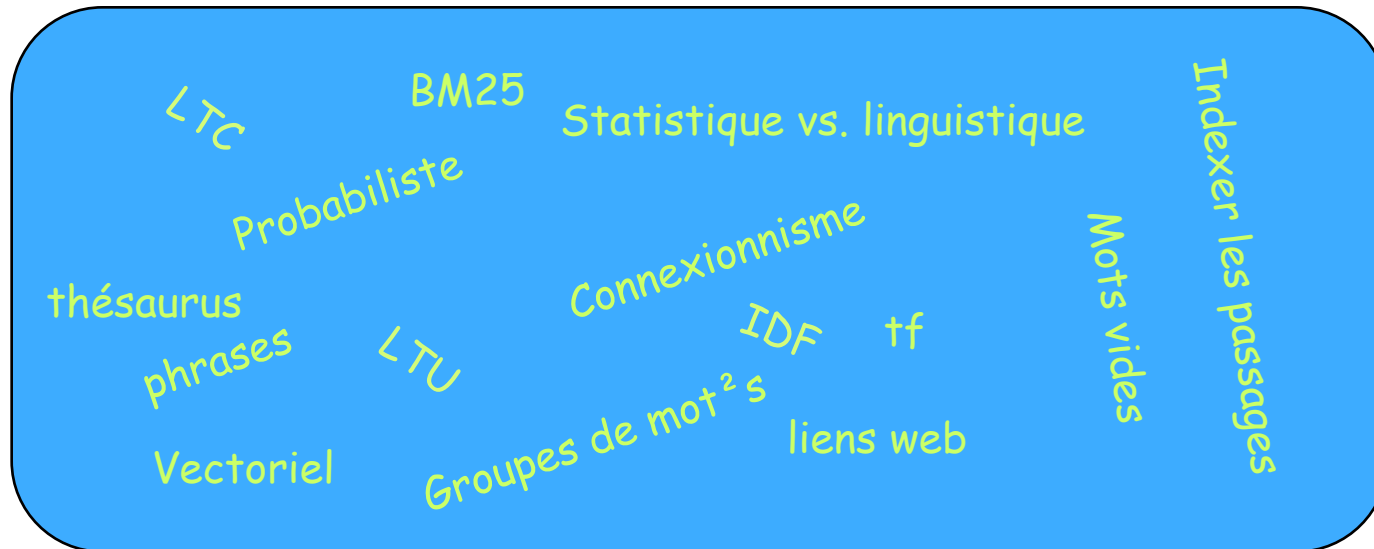


---

# Evaluation des performances dans les SRI

# Qu'est ce qui marche ?

---



Evaluer



# Objectif

---

- Evaluer la performance d'une approche, d'une technique, d'un système
  - En RI, on ne mesure pas la performance absolue d'un système/technique/approche car non significative
  - Mais, ..
    - Evaluation comparative entre approches
    - Mesurer la performance relative de A par rapport à B

# Critères d'évaluation

---

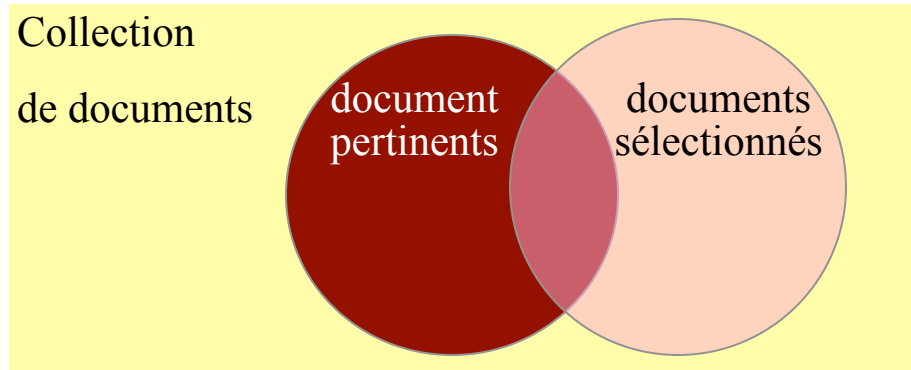
- Identifier la tâche à évaluer
- Identifier les critères (Cleverdon 66)
  - Facilité d'utilisation du système
  - Coût accès/stockage
  - Présentation des résultats
  - Capacité d'un système à sélectionner des documents pertinents.

# Deux facteurs

---

- Rappel
  - La capacité d'un système à sélectionner **tous** les documents pertinents de la collection
- Précision
  - La capacité d'un système à sélectionner **que** des documents pertinents

# Précision et Rappel



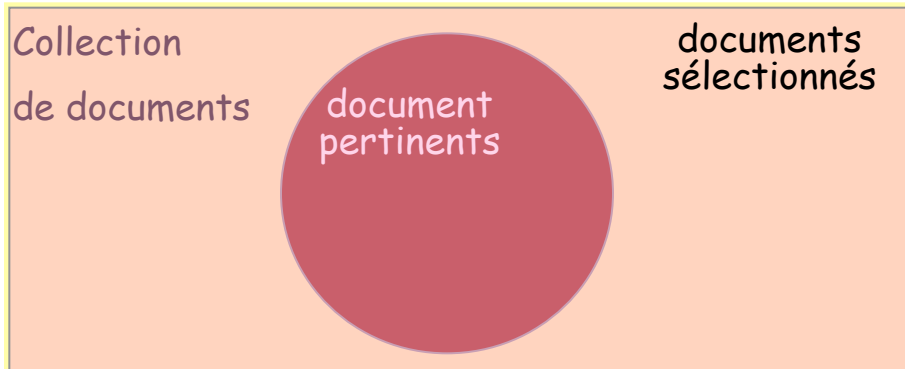
irrelevant	Sélection. & Non Pert.	Non sélection. & Non Pert.
	Sélection. & Pert	not sélection. mais Pert.
	retrieved	not retrieved

$$\text{rappel} = \frac{\text{Nombre de documents pertinents sélectionnés}}{\text{Nombre total de documents pertinents}}$$

$$\text{précision} = \frac{\text{Nombre de documents pertinents sélectionnés}}{\text{Nombre total de documents sélectionnés}}$$

# Pourquoi deux facteurs ?

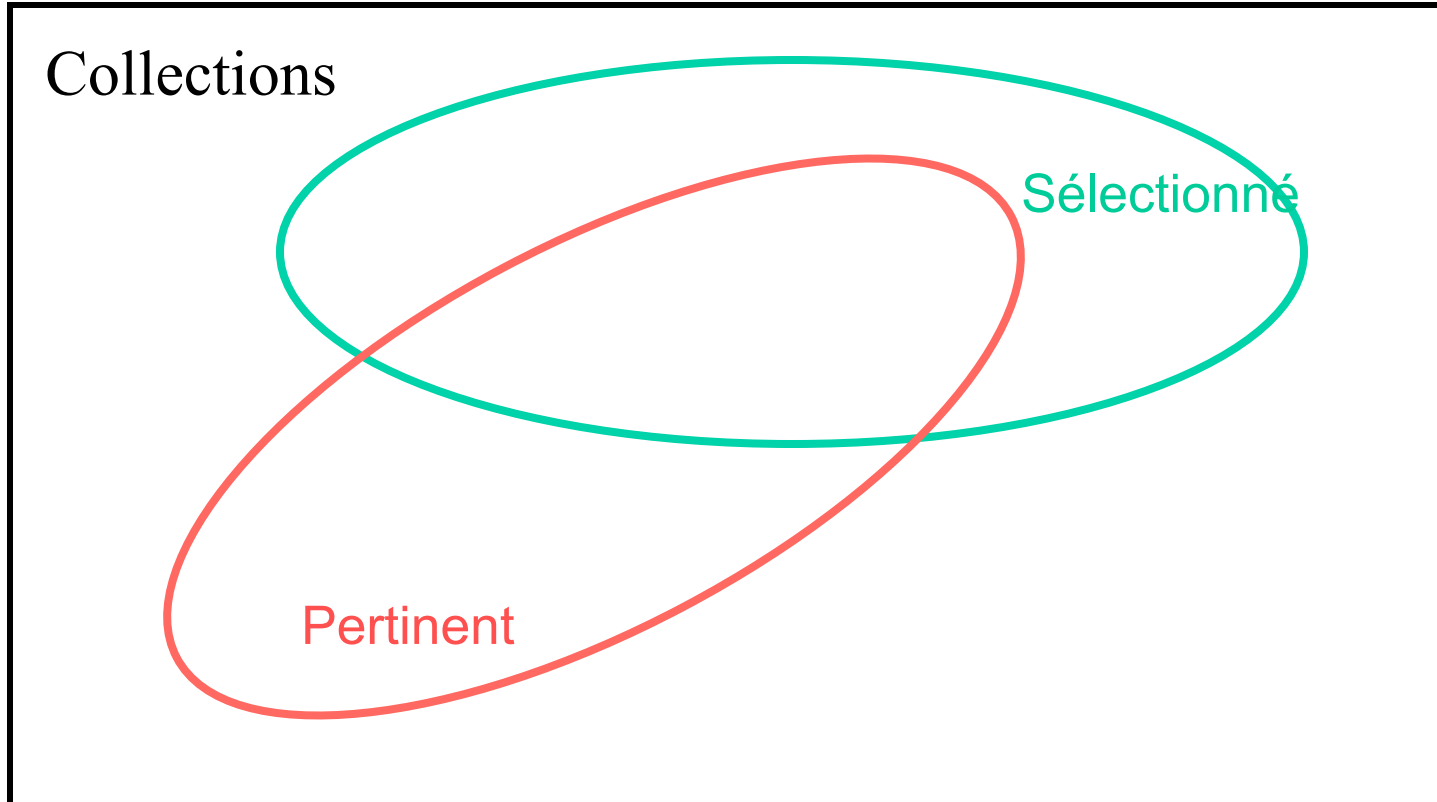
---



- **FACILE** de faire du rappel il suffit de sélectionner toute la collection
- **MAIS**, la précision sera très faible

# Pertinent vs. Sélectionné

---

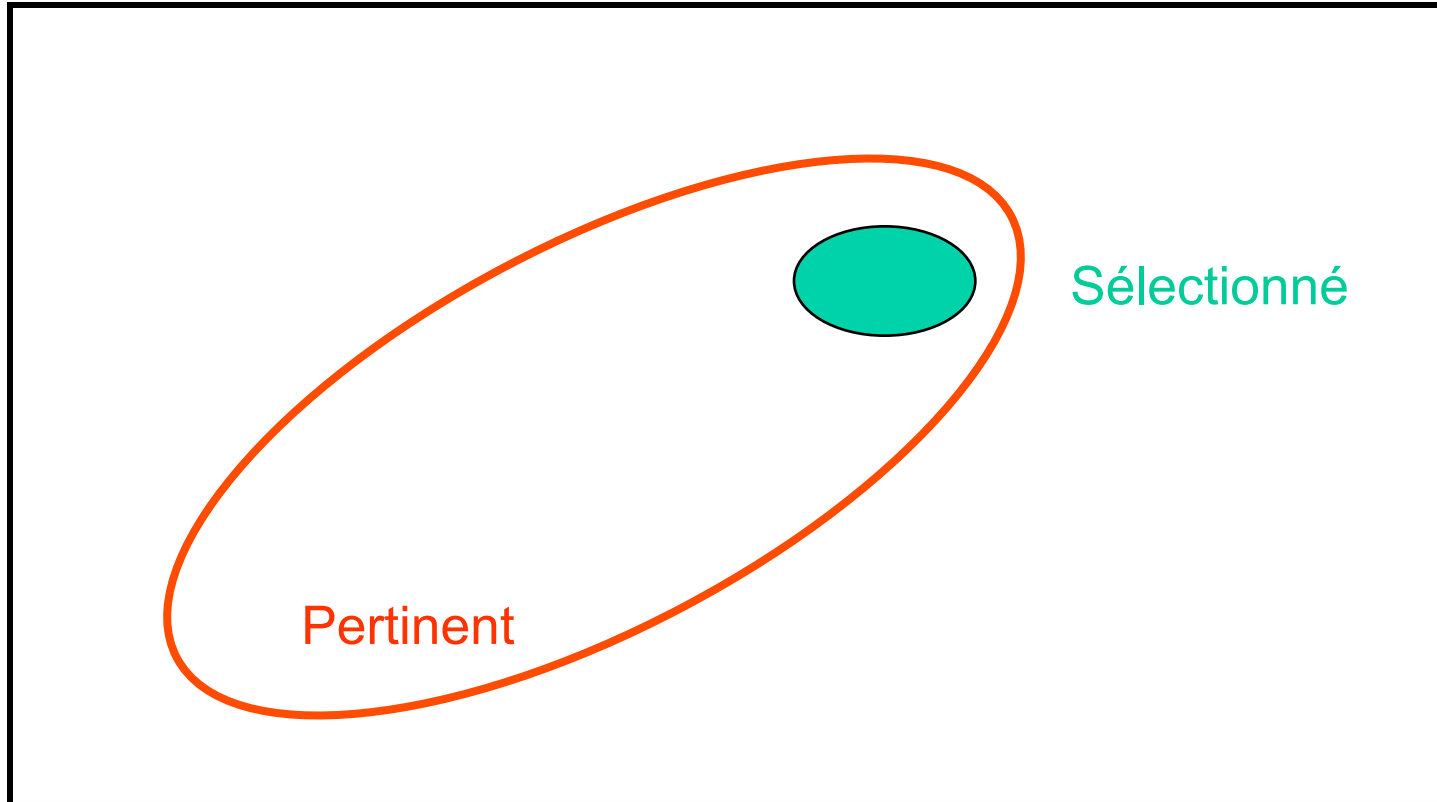




# Sélectionné vs. Pertinent

---

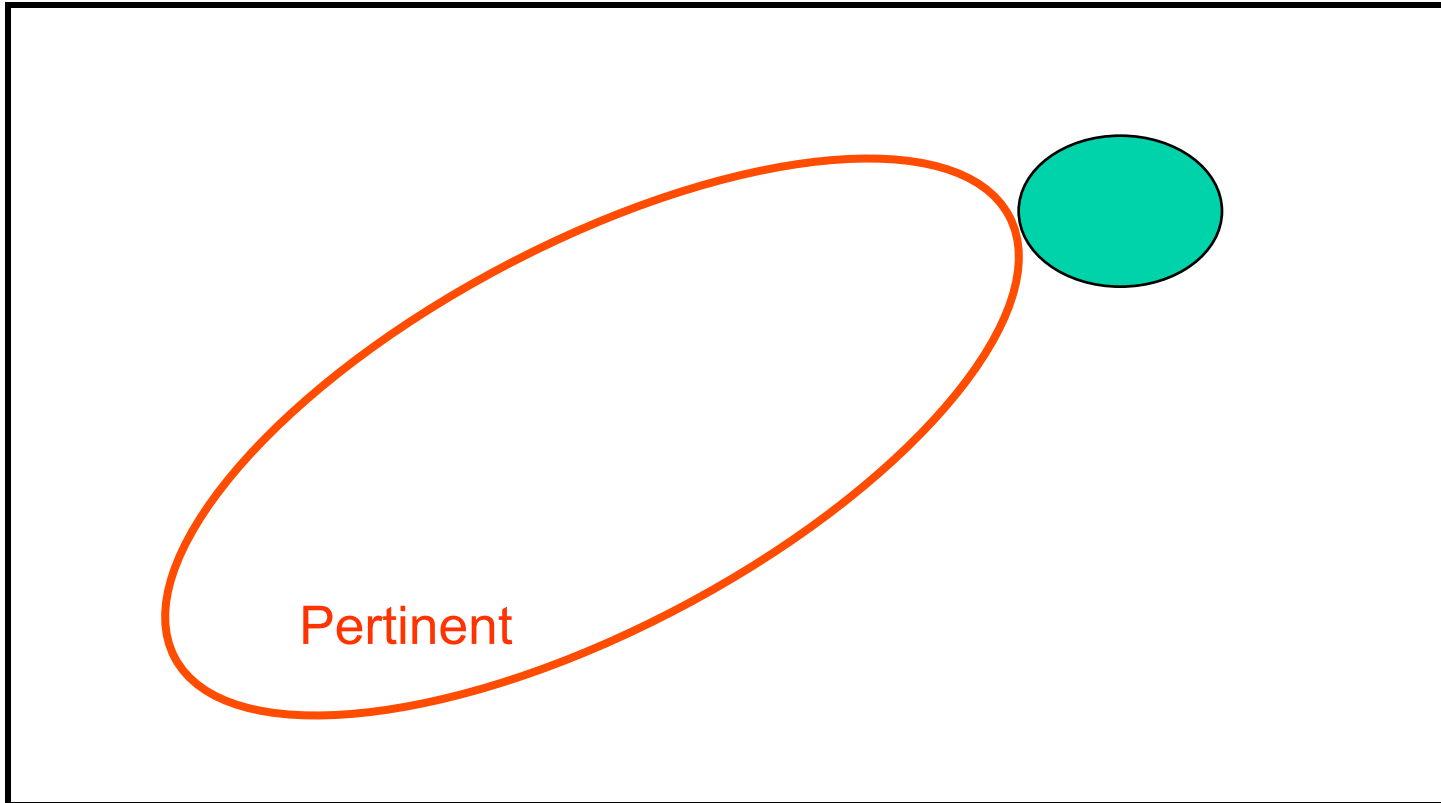
Précision très élevée, rappel très faible



# Sélectionné vs. Pertinent

---

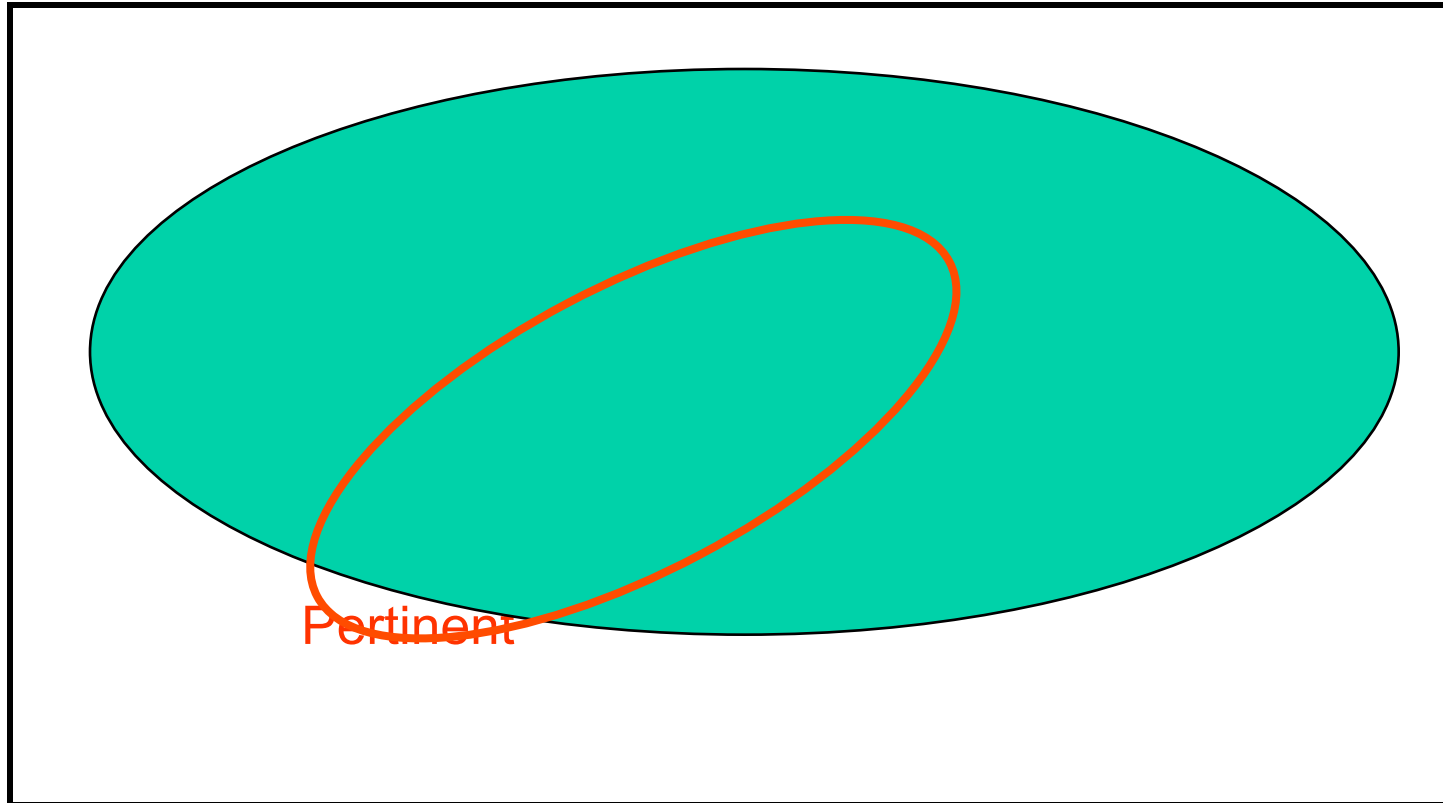
Précision très faible, rappel très faible (en fait, 0)



# Sélectionné vs. Pertinent

---

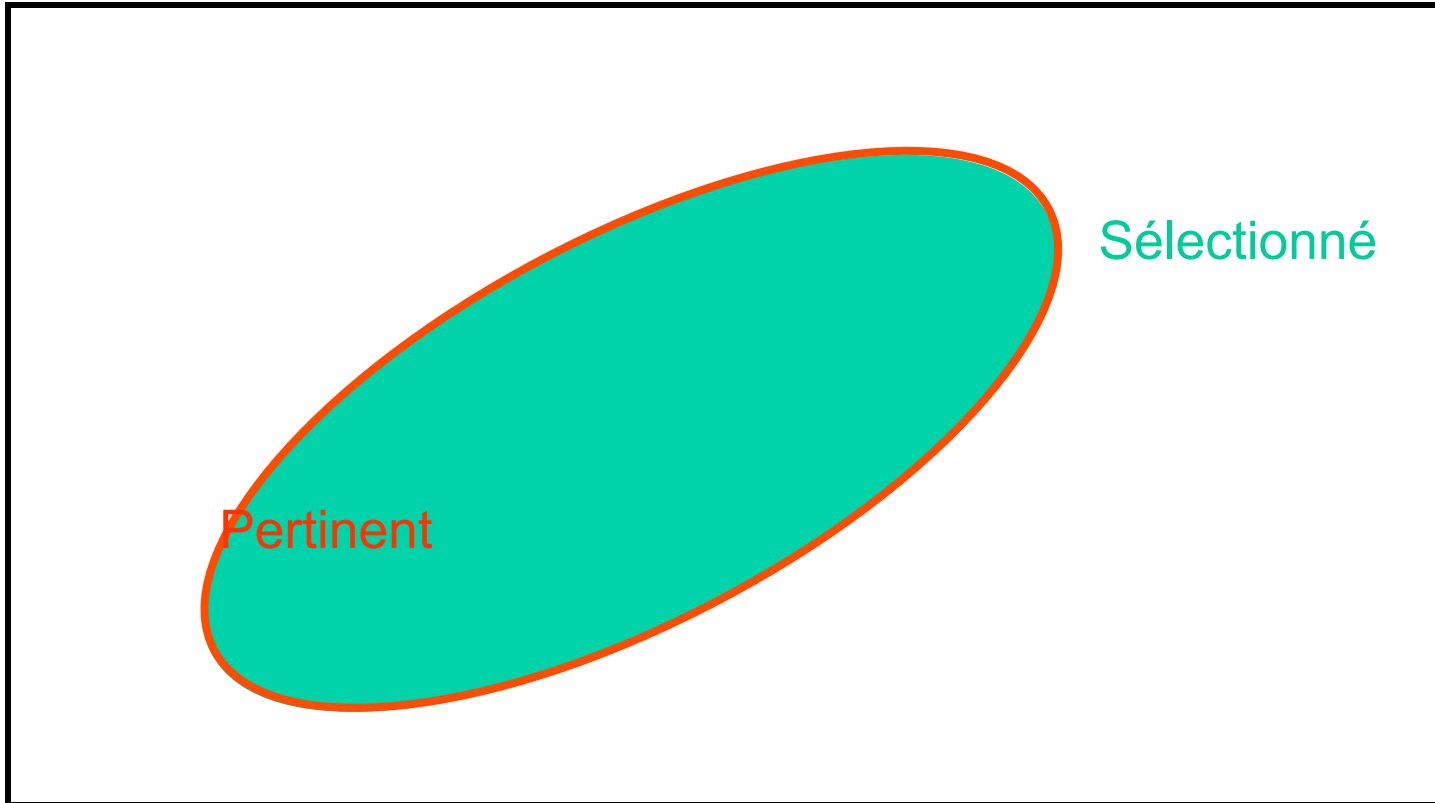
Rappel élevé, mais précision faible



# Sélectionné vs. Pertinent

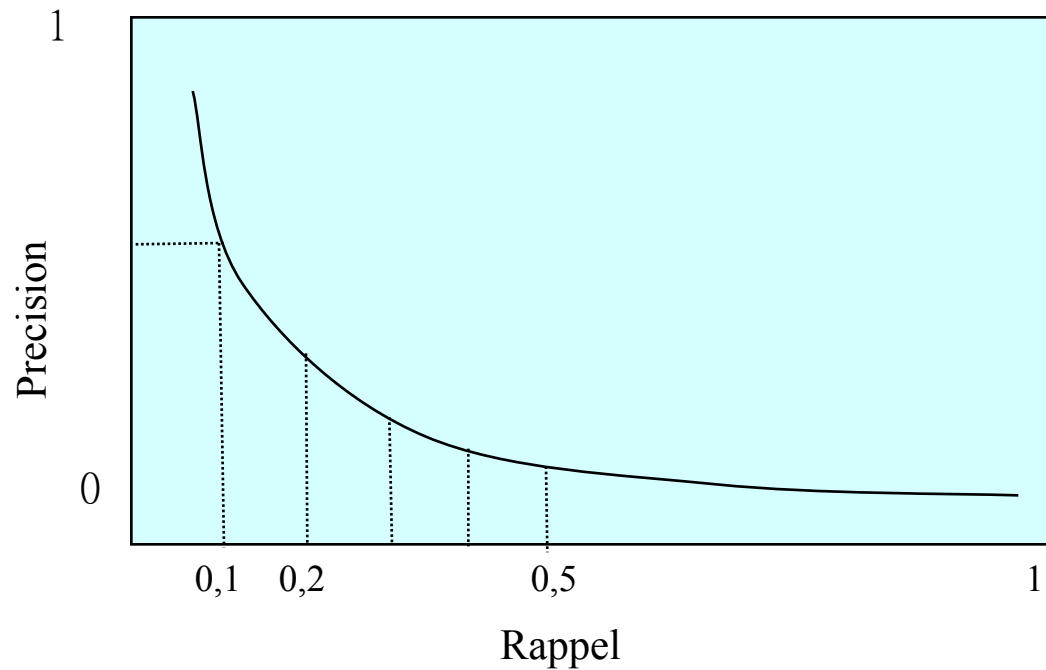
---

Précision élevée, rappel élevé (idéal, mais difficile)



# Lien entre Rappel et Précision

---



Précision moyenne : une seule valeur reliant le rappel et précision

# Démarche d'évaluation

---

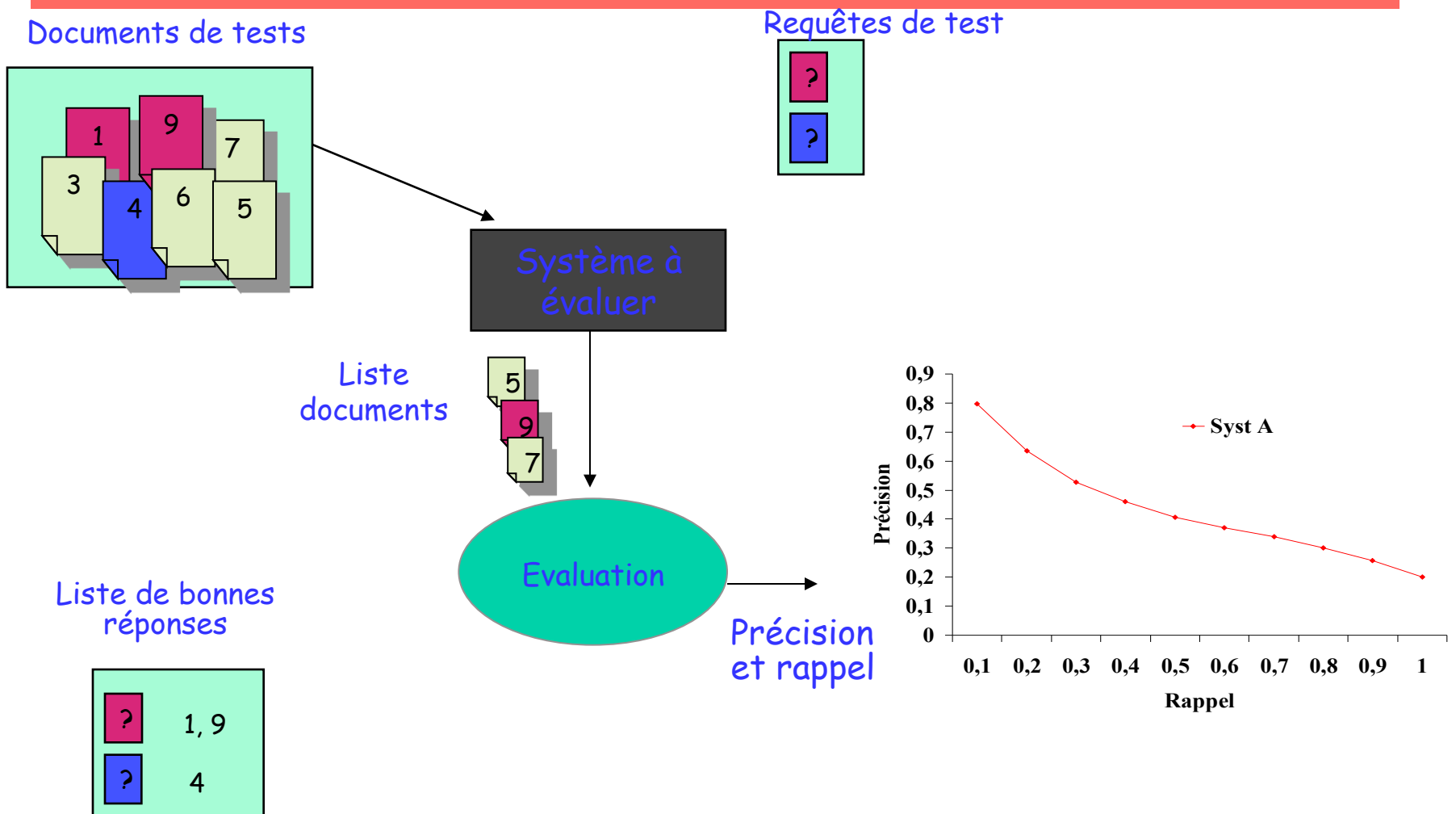
- **Démarche Analytique (formelle)** :
  - Difficile pour les SRI, car plusieurs facteurs : pertinence, distribution des termes, etc. sont difficiles à formaliser mathématiquement
- **Démarche Expérimentale**
  - par « **benchmarking** ».
  - Evaluation effectuée sur des collections de tests
  - Collection de test : un ensemble de documents, un ensemble de requêtes et des pertinences (réponses positives pour chaque requêtes)

# Démarche expérimentale

---

- Lancée dès les années 1960, par Cleverdon, dans le cadre du projet Cranfield
- Objectif du projet Cranfield
  - Construire des collections de test
  - Evaluer les systèmes sur ces collections de test
- Evaluation à la Cranfield

# Evaluation à la Cranfield





# Test Collections

---

- 

Collection Name	Number Of Documents	Number Of Queries	Raw Size (Mbytes)
CACM	3,204	64	1.5
CISI	1,460	112	1.3
CRAN	1,400	225	1.6
MED	1,033	30	1.1
TIME	425	83	1.5

---

- COLLECTION TREC

---

# Calcul du rappel et la précision

# Calcul du rappel et de la précision

---

- On suppose qu'on dispose d'une collection de test
  - Lancer chaque requête sur la collection de test
  - Marquer les documents pertinents par rapport à la liste de test.
  - Calculer le rappel et la précision à pour chaque document pertinent de la liste.

# Calcul du rappel et de la précision

## Exemple

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Le nombre total de documents pertinents est = 6

$R=1/6=0.167; P=1/1=1$

$R=2/6=0.333; P=2/2=1$

$R=3/6=0.5; P=3/4=0.75$

$R=4/6=0.667; P=4/6=0.667$

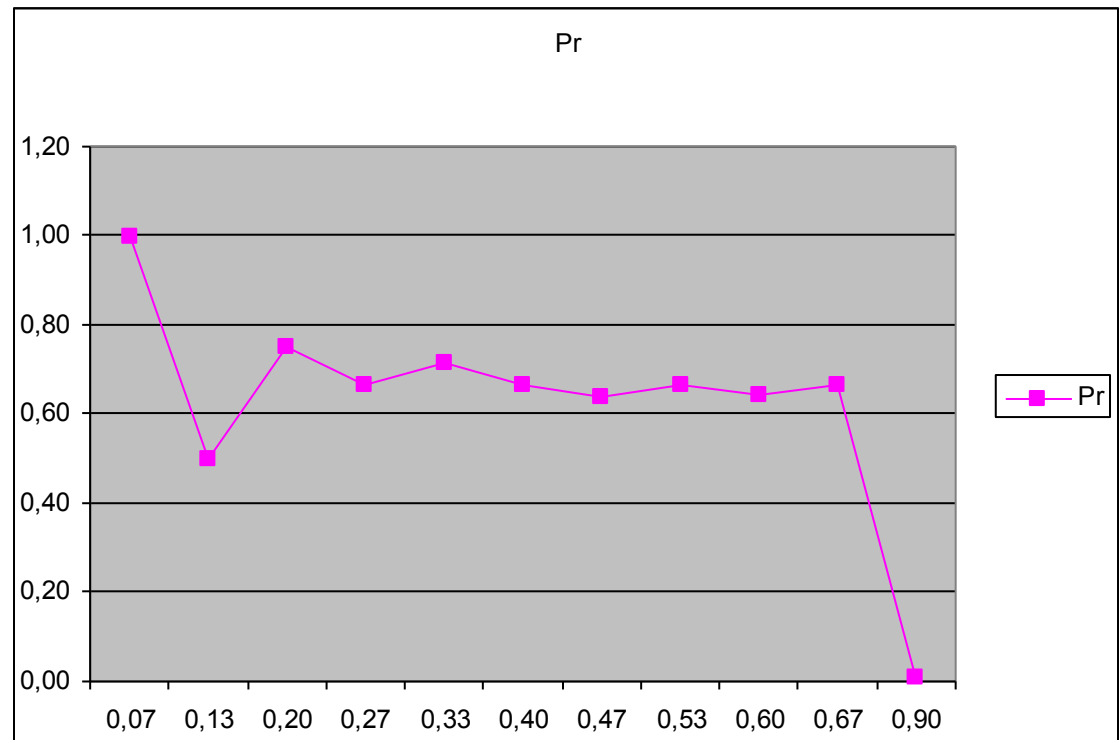
$R=5/6=0.833; p=5/13=0.38$

Il manque un document pertinent.  
On n'atteindra pas le 100% de rappel

# Calcul du rappel et de la précision

## Exemple 2

Ra	Pr
0,07	1,00
0,13	0,50
0,20	0,75
0,27	0,67
0,33	0,71
0,40	0,67
0,47	0,64
0,53	0,67
0,60	0,64
0,67	0,67
0,90	0,01



# Interpolation de la courbe

## Rappel/Précision

---

- Interpoler une précision pour chaque point de rappel :
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- La précision interpolée au point de rappel  $r_j$  est égale à la valeur maximale des précisions obtenues aux points de rappel  $r$ , tel que  $r \geq r_j$

$$P(r_j) = \max_{r \geq r_j} P(r)$$

# Exemple Interpolation des Précisions

---

Ra	Pr
0,07	1,00
0,13	0,50
0,20	0,75
0,27	0,67
0,33	0,71
0,40	0,67
0,47	0,64
0,53	0,67
0,60	0,64
0,67	0,67
0,90	0,01

Ra	Pr
0,0	
0,1	
0,2	
0,3	
0,4	
0,5	
0,6	
0,7	
0,8	
0,9	
1	

# Précision moyenne

---

- On souhaite souvent avoir une valeur unique
  - Par exemple pour les algorithmes d'apprentissage pour contrôler l'amélioration
- La précision moyenne est souvent utilisée en RI
- Plusieurs moyennes
  - Précision moyenne non interpolée (PrecAvg) :
    - Calculer la moyenne des précisions à chaque apparition d'un document pertinent



# Précision moyenne non interpolée

## Exemple

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Le nombre total de document pertinent est = 6

$R=1/6=0.167; P=1/1=1$

$R=2/6=0.333; P=2/2=1$

$R=3/6=0.5; P=3/4=0.75$

$R=4/6=0.667; P=4/6=0.667$

$R=5/6=0.833; p=5/13=0.38$

$AP'' = AvgPrec = (1+1+0,75+0,667+0,38)/6$

# Autres mesures de moyennes

---

- F-Mesure
  - Mesure tenant compte à la fois du rappel et de la précision.
  - Introduite par van Rijbergen, 1979
  - Moyenne harmonique entre R et P

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

# Autres mesures de moyennes

---

- E-Mesure (F-Mesure paramétrique)
  - Une variante de F-Mesure qui tient compte du poids accordé à la précision vis-à-vis du rappel

$$E = \frac{(1 + \beta^2)PR}{\beta^2 P + R} = \frac{(1 + \beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$$

- $\beta$  contrôle le compromis R, P:
  - $\beta = 1$ : même poids précision et recall (E=F).
  - $\beta > 1$ : privilégie la précision au rappel
  - $\beta < 1$ : plus d'importance au rappel.

# Exemple de résultats renvoyés par le Programme TREC\_EVAL

---

Total number of documents over all queries

Retrieved: 1000

Relevant: 80

Rel\_ret: 30

Interpolated Recall - Precision Averages:

at 0.00 0.4587

at 0.10 0.3275

at 0.20 0.2381

at 0.30 0.1828

at 0.40 0.1342

at 0.50 0.1197

at 0.60 0.0635

at 0.70 0.0493

at 0.80 0.0350

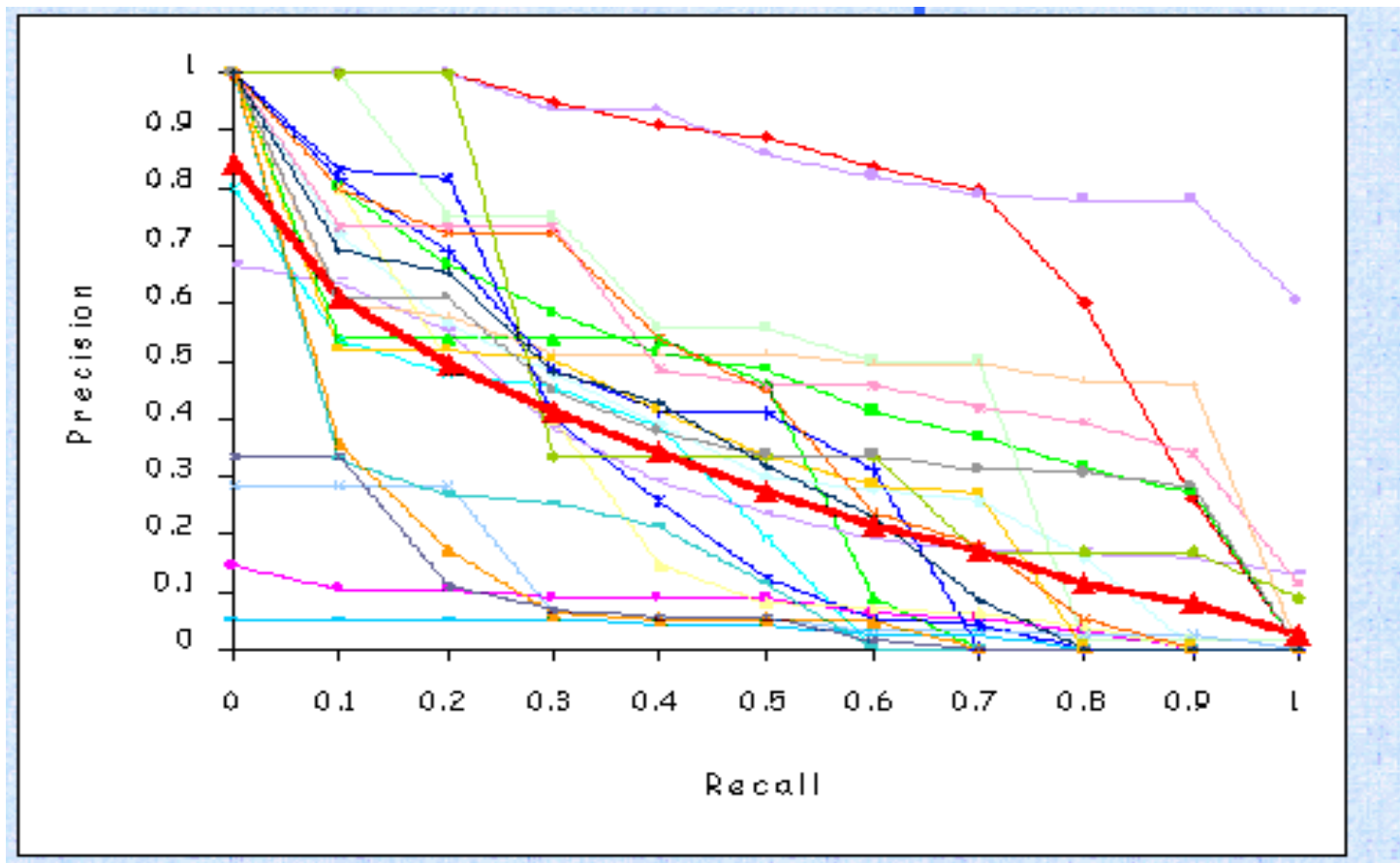
at 0.90 0.0221

at 1.00 0.0150

Average precision (non-interpolated) for all rel docs:

0.1311

# R-P courbes sur l'ensemble des requêtes



Illisible, difficile de comparer deux approches/systèmes requête par requête  
On a besoin d'une moyenne entre les requêtes

# Moyenne sur plusieurs requêtes

---

- Deux façons de calculer la moyenne
  - Micro-moyenne – chaque document pertinent est un point de la moyenne
  - Macro-moyenne – faire la moyenne par requête
- On calcule également la moyenne des précisions moyennes

# Courbe des moyennes sur plusieurs requêtes

---

- **Macro moyenne**
  - Calculer la précision moyenne à chaque point de rappel pour l'ensemble des requêtes.
  - Tracer la courbe rappel-précision

# Exemple

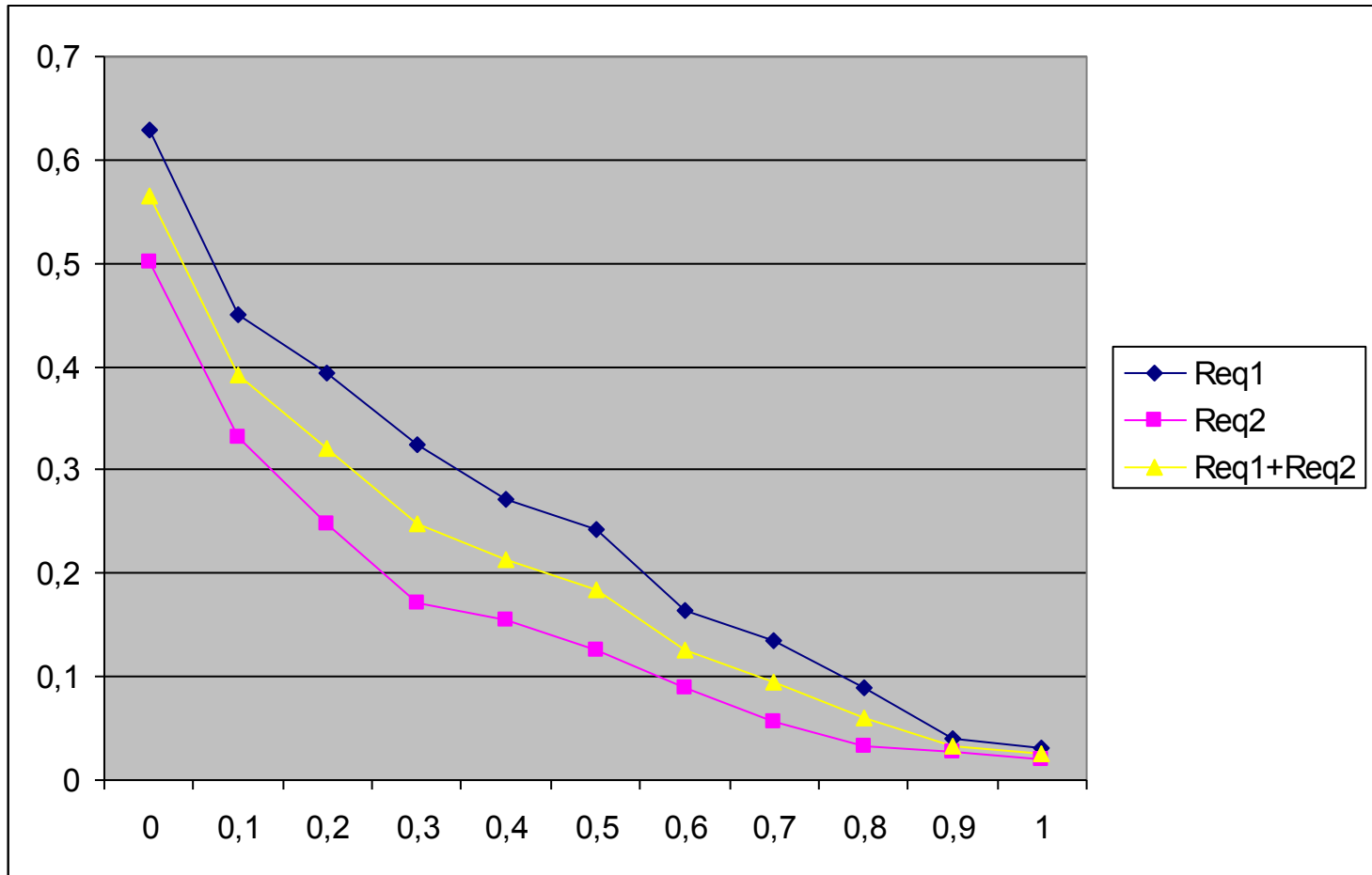
Requete1	
R	Pr
0	0,629
0,1	0,451
0,2	0,393
0,3	0,3243
0,4	0,271
0,5	0,2424
0,6	0,164
0,7	0,134
0,8	0,09
0,9	0,04
1	0,031
<b>AvrPrec</b>	<b>0,2329</b>

Requete2	
R	Pr
0	0,5017
0,1	0,332
0,2	0,248
0,3	0,171
0,4	0,155
0,5	0,125
0,6	0,089
0,7	0,056
0,8	0,032
0,9	0,027
1	0,02
<b>AvrPrec</b>	<b>0,1443</b>

Ens des requêtes	
R	Pr
0	0,56535
0,1	0,3915
0,2	0,3205
0,3	0,24765
0,4	0,213
0,5	0,1837
0,6	0,1265
0,7	0,095
0,8	0,061
0,9	0,0335
1	0,0255
<b>AvrPrec</b>	<b>0,1886</b>

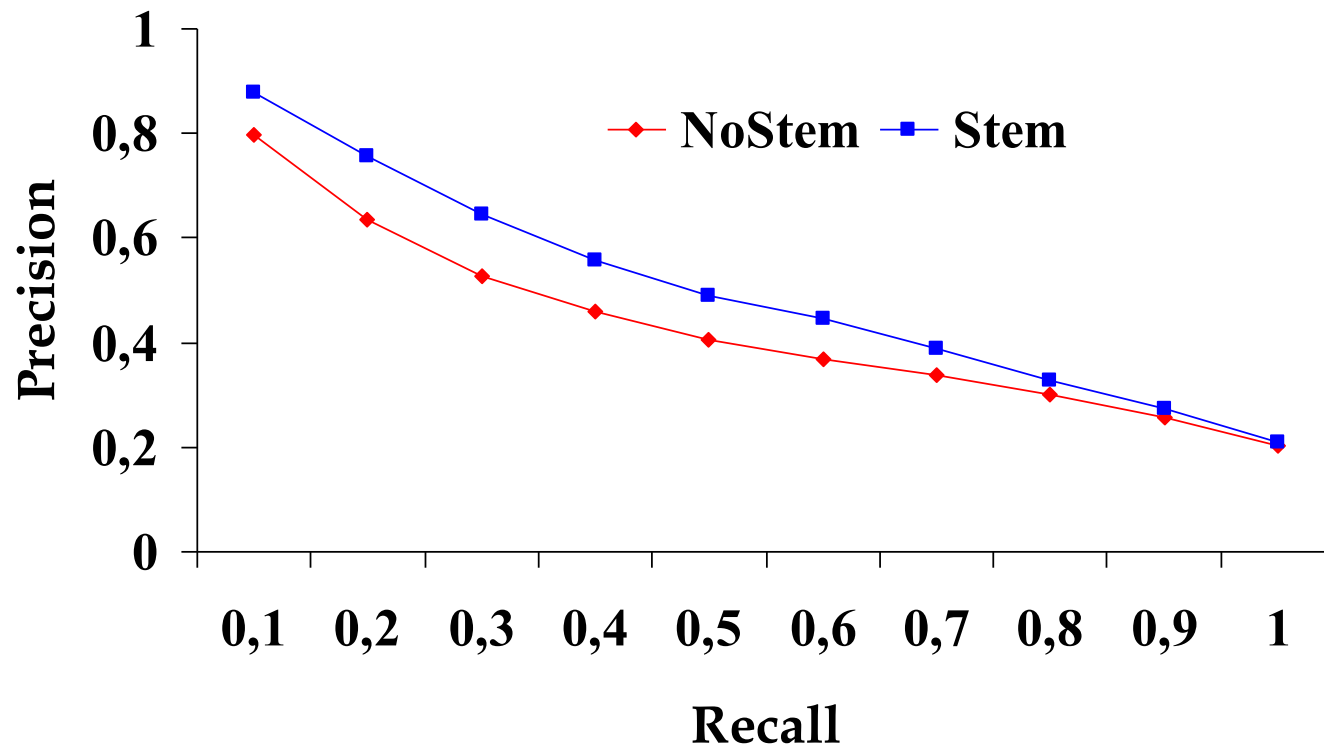


# Exemple



# Comparaison de deux systèmes sur un ensemble de requêtes

---



# Mesures focalisées sur le “top” de la liste

---

- Les utilisateurs se focalisent davantage sur les documents pertinents se trouvant en “top” des résultats
- La mesure de rappel  $n'$  est pas toujours appropriée
  - Il existe des stratégies de recherche pour lesquelles il y a une réponse unique
  - e.g., navigational search, question answering
- Solution : mesurer plutôt la capacité d'un SRI à trouver les documents pertinents en top de la liste

# Mesures focalisées sur le “top” de la liste

---

- Precision au Rang  $X$  (Precision at rank  $X$ )
  - $X = 5, 10, 20$
- Discounted Cumulative Gain
  - Prise en compte de la pertinence graduelle des documents
  - Les documents très pertinents sont plus utiles que ceux qui sont marginalement pertinents
- Reciprocal Rank
  - Rang inverse du premier document pertinent sélectionné

# Précision à X documents

- Précision à différent niveau de documents
  - Précision calculée à 5 docs, 10 docs, 15docs, ...

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Prec. à 5 docs = 3/5

Prec. à 10 docs = 4/10

# R- Précision

- Une façon de calculer une valeur de précision unique : précision au R ème document de la liste des documents sélectionné par la requête ayant R documents pertinents dans la collection.

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

$$R = \# \text{ documents pertinents} = 6$$

$$\text{R-Precision} = 4/6 = 0,66$$

# Exemple R-précision

---

	Précision
at 5 docs	0,224
at 10 docs	0,177
at 15 docs	0,142
at 30 docs	0,114
at 100 docs	0,073
at 200 docs	0,053
at 500 docs	0,013
R-précision= Précision Exacte	0,144

# Discounted Cumulative Gain

---

- Deux hypothèses:
  - Prise en compte de la pertinence graduelle des documents
  - Les documents très pertinents sont plus utiles que ceux qui sont marginalement pertinents
- Plus un document pertinent est loin du début de la liste moins il est utile pour l'utilisateur, car il a peu de chance d'être examiné



# Discounted Cumulative Gain

---

- Utilise la pertinence graduelle comme une mesure de l'utilité, ou du gain, obtenu en examinant un document
- Le gain est accumulé en commençant par le haut du classement et il est réduit (diminué) au fur et à mesure on l'on va vers le fond de la liste
- La réduction peut être de  $1/\log(\text{rang})$ 
  - Avec un log base 2, une réduction au rang 4 serait de  $1/2$ , au rang 8 de  $1/3$

# Discounted Cumulative Gain

---

- Le gain cumulé au rang  $p$

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i}$$

- Autre formulation

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log(1+i)}$$

# DCG Exemple

---

- Soit une liste de 10 documents jugés sur une échelle de 4 : 0, 1, 2, 3:  
3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- Gain réduit (DG) =  $\text{rel}/\log(\text{rank})$   
3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0  
= 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:  
3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

# DCG normalisé

---

- Moyenne des DCG sur un ensemble de requêtes
  - e.g., DCG au 5 est 6.89 et 9.61 rang 10
- Les valeurs de DCG sont souvent normalisées *selon la valeur DGC du classement parfait (Ideal DCG)*

# NDCG Exemple

---

- Classement parfait:

3, 3, 3, 2, 2, 2, 1, 0, 0, 0

- DCG idéal :

3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10

- NDCG (valeurs de DCG normalisées)

- $NDCG_p = DCG_p / iDCG_p$

- 1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

- $NDCG \leq 1$

# Retour sur la comparaison de systèmes

---

- L'évaluation en RI est comparative
  - Vérifier si le système A est meilleur que le système B ?
  - Quelle est la démarche ?
    - Comparer les performances en termes de (précisions moyennes, R, F) des deux systèmes
      - $(\text{Val}(A) - \text{Val}(B)) / \text{Val}(B) * 100$
      - .... partir de 5% on peut considérer que A est meilleur que B
    - Comparer les courbes R/P
      - La courbe de A est toujours supérieure à celle de B
- Que se passe t-il quand on change de collection?

# Retour sur la comparaison de systèmes : test statistiques

---

- Vérifier si la différence de performances entre deux systèmes est significative → Test statistique
- Test de “significativité” permet de
  - Rejeter l’hypothèse nulle (pas de différence entre A et B)
  - Accepter l’hypothèse alternative (A et B sont différents)
- Plusieurs facteurs :
  - t-test, test de wilcoxon, Kendall (tau)

# t-Test

---

- Calculer la performance (par exemple P@10, MAP, ...) pour chaque requête
- Calculer la différence entre les valeurs de de A et B
- On calcule

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} \cdot \sqrt{N}$$

*Degré de liberté (v=N-1)  
N=taille de l'échantillon)  
à lire dans la table de la loi de student*

- H0 est rejetée à  $\alpha$  si  $t_{\text{calc}} \leq t_{(\alpha:v)}$  



# t-Test

---

- Dire qu'une différence est **significative** à
  - $x\%$  si  $p\text{-value} < x/100$
  - Valeurs de  $x$  considérées
  - 0.05 (ou 5%) signifie : il y a 95% de chances que la différence ne soit pas due au hasard
  - 0.01 (ou 1%) signifie : il y a 99% de chances que la différence ne soit pas due au hasard
  - 0.10 (ou 10%) signifie : il y a 90% de chances que la différence ne soit pas due au hasard

# Exemple

---

Query	A	B	B-A
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

# Exemple : t-Test

---

- Hypothèse Null : la moyenne de la distribution de ces différences est zéro
- Calculer t

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} \cdot \sqrt{N}$$

## – Exemple

$$\overline{B-A} = 21.4, \sigma_{B-A} = 29.1, t = 2.33, \text{p-value} = .02$$

*Chercher la valeur dans la table, alpha=98%,  $t_{table\_student}(98,9)=2.821$*

*$t_{cal} < t_{table}$*

*P-value=0.02*

*L'hypothèse est rejetée à 98% donc les deux distributions sont différentes*

# Questions

---

- Comment construire une collection de test ?
  - Quels / combien de documents ?
  - Quelles / combien de requêtes ?
  - Comment identifier les documents pertinents pour chaque requête ?
- Evaluer la validité de la collection

# Comment identifier les documents pertinents ?

---

- Pour répondre d'une façon sûre, il faut
  - Juger tous les documents de la collection pour chaque requête
  - Qui juge ?
    - Humain : 1, 2 .. n personnes
    - Faisable pour des petites collections
    - Impossible sur des collections volumineuses
      - TREC collections ont plus d'un millions de documents

# Comment identifier les documents pertinents ?

---

- Autres approches
  - Pooling
  - Sampling

# Comment identifier les documents pertinents ?

---

- Pooling
  - Pour chaque requête
    - Sélectionner des documents en utilisant différents techniques
    - Juger les n meilleurs documents obtenus par chaque technique
    - La liste des documents pertinents = l'union des documents pertinents d de chaque technique
    - Sous ensemble de vrai jugement de pertinence
- Echantillonnage
  - Possible d'estimer le nombre de documents pertinents par des techniques d'échantillonnage
- Incomplète, problème ?
  - Comment doit-on traiter les documents non jugés
  - Comment ceci peut affecter les performances calculées ?

# Avantages et inconvénients des collections de tests

---

- **Avantages**
  - Mesures de performances
  - Possibilité de comparaison avec d'autres travaux
- **Inconvénients**
  - Les résultats obtenus sont propres à la collection.
  - Ne répondent pas à toutes les tâches de RI, notamment celles orientées utilisateur



---

TREC

Expérience de TREC

# TREC

## The Text REtrieval Conference

---

- Competition/collaboration between IR research groups worldwide
- Run by NIST, just outside Washington DC
- Common tasks, common test materials, common measures, common evaluation procedures
- Now various similar exercises (CLEF, INEX, NCTIR etc.)

# The TREC Benchmark

---

- TREC: **T**ext **RE**trieval **C**onference (<http://trec.nist.gov/>)  
Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).
- Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.
- Uniform, appropriate scoring procedures

# The TREC Objective

---

- TREC is a modern example of the Cranfield Tradition
  - System evaluation based on text collections
- Sharing of resources and experiences in developing the benchmark.
  - With major sponsorship from government to develop large benchmark collections.
- Encourage participation from industry and academia.
- Development of new evaluation techniques, particularly for new applications.
  - Retrieval, routing/filtering, non-English collection, web-based collection, question answering.

# A Brief History of TREC

- 1992: first TREC conference
  - started by Donna Harman and Charles Wayne as 1 of 3 evaluations in DARPA's TIPSTER program
  - first 3 CDs of documents from this era, hence known as the "TIPSTER" CDs
  - open to IR groups not funded by DARPA
    - 25 groups submitted runs
  - two tasks: ad hoc retrieval, routing
    - 2GB of text, 50 topics
    - primarily an exercise in scaling up systems

# A Brief History of TREC

- 1993 (TREC-2)
  - true baseline performance for main tasks
- 1994 (TREC-3)
  - initial exploration of additional tasks in TREC
- 1995 (TREC-4)
  - official beginning of TREC track structure
- 1998 (TREC-7)
  - routing dropped as a main task, though incorporated into filtering track
- 2000 (TREC-9)
  - ad hoc main task dropped; first all-track TREC