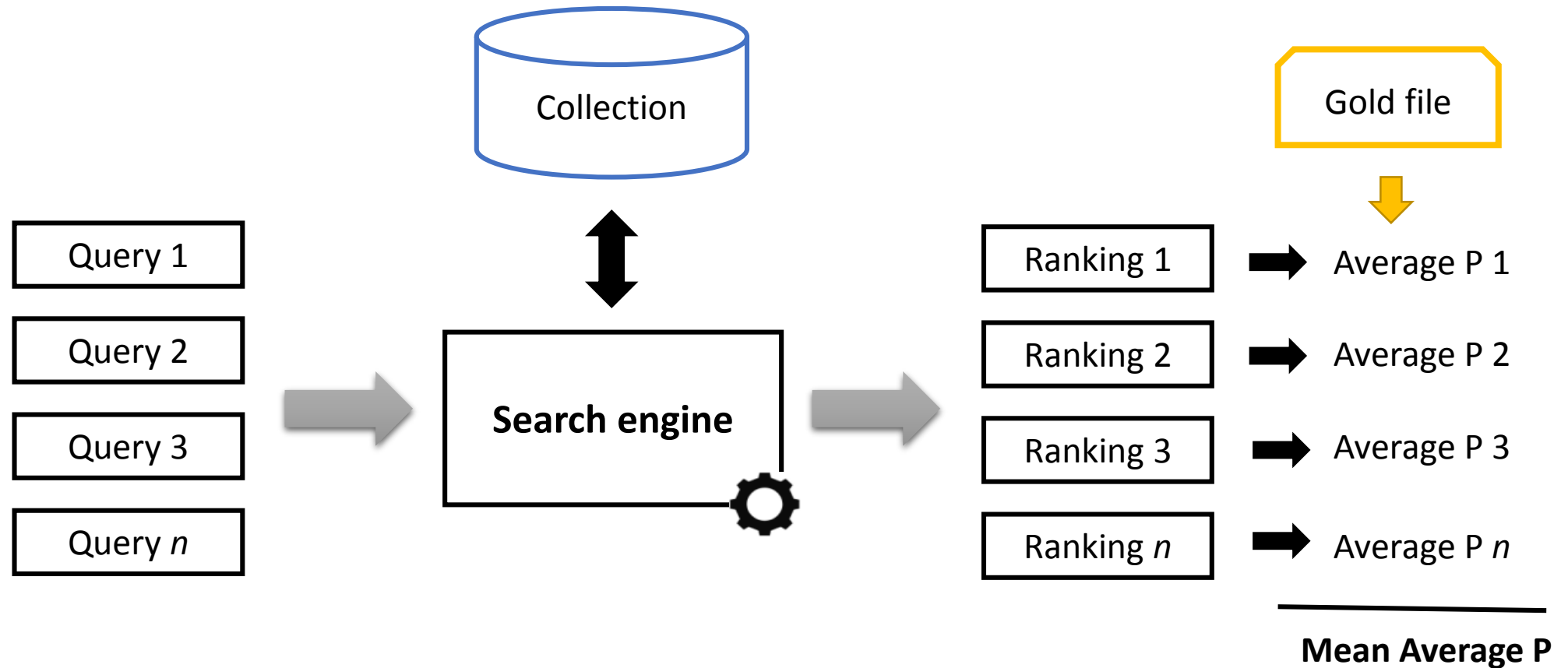
➔ File with all judgements : gold file, qrel, gold standard...



Benchmark

- A benchmark contains:
 1. A collection of documents (the same for all teams),
 2. Representative queries (~30-50),
 3. Relevance judgements (gold file): relevant documents for each query.
- Ex: TREC benchmarks

Benchmark in figure



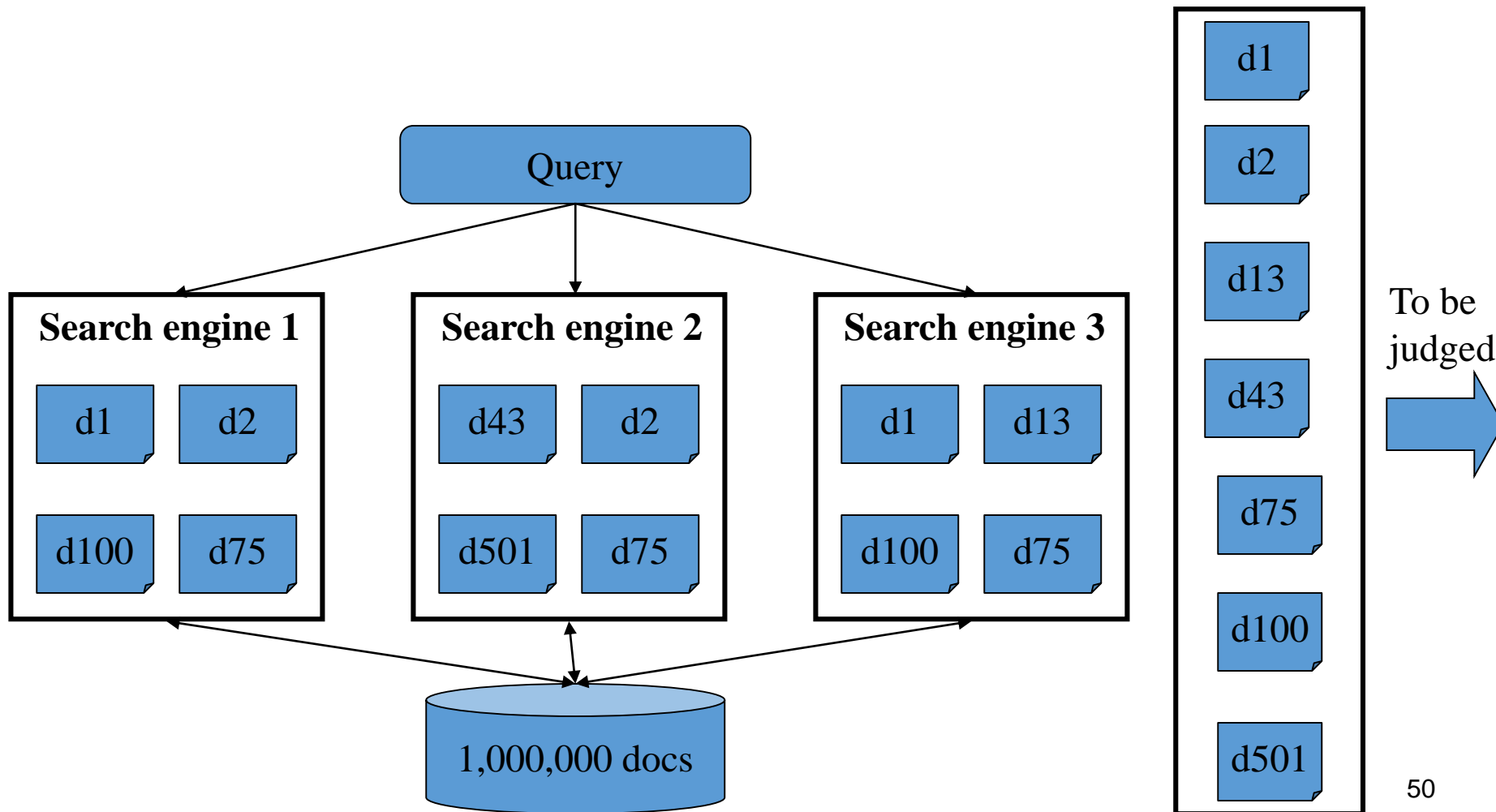
Gold file: pooling

- TREC CDS 2014
 - 30 queries, 80 runs of 1000 docs, 733'000 docs, 22M judgements...
 - How to deal with ?

➔ Pooling :

- Judgements only for retrieved documents
- Judgements only for a sample of retrieved documents :
 - Tops documents for all runs (ex: 20)
 - Then 80% of docs from rank 21 to 100
 - Then 50% of docs from rank 101 to 500...

Pooling: schéma



Shall we give the rank to the judge ?

- ➔ NO (bias)
- ➔ Random

Gold file: kappa

- Several judgements for a couple (requête,document)
 - Vote, union, inter...

➔ Kappa :

- Inter-Annotator Agreement (IAA)
- Observed agreement $P(A)$ compared to random $P(E)$

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

If IAA is 80%, can we have Recall up to 1 ?

➔ NO. Expert jugdements are subjective.

In practice: trec_eval

- Even the program which computes metrics is standard !
- <http://eagl.unige.ch/download/TRECCDS14.zip>
- password : TREC

Collection

```
<!DOCTYPE article PUBLIC "-//NLM//DTD Journal Archiving and Interchange DTD v2.3  
20070202//EN" "archivearticle.dtd"><article xmlns:xlink=  
"http://www.w3.org/1999/xlink" article-type="research-article"><?properties  
open_access?><front><journal-meta><journal-id journal-id-type="nlm-ta">Diabetes Care  
</journal-id><journal-id journal-id-type="publisher-id">diacare  
</journal-id><journal-title>Diabetes Care</journal-title><issn pub-type="ppub">  
0149-5992</issn><issn pub-type="epub">1935-5548</issn><publisher><publisher-name>  
American Diabetes Association  
</publisher-name></publisher></journal-meta><article-meta><article-id pub-id-type=  
"pmid">19106380</article-id><article-id pub-id-type="pmc">2646014  
</article-id><article-id pub-id-type="publisher-id">323387</article-id><article-id  
pub-id-type="doi">10.2337/dc08-0800</article-id><article-categories><subj-group  
subj-group-type="heading"><subject>Clinical Care/Education/Nutrition/Psychosocial  
Research</subject></subj-group></article-categories><title-group><article-title>  
Comparison of Glycemic Variability Associated With Insulin Glargine and  
Intermediate-Acting Insulin When Used as the Basal Component of Multiple Daily  
Injections for Adolescents With Type 1 Diabetes  
</article-title></title-group><contrib-group><contrib contrib-type="author"  
><name><surname>White</surname><given-names>Neil H.</given-names></name><degrees>MD,
```

Queries

<topic number="17" type="test">

<description>A 48-year-old white male with history of common variable immunodeficiency (CVID) with acute abdominal pain, fever, dehydration, HR of 132 bpm, BP 80/40. The physical examination is remarkable for tenderness and positive Murphy sign. Abdominal ultrasound shows hepatomegaly and abundant free intraperitoneal fluid. Exploratory laparotomy reveals a ruptured liver abscess, which is then surgically drained. After surgery, the patient is taken to the ICU.

</description>

<summary>48-year-old man with common variable immunodeficiency presents with abdominal pain and fever. Ultrasound reveals hepatomegaly and free intraperitoneal fluid. A ruptured liver abscess is found and drained during exploratory laparotomy.

</summary>

</topic>

<topic number="18" type="test">

<description>A 6-month-old male infant has a urine output of less than 0.2 mL/kg/hr shortly after undergoing major surgery. On examination, he has generalized edema. His blood pressure is 115/80 mm Hg, his pulse is 141/min, and his respiratory rate is 40/min. His blood sugar is 120 mg/dL, and his serum

Qrel (gold file) : 38000 lines!!

69	1	0	1379641	0	
70	1	0	1382225	0	For query 1, document 1382225 is not relevant
71	1	0	140012	0	
72	1	0	1402267	2	For query 1, document 1402267 is very relevant
73	1	0	1402280	1	
74	1	0	1413531	0	
75	1	0	1413554	0	
76	1	0	1413557	0	
77	1	0	1420300	0	

Run: 1000 retrieved docs per query

1	1	Q0	2790183	1	2.224	BiTeMSIBtex2
2	1	Q0	2902051	2	1.964	BiTeMSIBtex2
3	1	Q0	2572041	3	1.841	BiTeMSIBtex2
4	1	Q0	2778463	4	1.821	BiTeMSIBtex2
5	1	Q0	2672240	5	1.810	BiTeMSIBtex2
6	1	Q0	2922325	6	1.790	BiTeMSIBtex2
7	1	Q0	3098417	7	1.752	BiTeMSIBtex2
8	1	Q0	2895230	8	1.726	BiTeMSIBtex2
9	1	Q0	3696271	9	1.726	BiTeMSIBtex2
10	1	Q0	3874663	10	1.724	BiTeMSIBtex2
11	1	Q0	2983030	11	1.712	BiTeMSIBtex2

For query 1, my engine returns
the document 2672240 at rank 5
(score 1.810)

Stats computed with trec_eval

```
1
2 Queryid (Num) : ..... 30
3 Total number of documents over all queries
4 .... Retrieved: .... 30000
5 .... Relevant: ..... 3356
6 .... Rel_ret: ..... 1653
7 Interpolated Recall - Precision Averages:
8 .... at 0.00 ..... 0.5600
9 .... at 0.10 ..... 0.3092
10 .... at 0.20 ..... 0.2289
11 .... at 0.30 ..... 0.1562
12 .... at 0.40 ..... 0.1272
13 .... at 0.50 ..... 0.1042
14 .... at 0.60 ..... 0.0873
15 .... at 0.70 ..... 0.0716
16 .... at 0.80 ..... 0.0332
17 .... at 0.90 ..... 0.0062
18 .... at 1.00 ..... 0.0000
19 Average precision (non-interpolated) over all rel docs
20 ....|.....|.....| 0.1304
21 Precision:
22 .... At ..... 5 docs: .... 0.3600
```

Other quality criteria for users

- Interface
- Speed
- Clarity

➔ Not linked with relevance

Conclusion...

- There is not one perfect metric
 - Different metrics for different tasks
 - Different metrics for different engine properties
- Benchmark :
 - Representative queries (>30)
 - Gold file (kappa)