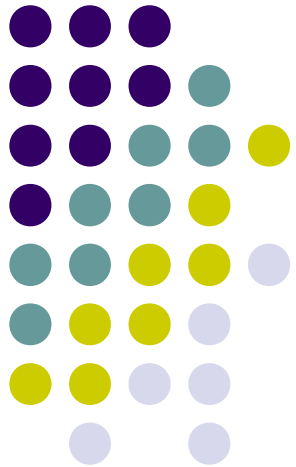
Introduction à la recherche d'information

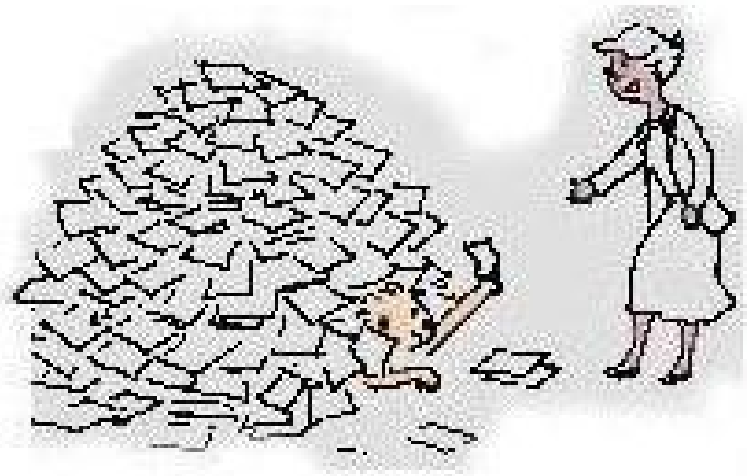


Plan:

- Qu'est ce que la RI ?
- Interrogation
- Indexation
- Modèles de RI
- Visualisation
- Évaluation des performances
- État des lieux et perspectives

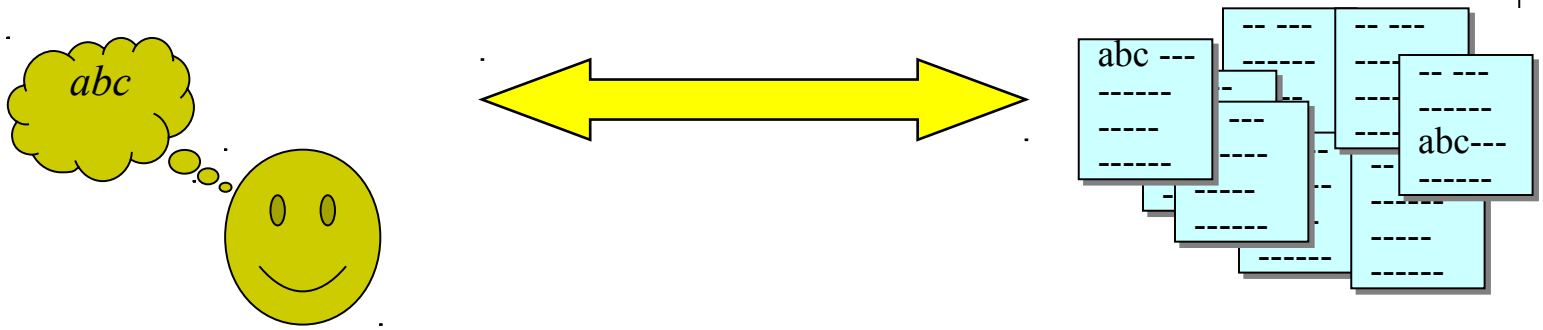
Objectif de la RI

- Sélectionner dans une collection de **documents**
  - Les **informations**
  - … **pertinentes** répondant à des
  - … **besoins en information** d'utilisateurs



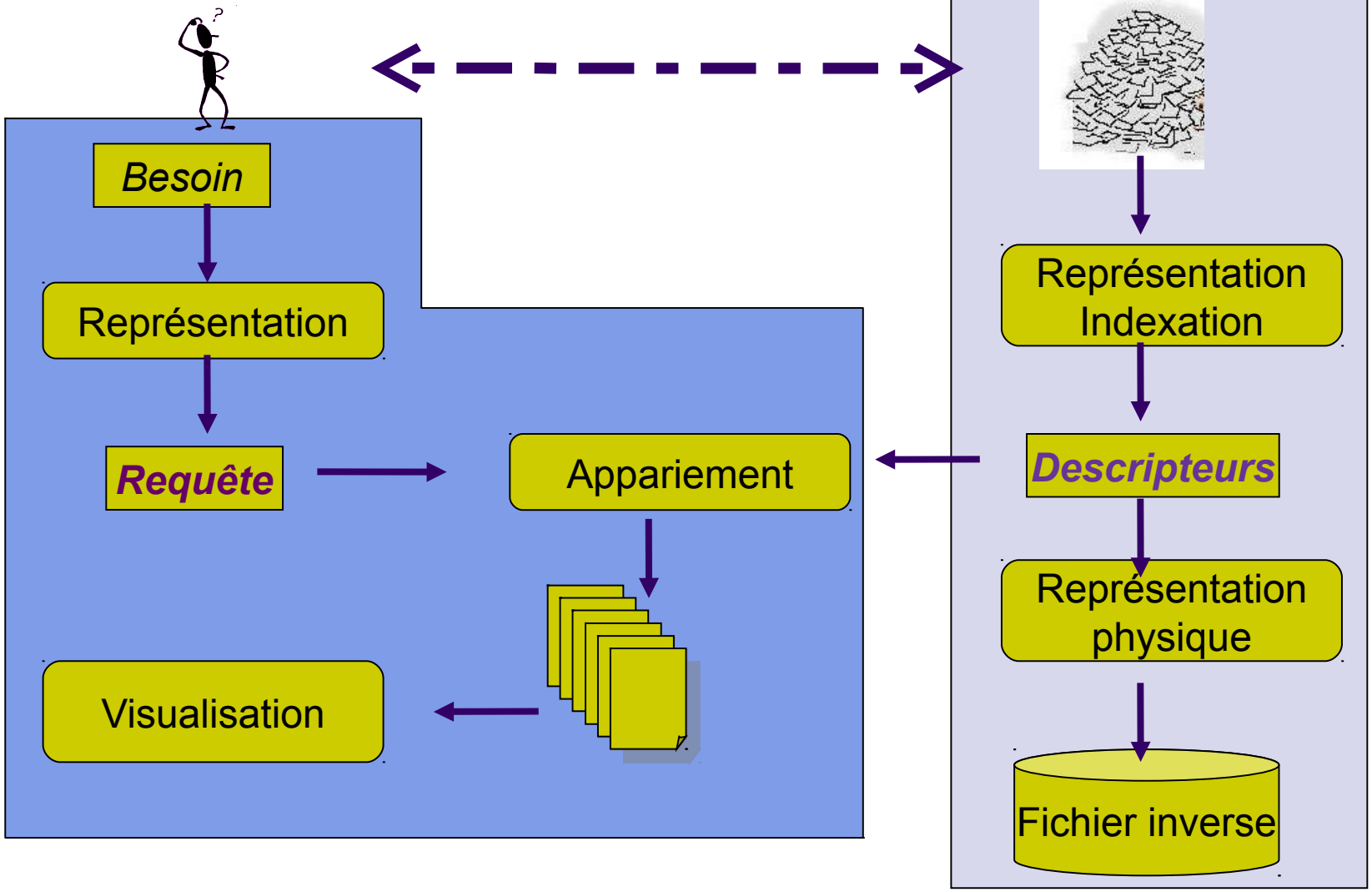
Comment sélectionner l'information pertinente ?

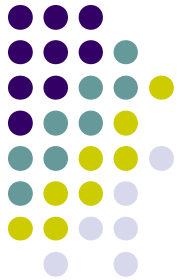
Requête : liste mots clés



- Comment sélectionner les informations répondant à une requête ?
  - Une façon simple consiste à rechercher les mots de la requête dans tous les documents.
  - Solution lourde et pas pratique

Processus de RI



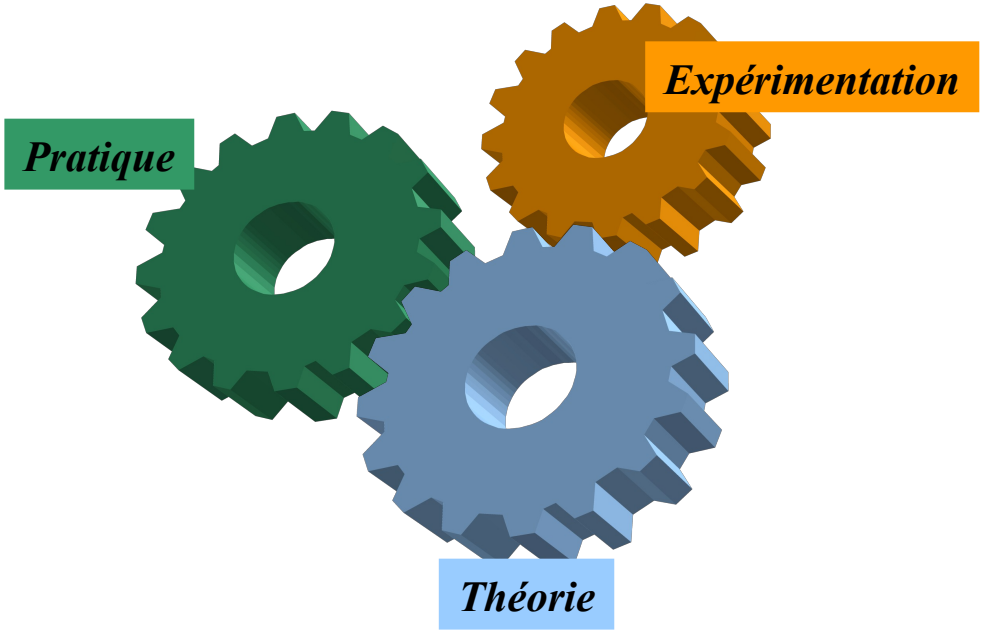


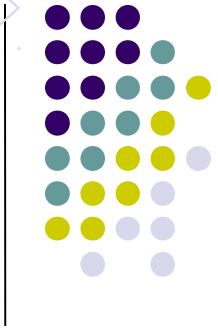
Problématique de la RI

- Représentation de l'information
  - Comment construire une représentation à partir de l'information?
  - Qu'est-ce qu'une « bonne » représentation ?
  - Quelle organisation physique pour ces index ?
- Représentation des besoins
  - Comment exprimer le besoin (langage de requêtes)?
  - Comment représenter le besoin?
- Comparaison des représentations
  - Comment mesurer (décider) la pertinence d'un document ?
- Évaluation des performances
  - Comment décider que l'approche A est mieux que B?
  - Quelle démarche ?
  - Quelles métriques ?

Problématique de la RI

- En RI on a besoin de :
  - théorie, pratique et expérimentation.





Indexation automatique

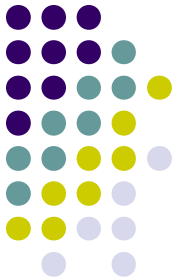
- Indexation = créer une représentation (descripteurs) des documents

- Approches
  - Statistique (distribution des mots) et/ou TALN (analyse du texte)
  - Approche courante est plutôt statistique avec des hypothèses simples
    - Redondance d'un mot marque son importance
    - Cooccurrence des mots marque le sujet d'un document

Indexation automatique: démarche

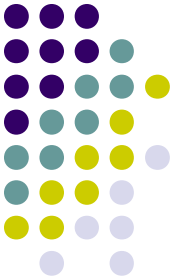
- Etape 1: extraction des termes
- Etape 2: normalisation des termes
  - Regrouper les variantes d'un même terme
- Etape 3: pondération
  - Discrimination entre
    - les termes clés/ importants/ significatifs
    - et les autres
- Etape 4: construction du fichier inverse

Indexation automatique
Etape 1: extraction des mots



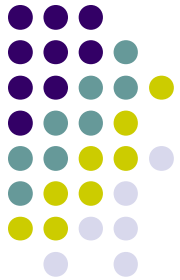
- Extraire les mots clés
  - Mot simple ou composé
  - Mot: suite de caractères séparés par blanc (ou signe de ponctuation, caractères spéciaux,…), nombres
- Dépend de la langue
  - Langue française
    - Pomme de terre ? Un mot clé, deux mots clés ou trois ?
  - Langue allemande
    - Les mots composés ne sont pas segmentés
      - Lebensversicherunggesellschaftsangesteller
      - ‘employé d’une compagnie d’assurance-vie’

Etape 1: Extraction des mots (suite)



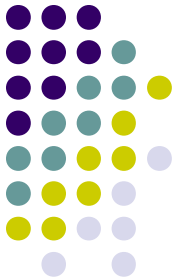
- Suppression des mots « vides » (Stop list)
  - Mots trop fréquents mais pas utiles
  - Exemples:
    - Anglais: the, or, a, you, I, us,…
    - Français: le, la , de, des, je, tu,…

- Attention à:
  - US: « USA », « give us information »
  - A (de vitamine A)



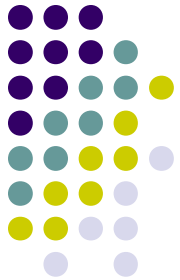
Etape 2: normalisation

- Lemmatisation / (radicalisation/racinisation) / stemming
  - Processus morphologique permettant de regrouper les variantes d'un mot
    - Ex: économie, économiquement, économiste -> économ
    - Pour l'anglais: retrieve, retrieving, retrieval, retrieved, retrieves -> retriev



Etape 2: normalisation

- Utilisation de règles de transformations
  - Règle de type: condition action
    - Ex: si le mot se termine par s supprimer la terminaison
  - Technique utilisée principalement pour l'anglais
    - L'algorithme le plus connu est celui de Porter
- Analyse grammaticale
  - Utilisation de lexique (dictionnaire)
  - Tree-tagger (gratuit sur le net)
- Troncature
  - Pour le français : tronquer à 7 caractères…



Etape 3: pondération des mots

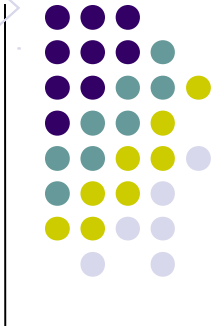
- Comment caractériser l'importance des termes dans un document ?
  - Associer un (ou plusieurs) poids à un terme
  - Idée sous-jacente:
    - Les termes importants doivent avoir un poids fort

Etape 3: Pondération: TF.IDF

- ***TF (Term Frequency):***
  - Idée sous-jacente: plus un terme est fréquent dans un document plus il est important dans la description de ce document
  - Exemple de TF: fréquence du terme dans le doc
  - Robertson TF: tf/ (K+tf)
    - K introduit pour tenir compte de la longueur des documents
    - TF=fréq/ (fréq+0.5+1.5*(longueur_doc/longueur_moy_doc))

Etape 3: Pondération: TF.IDF

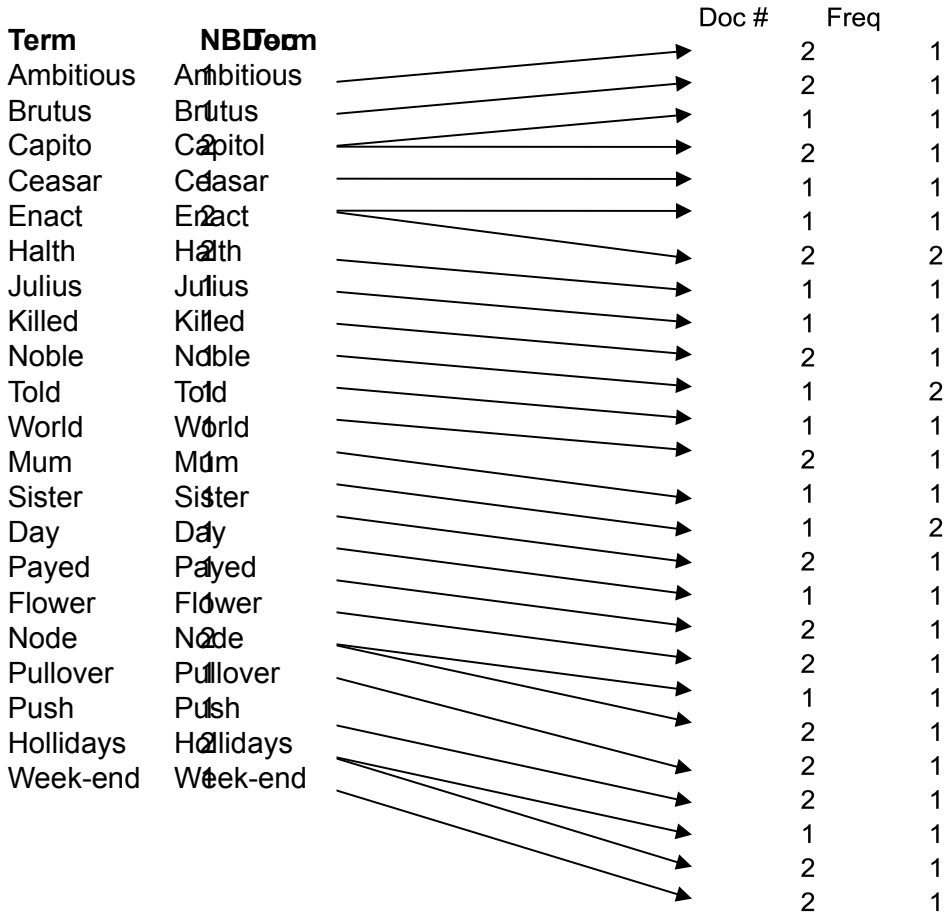
- ***IDF (Inverse Document Frequency)***
  - Idée sous-jacente: plus un terme est fréquent dans une collection, moins il est important dans la description de ce document
  - Log (N/ni)
    - Avec:
      - *N* la taille de la collection
      - *ni* le nombre de documents contenant le terme ti



Indexation automatique

- Une fois les documents indexés :
  - chaque document aura donc un descripteur
    - Liste de mots
    - Fréquence de chaque mot (poids)
    - Exemple : systeme 1, recherc 1, informa 1, documen 3, sri 1, base 1, donnee 1, analyse1, indexer 1, retrouv 1, pertine 1, reponda 2, besoin 3, utilisa 1
  - Ces termes sont ensuite stockés dans une structure appelée **fichier inverse**

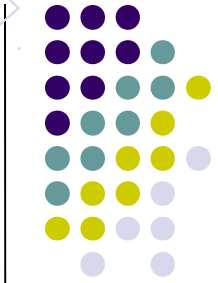
Etape 4: fichier inverse



- La construction d'un fichier inverse est une étape importante
- Elle peut prendre énormément de temps
- Information supplémentaire : position du terme dans le document

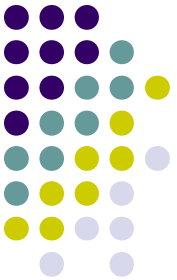
=> gestion des expressions

Généralement stocké dans une BDR



Qualité de l'indexation

- Exhaustive (cf. rappel)
  - Complétude, nombre d'éléments (sujets, concepts) indexés
  - Limiter le silence
- Spécificité (cf. précision)
  - Exactitude (précision) des index
  - Limiter le bruit



Qu'est ce qu'un modèle de RI ?

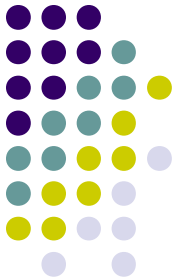
- Un modèle est une abstraction d'un processus (ici recherche d'info)

- Les modèles mathématiques sont souvent utilisés pour
  - formaliser les propriétés d'un processus,
  - élaborer des conclusions, faire des prévisions, etc.

- Les conclusions dérivées d'un modèle dépendent de la qualité du modèle
  - Question : est ce que le modèle est une bonne approximation du processus ?

Qu'est ce qu'un modèle de RI ?

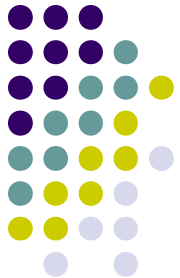
- Modèle de RI est défini par :
  - Représentation des documents
  - Représentation de la requête
  - Mesure de la pertinence requête-document



Pertinence requête-document

- Deux types de mesure
  - Exact Vs. Approché (Exact-Matching ou Best Matching)

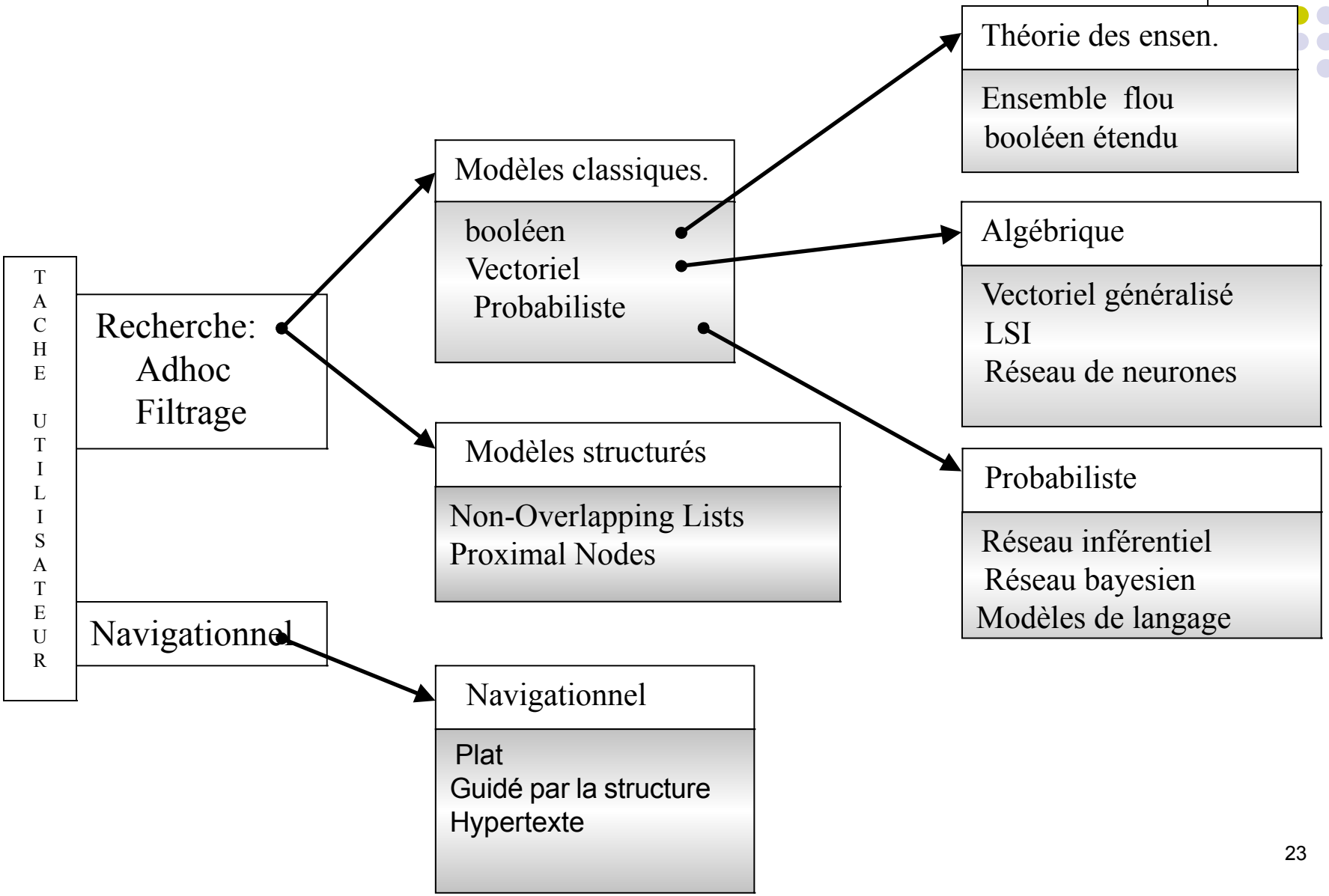
- Appariement exact
  - Sélectionner les documents respectant exactement la requête spécifiée avec des critères précis
- Appariement approché
  - Sélectionner les documents selon un degré de pertinence (calculé)



Concepts de base

- De manière générale, la majorité des approches considère que
  - Chaque document est représenté par une liste de termes d'indexation (mots clés, termes)
  - Les termes n'ont pas la même importance dans un document
  - L'importance d'un terme dans un document est représentée par un poids
- Les modèles de RI diffèrent principalement dans leur manière de mesurer la pertinence requête/document

Modèles de RI



Le Modèle Booléen

- Le premier modèle de RI
- Basé sur la théorie des ensembles
- Un document est représenté par un ensemble de termes
  - Ex : d1(t1,t2,t5); d2(t1,t3,t5,t6); d3(t1,t2,t3,t4,t5)
- Une requête est un ensemble de mots avec des opérateurs booléens : AND (∧), OR(∨), NOT (¬)
  - Ex: $q = t1 \wedge (t2 \vee \neg t3)$
- Appariement Exact basé sur la présence ou l'absence des termes de la requête dans les documents
  - Appariement (q,d) = RSV(q,d)=1 ou 0

Appariement(q,d1)=1

Appariement(q,d2)=0

Appariement(q,d1)=1

Inconvénients du Modèle Booléen

- La sélection d'un document est basée sur une décision binaire
- Pas d'ordre pour les documents sélectionnés
- Formulation de la requête difficile pas toujours évidente pour beaucoup d'utilisateurs
- Problème de collections volumineuses : le nombre de documents retournés peut être considérable

Modèle Vectoriel (VSM)

- Proposé par Salton dans le système SMART (Salton, 1970)

- Idée de base :
  - représenter les documents et les requêtes sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents

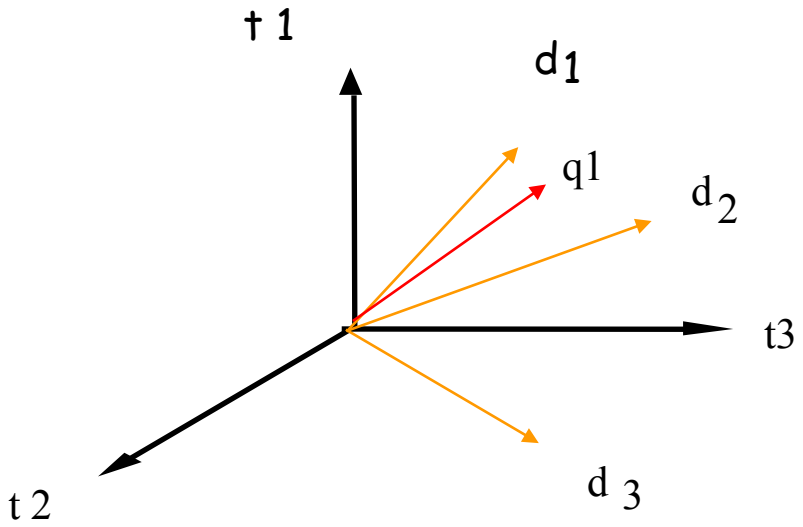
  - Un terme de l'index = une dimension

Modèle Vectoriel (VSM)

- Soit T(t1,t2, …, tM) : ensemble des M termes de la collection

dj= (w1j, w2j, …, wMj)

q= (w1q, w2q, …, wTq)

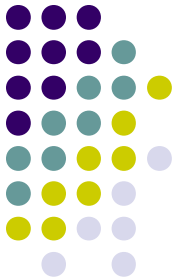


Modèle Vectoriel

- Une collection de *n* documents et *t* termes distincts peut être représentée sous forme de matrice

$$\begin{array}{c|cccc} & T_1 & T_2 & .... & T_t \\ D_1 & w_{11} & w_{21} & \dots & w_{t1} \\ D_2 & w_{12} & w_{22} & \dots & w_{t2} \\ : & : & : & & : \\ : & : & : & & : \\ D_n & w_{1n} & w_{2n} & \dots & w_{tn} \end{array}$$

- La requête est également représentée par un vecteur.

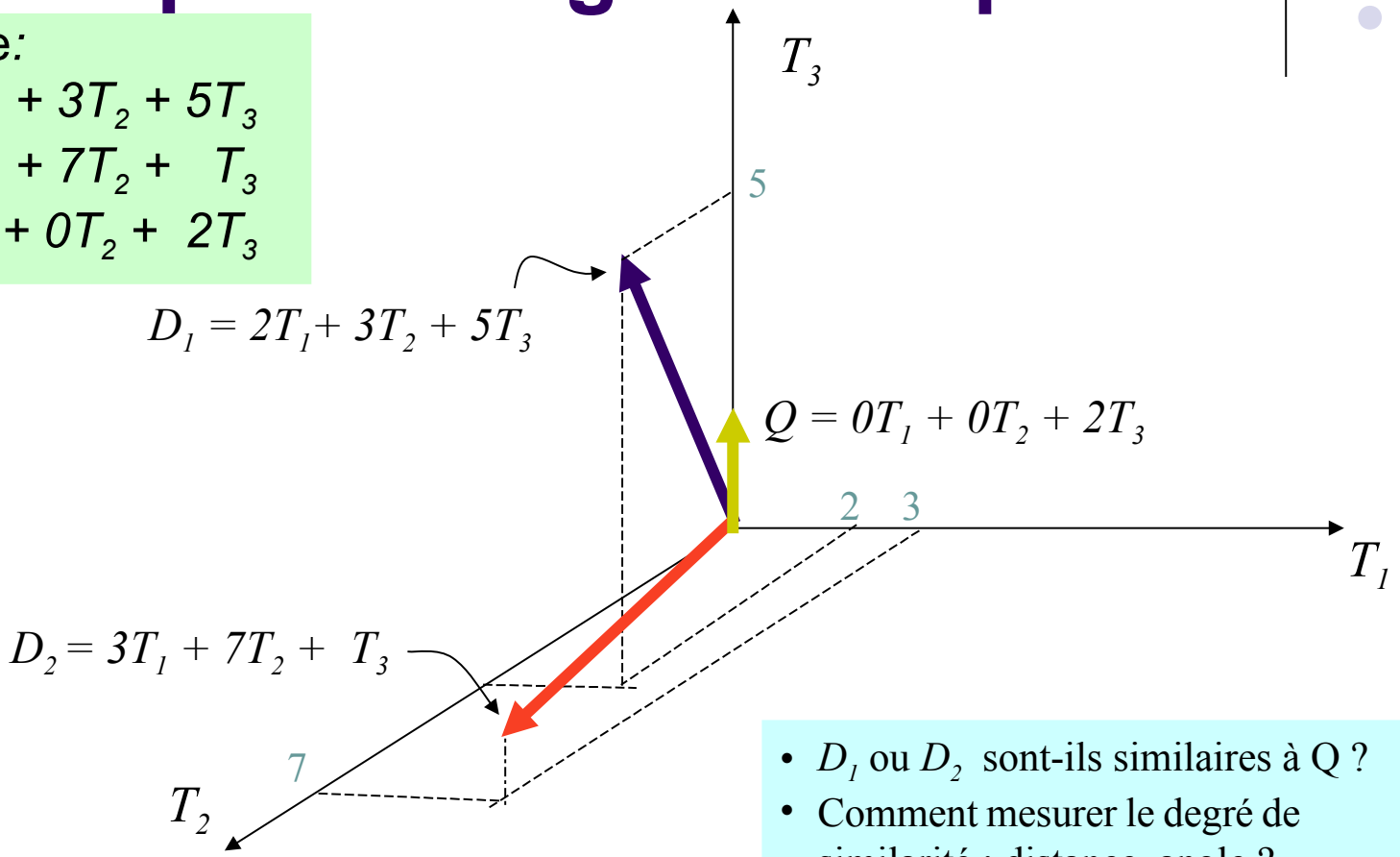


Modèle Vectoriel (VSM)

- Exemple :
  - T(document, web, information, recherche,image,contenu) : ensemble des termes d'indexation
  - d1(document 2,web 1)
  - d2(information 1, document 3, contenu 2)
  - q1 (image web); q2(recherche, documentaire)
  - Représentation vectorielle
    - d1 (2,1,0,0,0,0)
    - d2 (3,0,1,0,0,2)
    - q1 (0,1,0,0,1,0)
    - q2(0,0,0,1,0,0)

Le modèle vectoriel
Interprétation géométrique

Exemple:
$D_1 = 2T_1 + 3T_2 + 5T_3$
$D_2 = 3T_1 + 7T_2 + T_3$
$Q = 0T_1 + 0T_2 + 2T_3$



- D_1 ou D_2 sont-ils similaires à Q ?
- Comment mesurer le degré de similarité : distance, angle ?

La pertinence est traduite en terme de similarité vectorielle :
deux vecteurs sont d'autant plus similaires qu'ils sont proches l'un de l'autre

Le Modèle Vectoriel: mesure de similarité

Soient X et Y deux vecteurs, Sim(X,Y)=

Soient $(x_1,..,x_m)$ les coordonnées de X et $(y_1,..,y_m)$ les coordonnées de Y

Inner product	$\|X \cap Y\|$	$\sum x_i * y_i$
Coef. de Dice	$\frac{2*\|X \cap Y\|}{\|X\|+\|Y\|}$	$\frac{2*\sum x_i * y_i}{\sum x_i^2 + \sum y_j^2}$
Mesure du cosinus	$\frac{\|X \cap Y\|}{\sqrt{\|X\|} * \sqrt{\|Y\|}}$	$\frac{\sum x_i * y_i}{\sqrt{\sum x_i^2 * \sum y_j^2}}$
Mesure du Jaccard	$\frac{\|X \cap Y\|}{\|X\|+\|Y\|-\|X \cap Y\|}$	$\frac{\sum x_i * y_i}{\sum x_i^2 + \sum y_j^2 - \sum x_i * y_i}$

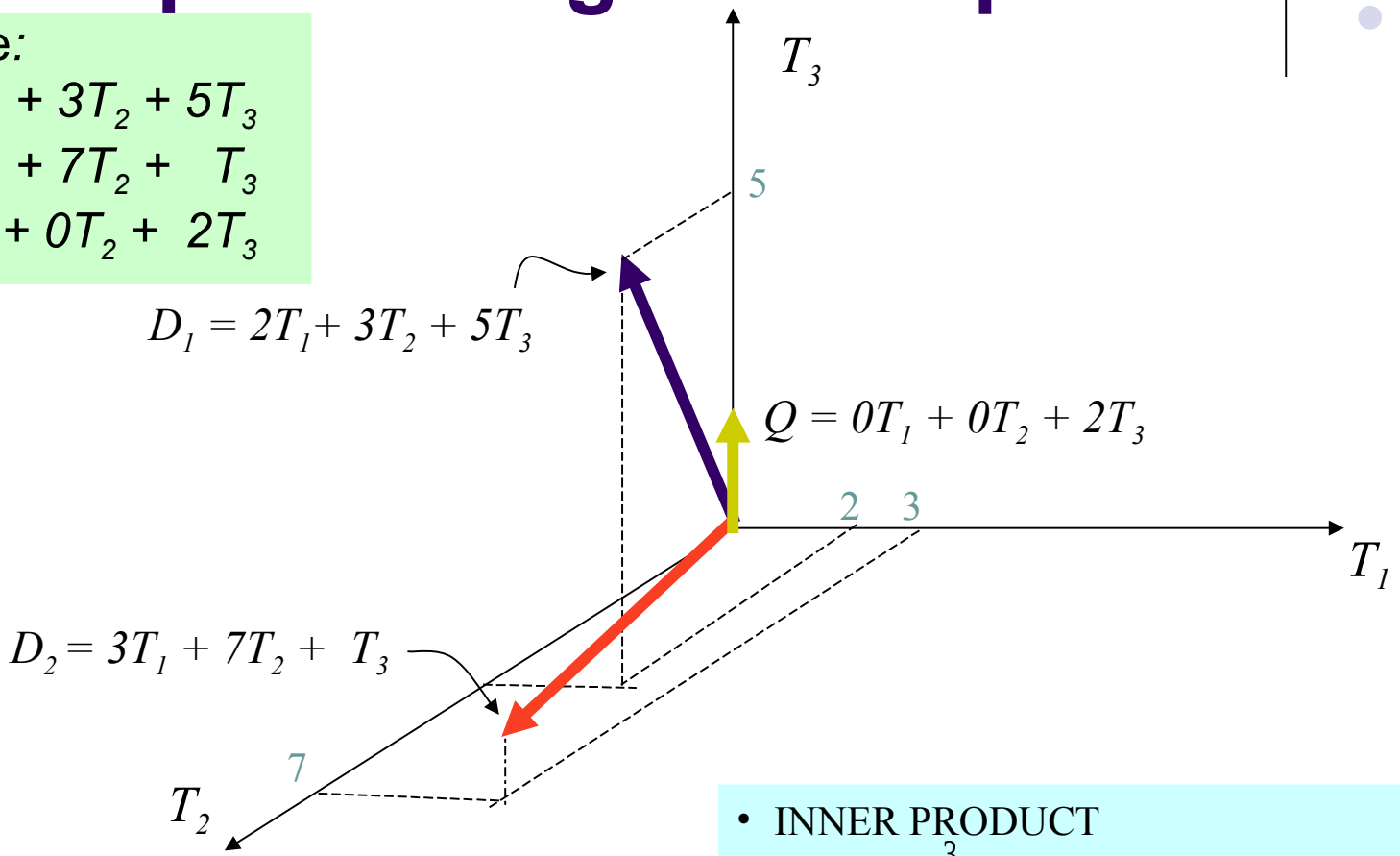
Le modèle vectoriel Interprétation géométrique

Exemple:

$D_1 = 2T_1 + 3T_2 + 5T_3$

$D_2 = 3T_1 + 7T_2 + T_3$

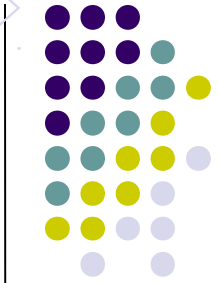
$Q = 0T_1 + 0T_2 + 2T_3$



- INNER PRODUCT

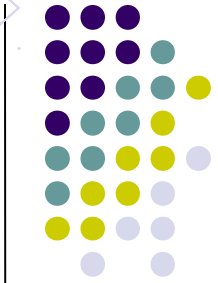
$$\text{sim}(q,d1) = \sum_{i=1}^{3} qi * d_1 i \quad = q3 * d_1 3 = 2*5 = 10$$

+> Dans l'implémentation, somme sur le poids des termes présents dans la requête



Le Modèle Vectoriel

- Avantages:
  - La pondération améliore les résultats de recherche
  - La mesure de similarité permet d'ordonner les documents selon leur pertinence vis à vis de la requête
- Inconvénients:
  - La représentation vectorielle suppose l'indépendance entre termes



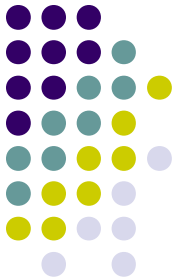
Modèles, mais encore…

- LSI (Latent Semantic Indexing)
- Modèle probabiliste
- Modèles de langages
- …

Qu'est ce qui marche ?

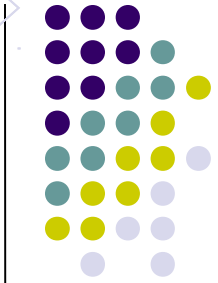
Logique floue
BM25
Statistique vs. linguistique
Indexer les passages
Probabiliste
Connexionnisme
Mots vides
thésaurus
troncature
IDF
tf
phrases
Groupes de mots
liens web
Vectoriel

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



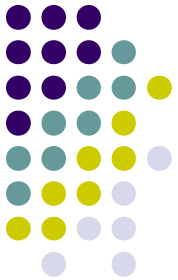
Critères d'évaluation

- Plusieurs critères
  - Facilité d'utilisation du système
  - Coût accès/stockage
  - Présentation des résultats

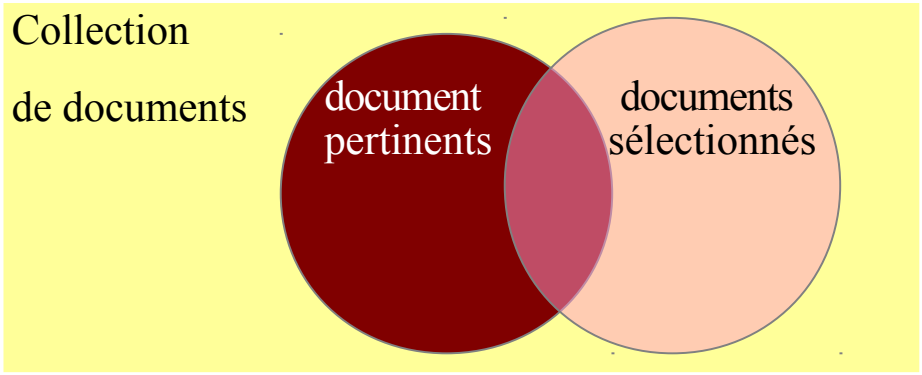
- Capacité d'un système à sélectionner des documents pertinents.

Deux facteurs

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



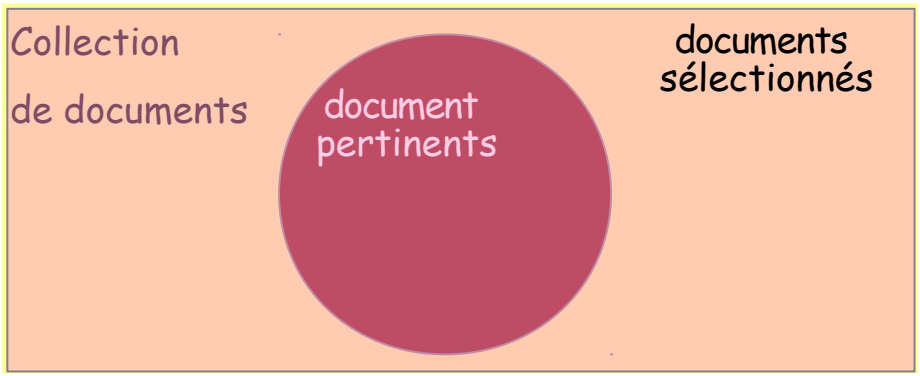
Précision et Rappel



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

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

Pourquoi deux facteurs ?

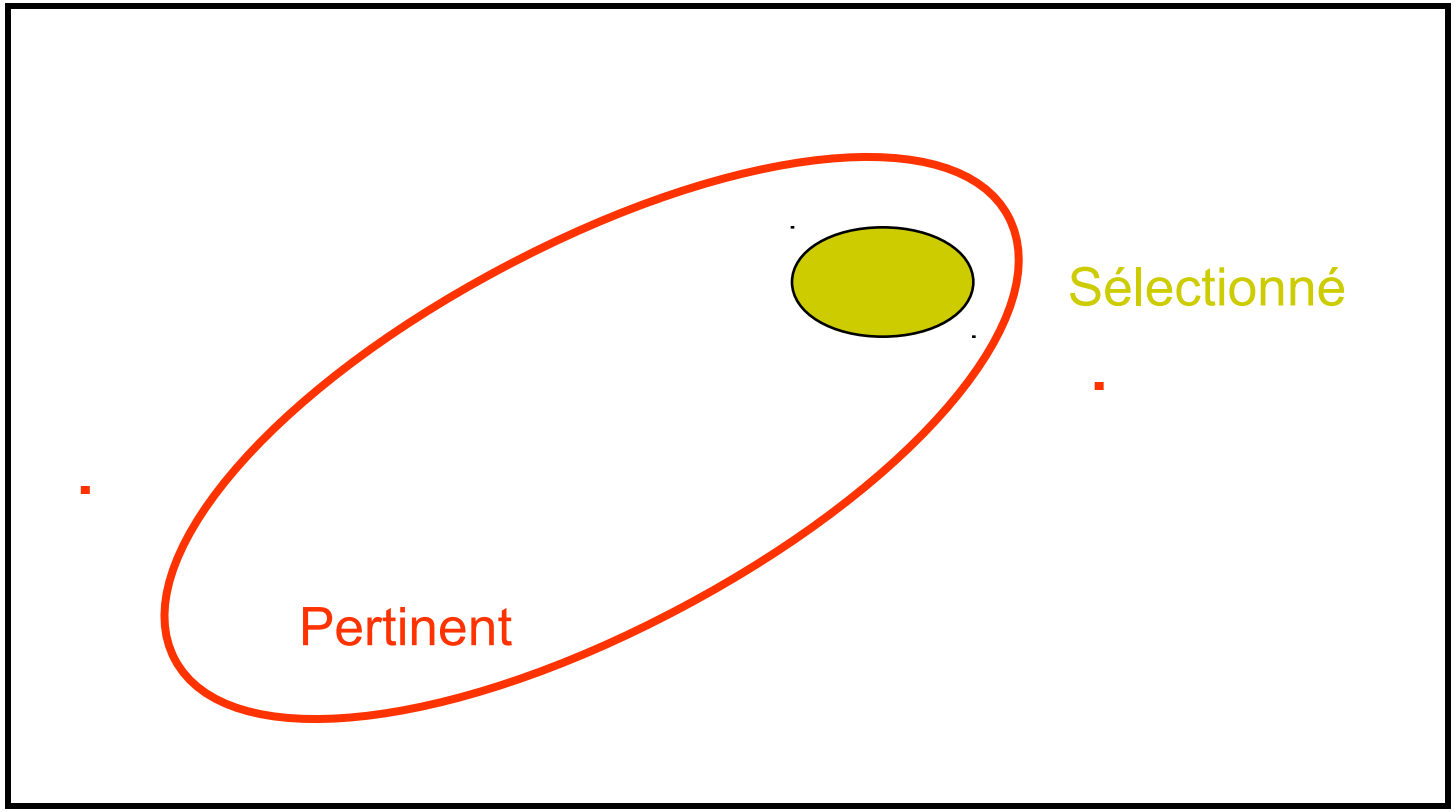


- FACILE de faire du rappel il suffit de sélectionner toute la collection
- MAIS, la précision sera très faible

Pertinent vs. Sélectionné

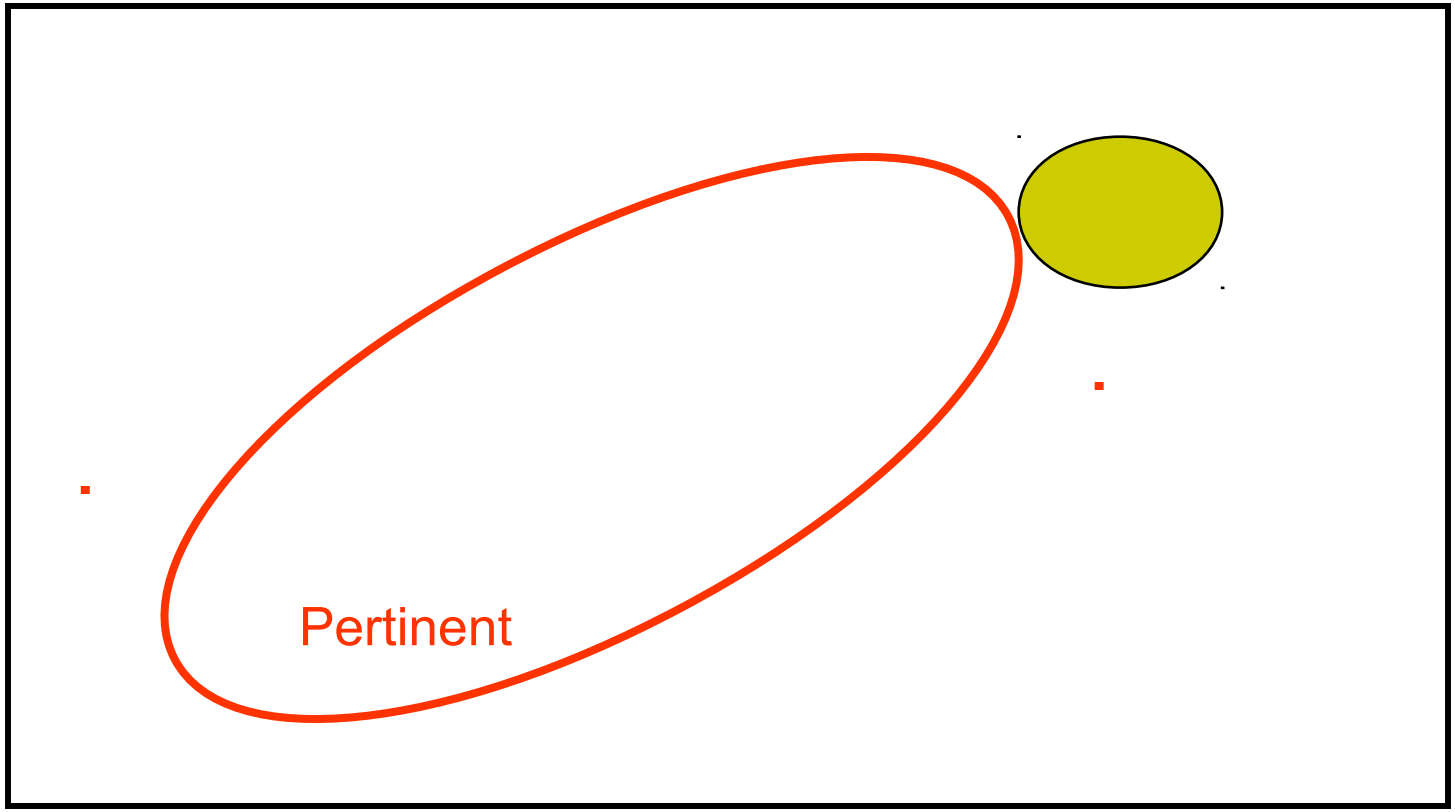
Sélectionné vs. Pertinent

Précision très élevée, rappel très faible



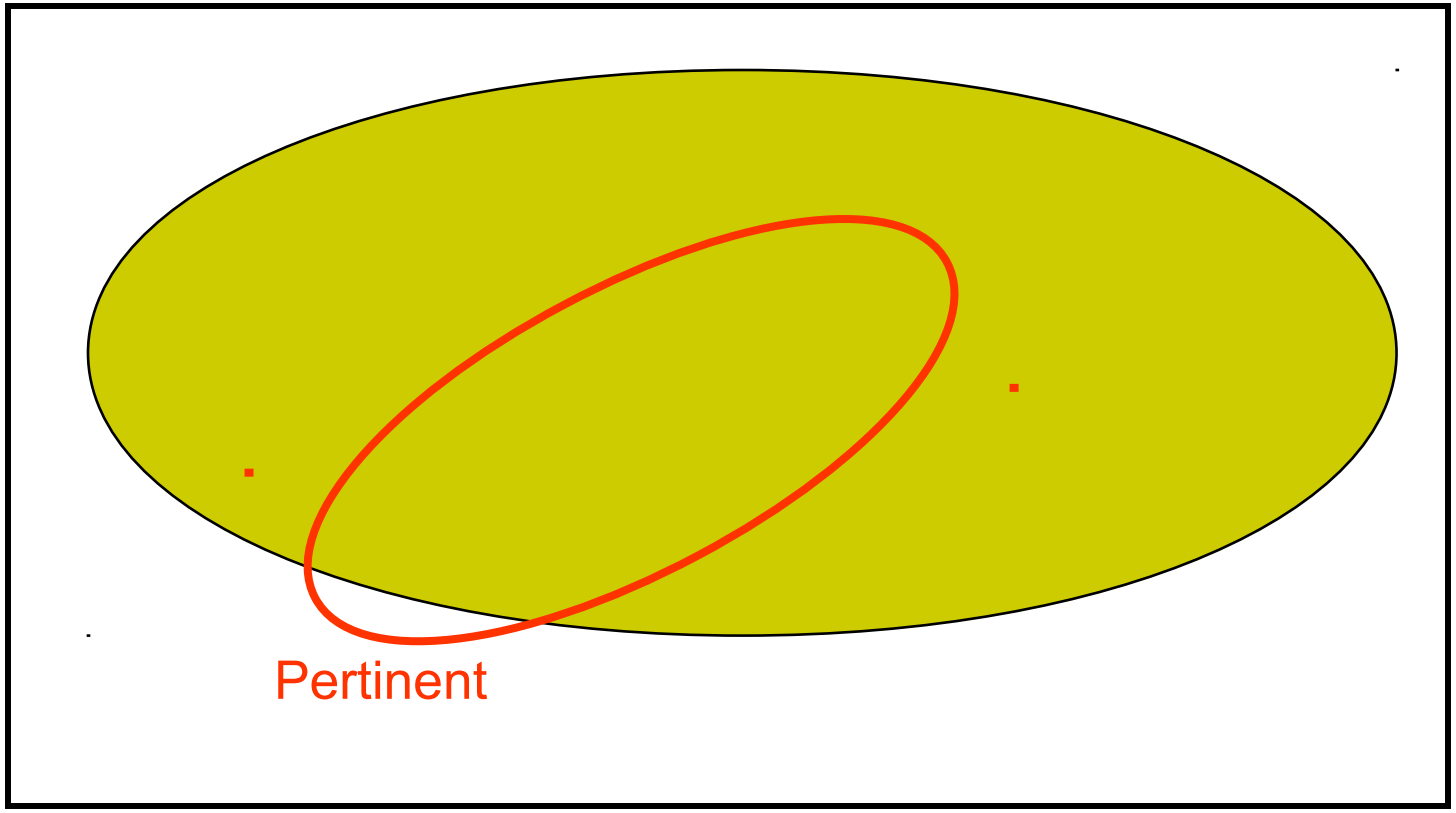
Sélectionné vs. Pertinent

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



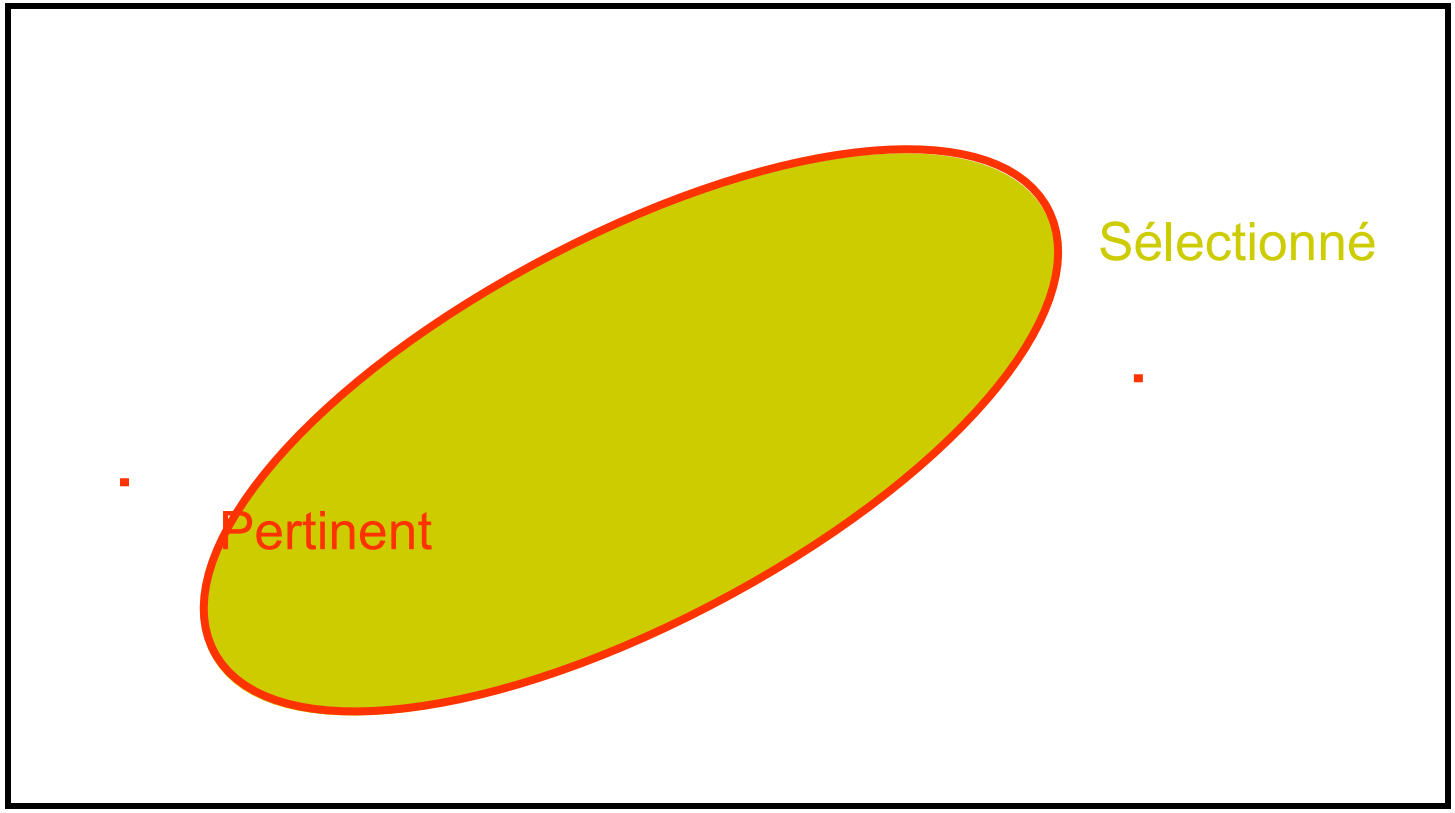
Sélectionné vs. Pertinent

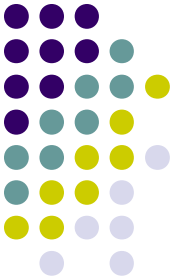
Rappel élevé, mais précision faible



Sélectionné vs. Pertinent

Précision élevée, rappel élevé (idéal, mais difficile)



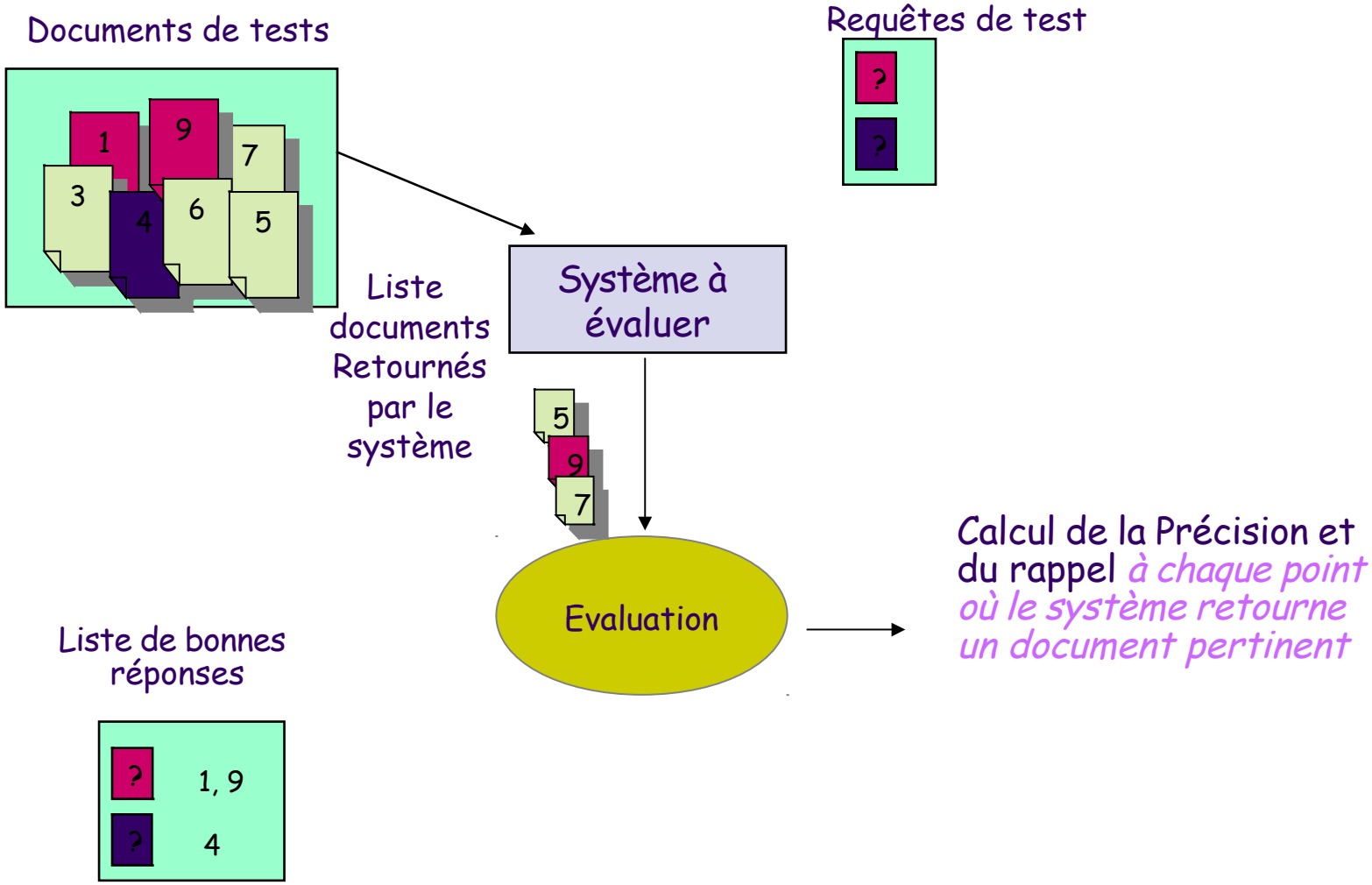


Démarche d'évaluation

- **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 requête)

Démarche expérimentale: Evaluation à la Cranfield



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

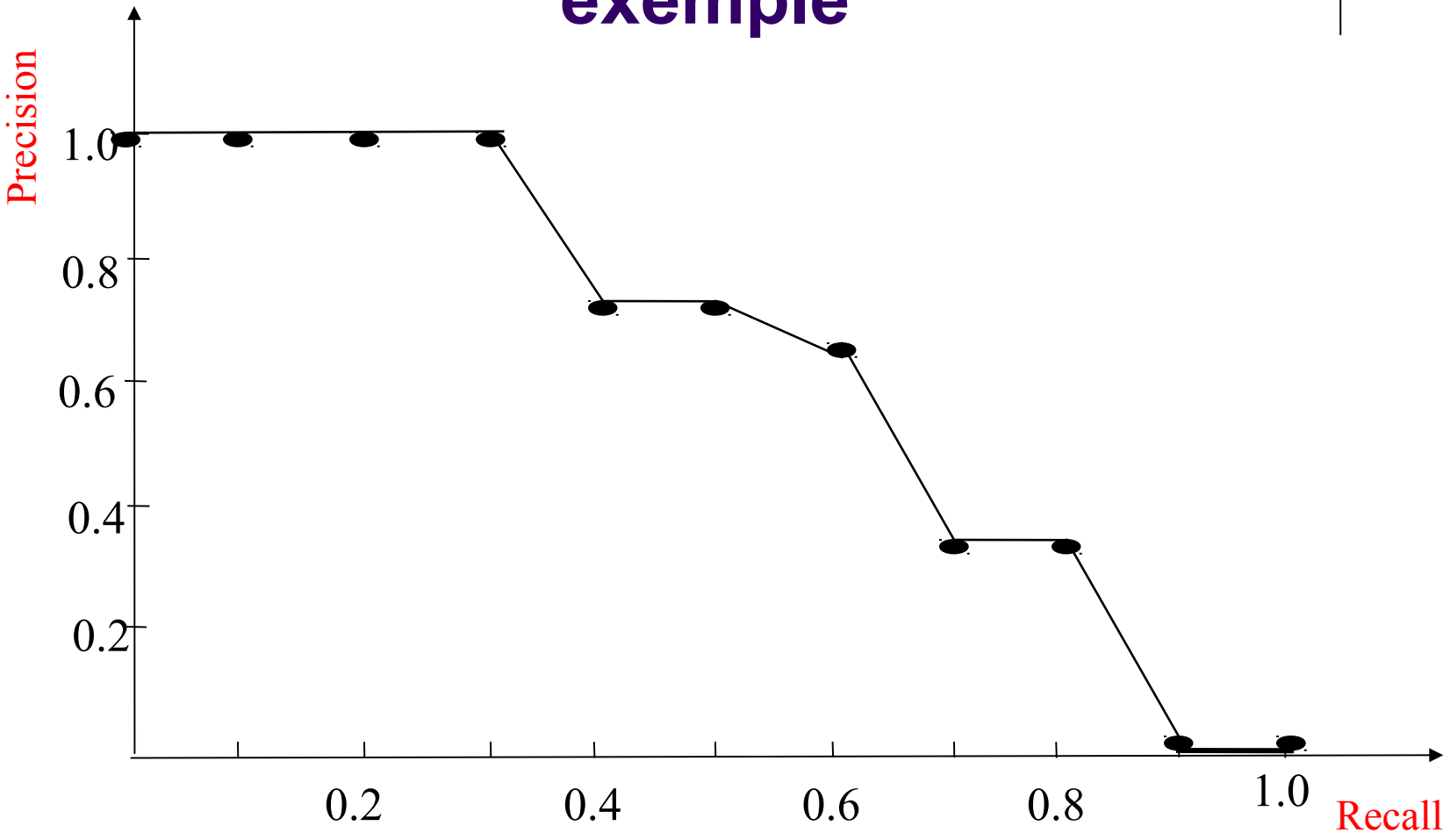
R=3/6=0.5; P=3/4=0.75

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

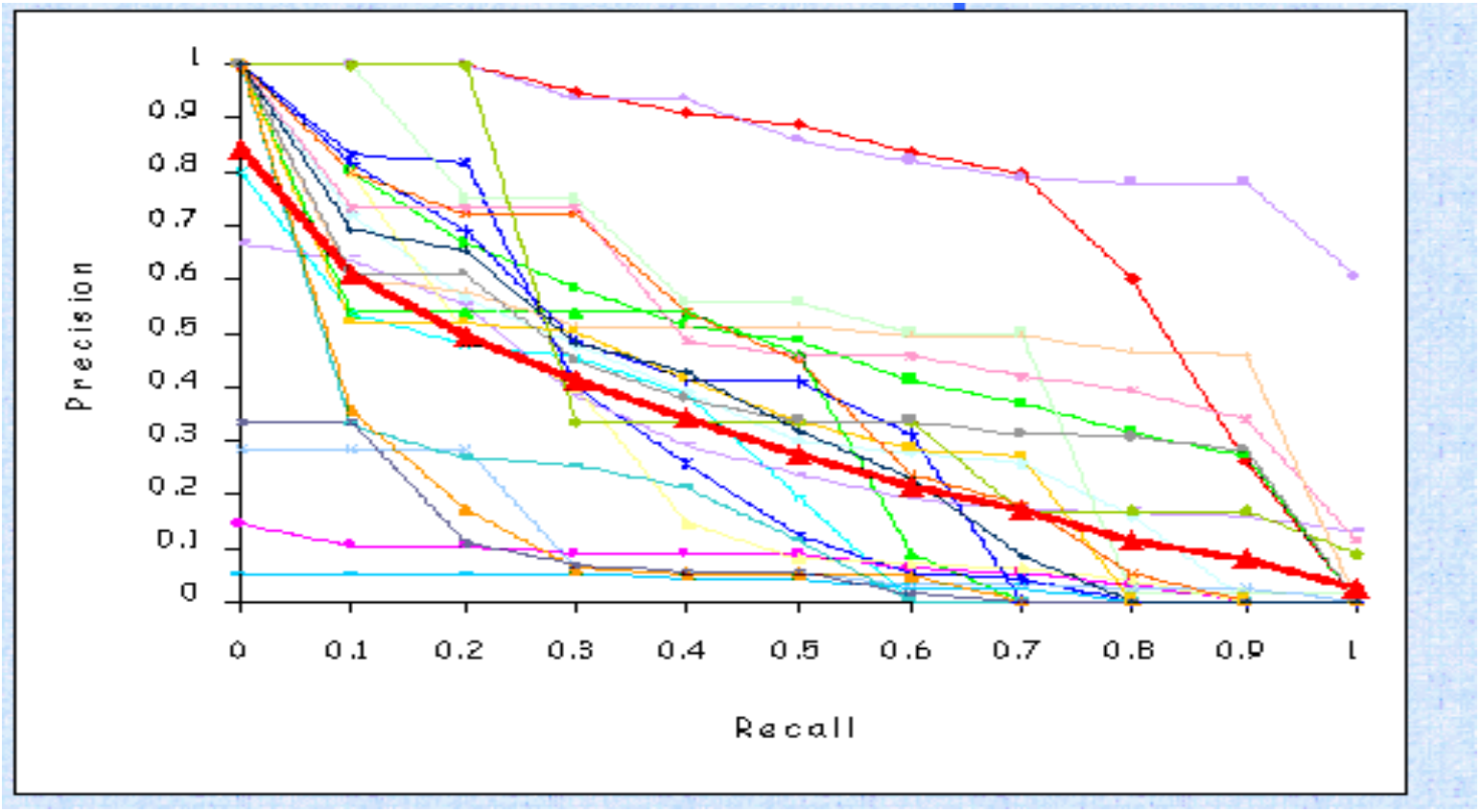
R=5/6=0.833; P=5/13=0.38

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

Interpolation de la courbe Rappel/Précision : exemple

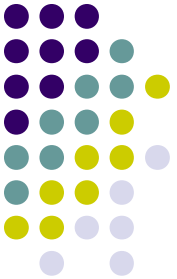


R-P courbes sur l'ensemble des requêtes



Illisible, difficile de comparer deux approches/systèmes requête par requête

On a besoin d'une moyenne entre les requêtes



Précision moyenne

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

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

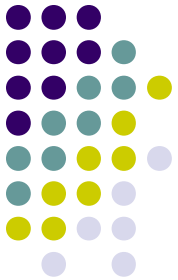
R=2/6=0.333; P=2/2=1

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

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

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

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

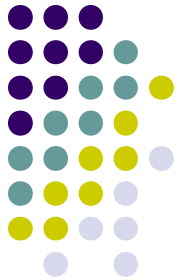


Précision à X documents

- Précision à différents niveaux de documents
  - Précision calculée à 5 docs, 10 docs, 15docs, …

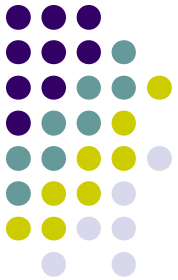
n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Prec. à 5 docs = 3/5
Prec. à 10 docs = 4/10



Pertinence

- Quelques suppositions « fausses »
  - Pertinence binaire (oui/non)
    - Les utilisateurs ne jugent pas souvent les documents par pertinent ou non pertinent
  - Pertinence d'un seul document peut être jugée indépendamment du contexte
    - Les utilisateurs peuvent juger différemment un document selon ce qu'ils ont vu au préalable.



En conclusion : problématique de la RI

- La RI est un domaine en pleine expansion
- De plus en plus important car
  - les masses d'information n'arrêtent pas d'augmenter
    - Web : plus de 170 Téraoctets (croissance exponentielle) (Web surfacique, sans compter les pages dynamiques)
    - Journaux : 25 Téraoctets (annuellement),... Documents (bureau) : 195 Téraoctets …"
    - on estime : 610 milliards emails sont envoyés chaque année soit 11 téraoctets"
  - … les demandes d'information (utilisateurs) n'arrêtent pas d'augmenter

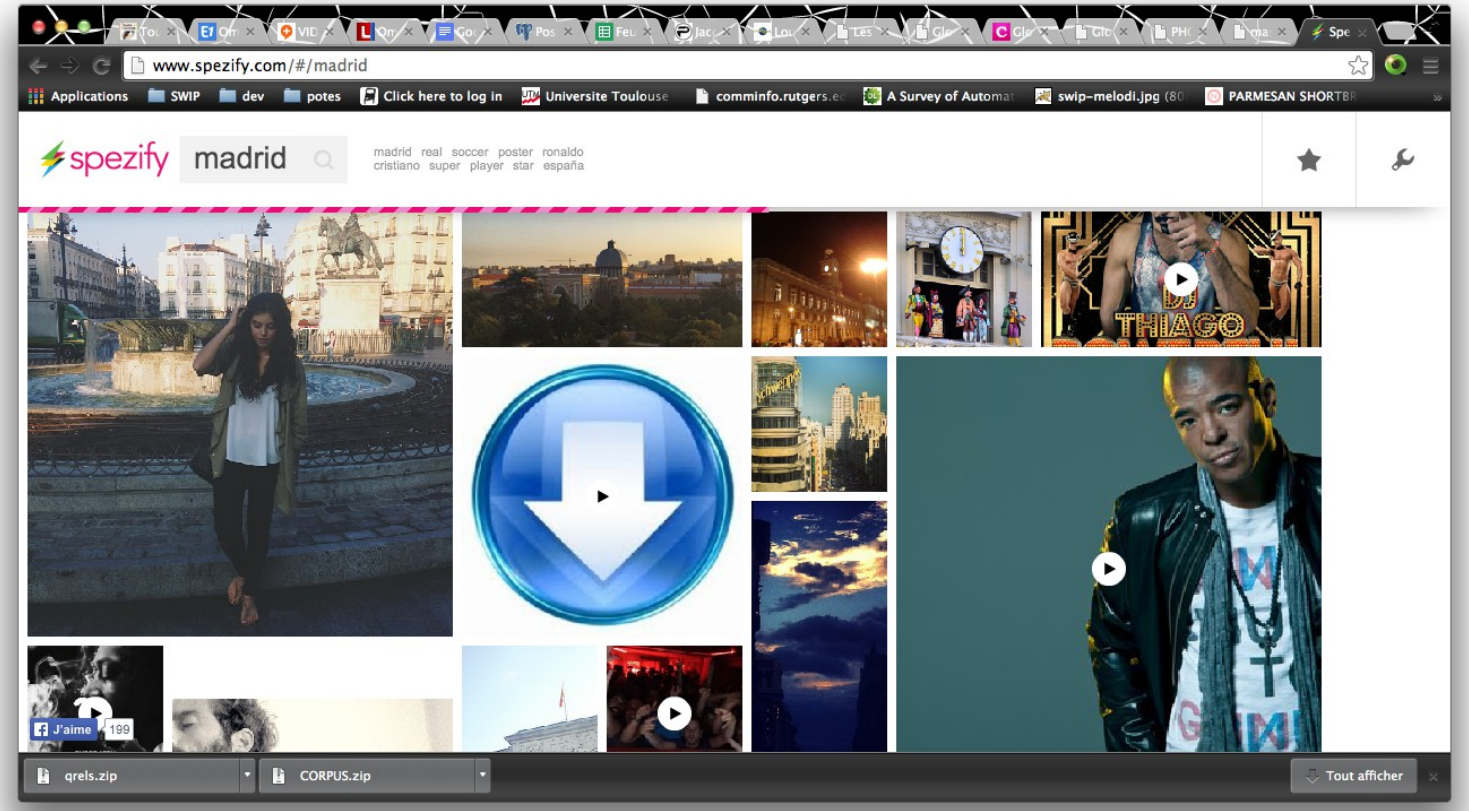
RI = plusieurs tâches et plusieurs problématiques

- **RI sur le Web**
  - Utilisation du contenu et des liens
    - PageRank
- **Accès personnalisé à l'information**
  - Prise en compte de l'utilisateur dans le processus de RI
  - Adaptabilité, flexibilité du processus
- **Recherche d'information multilingue**
  - Passer les barrières de la langue
  - Ex: Requête en français sur des documents en chinois
- **RI sur des documents structurés (XML)**
  - Combinaison de la structure et du contenu pour identifier les unités pertinentes
  - Quelles unités (éléments) à indexer ? quelles unités sélectionner ?

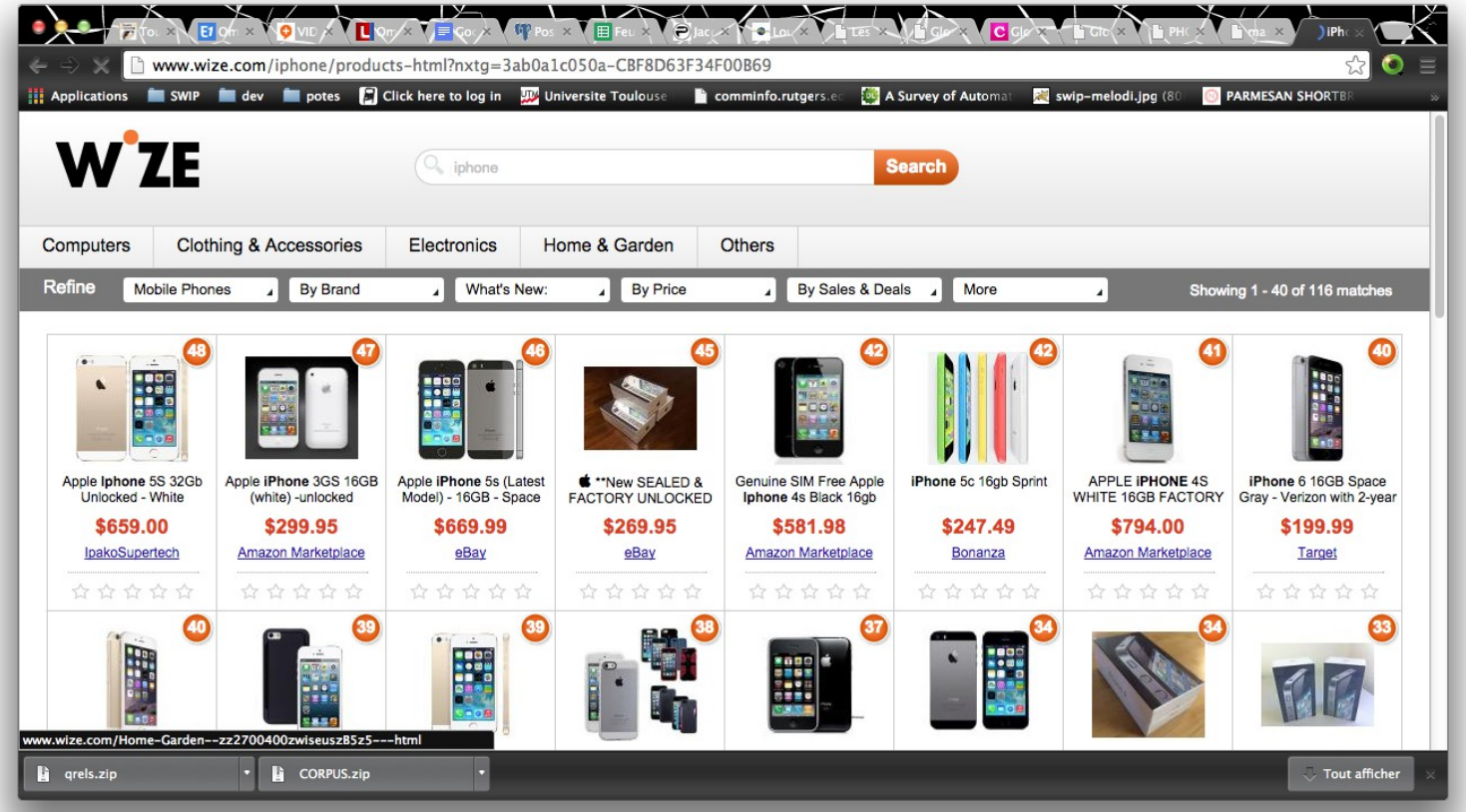
RI = plusieurs tâches et plusieurs problématiques

- **Filtrage / Recommandation d'information/**
  - Sélection de documents pertinents dans un flot de documents
- **Classification et catégorisation**
  - Regrouper les documents dans des classes
- **Présentation et Visualisation des résultats**
- **Questions / réponses**
  - Trouver des réponses à des questions
- **Passage à l'échelle**
  - Accès à plusieurs milliards de documents
- **Recherche agrégée**
- **Évaluation des performances**
  - Métriques, Benchmark, etc. pour l'évaluation de l'efficacité
  - Plusieurs Campagnes d'évaluation ….

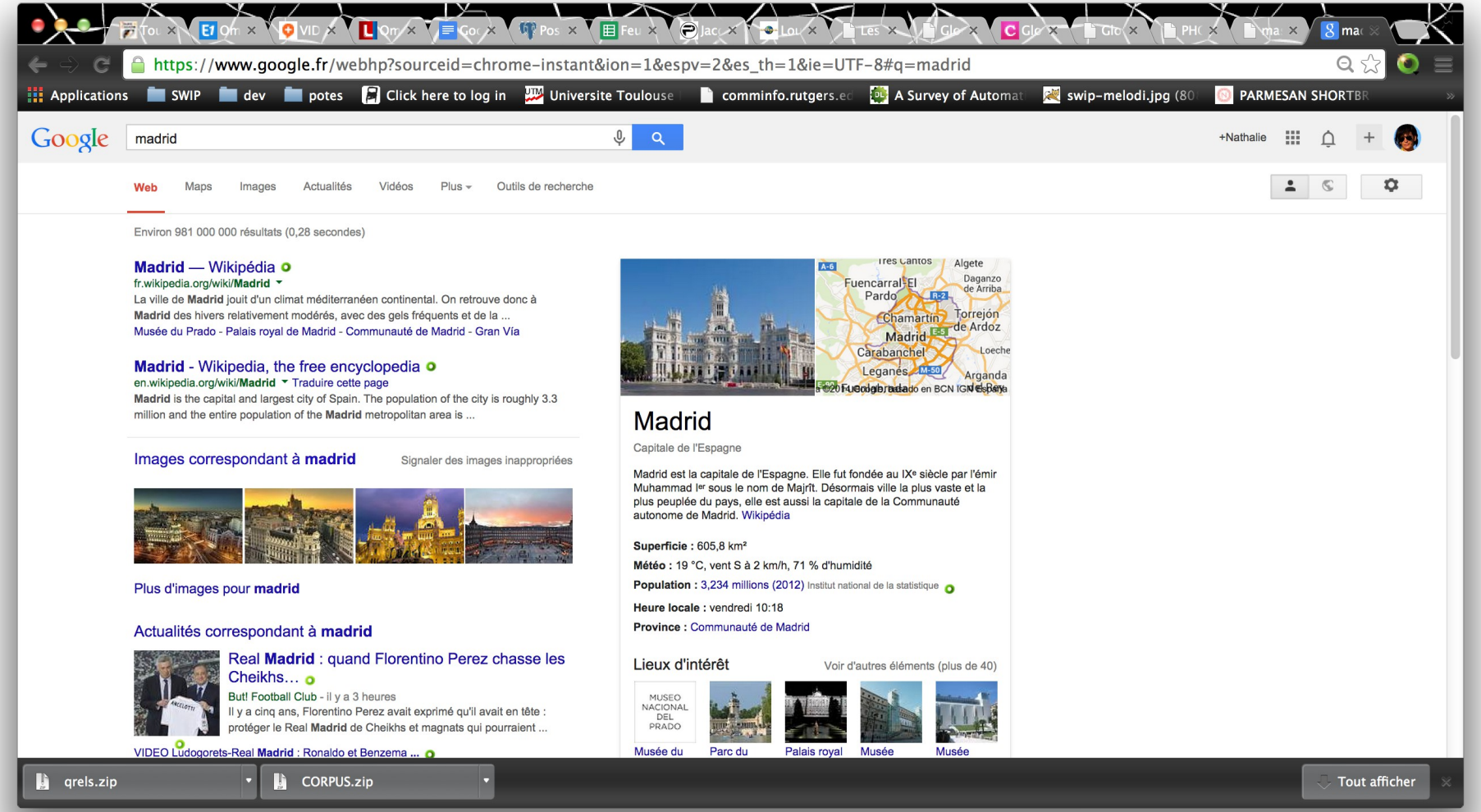
Zoom sur la visualisation

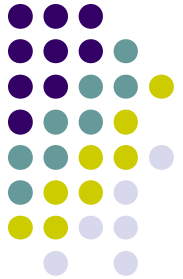


Zoom sur la visualisation / aggrégation



Vers le web sémantique !





Ce support a été réalisé par Karen Pinel

Un grand merci pour m’avoir permis de l’utiliser