UNIVERSITÉ DES SCIENCES ET TECHNOLOGIE HOUARI BOUMEDIENE

FACULTÉ D'INFORMATIQUE - MASTER 2 SII

RECHERCHE D'INFORMATION
INFORMATION RETRIEVAL

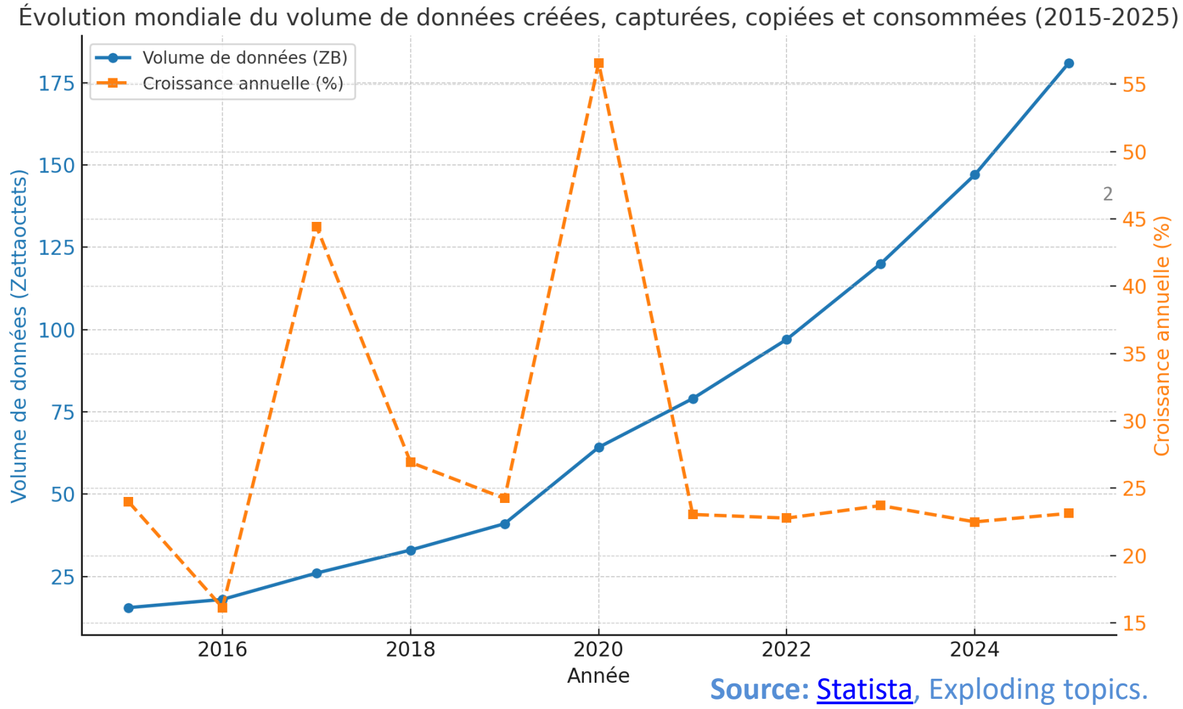
CHAPITRE 1: INTRODUCTION A LA RI

22 Septembre 2025

I. CONTEXTE ET MOTIVATION

EXPLOSION DES DONNÉES NUMÉRIQUES & BIG DATA

La quantité de données numériques produites dans le monde provenant d'une grande diversité de format: documents textuels, images, vidéos, données issues de capteurs, etc., connaît une **croissance exponentielle**.



Source: Statista, Exploding topics.

Nous sommes passé d'environ **15 zettaoctets en 2015** à une estimation de plus de **180 zettaoctets en 2025**.

La **Recherche d'Information (RI)** est devenue indispensable. Plus les données augmentent, plus il est difficile de trouver l'information pertinente.

II. DÉFINITIONS ET TERMINOLOGIE

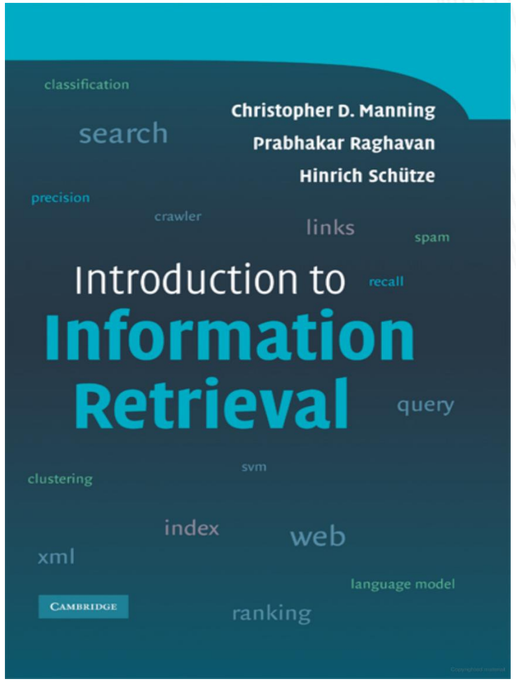
La Recherche d'information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information **pertinente** pour un utilisateur pour répondre à un besoin d'information précis.

Cela implique la formulation d'une requête, l'utilisation d'outils de recherche, la sélection et l'évaluation des documents trouvés.

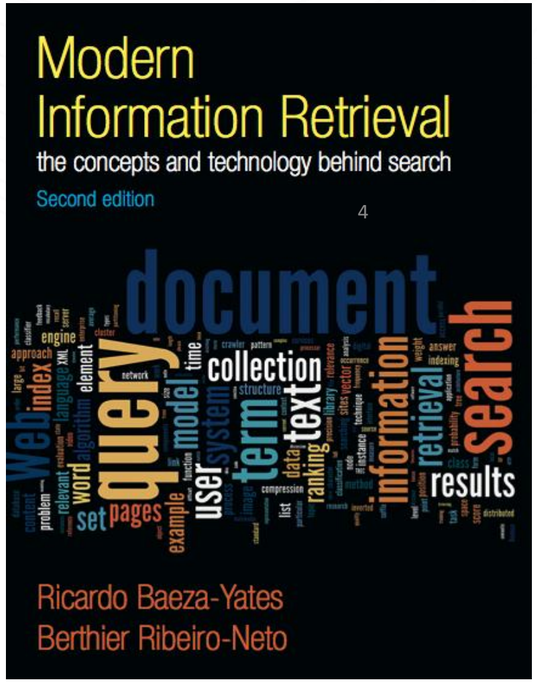
Terminologie

- Recherche d'information
- Informatique documentaire
- Information retrieval
- Textual information retrieval
- Document retrieval

II. DÉFINITIONS ET TERMINOLOGIE



IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).



Information retrieval deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organization of the information items should be such as to provide the users with easy access to information of their interest.

III. DOMAINE D'APPLICATION

Web

- Moteurs de recherche (Google, Bing, etc.)
- Recherche d'images, de vidéos et de contenus multimédias (Google Images, YouTube, etc.)

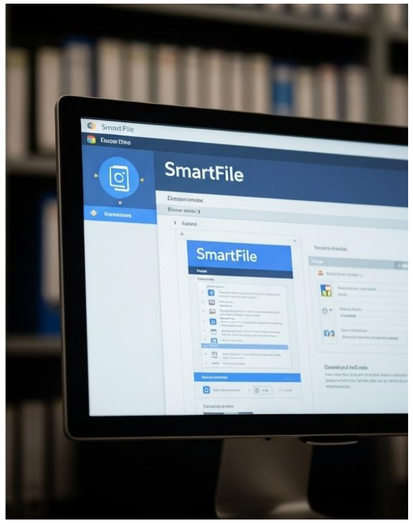
Bibliothèques numériques (Digital Libraries)

- Accès structuré et intelligent aux archives scientifiques (arXiv, Scopus, DBLP)
- Recherche de documents multilingues et interconnectés (Google Scholar)

III. DOMAINE D'APPLICATION

Entreprises

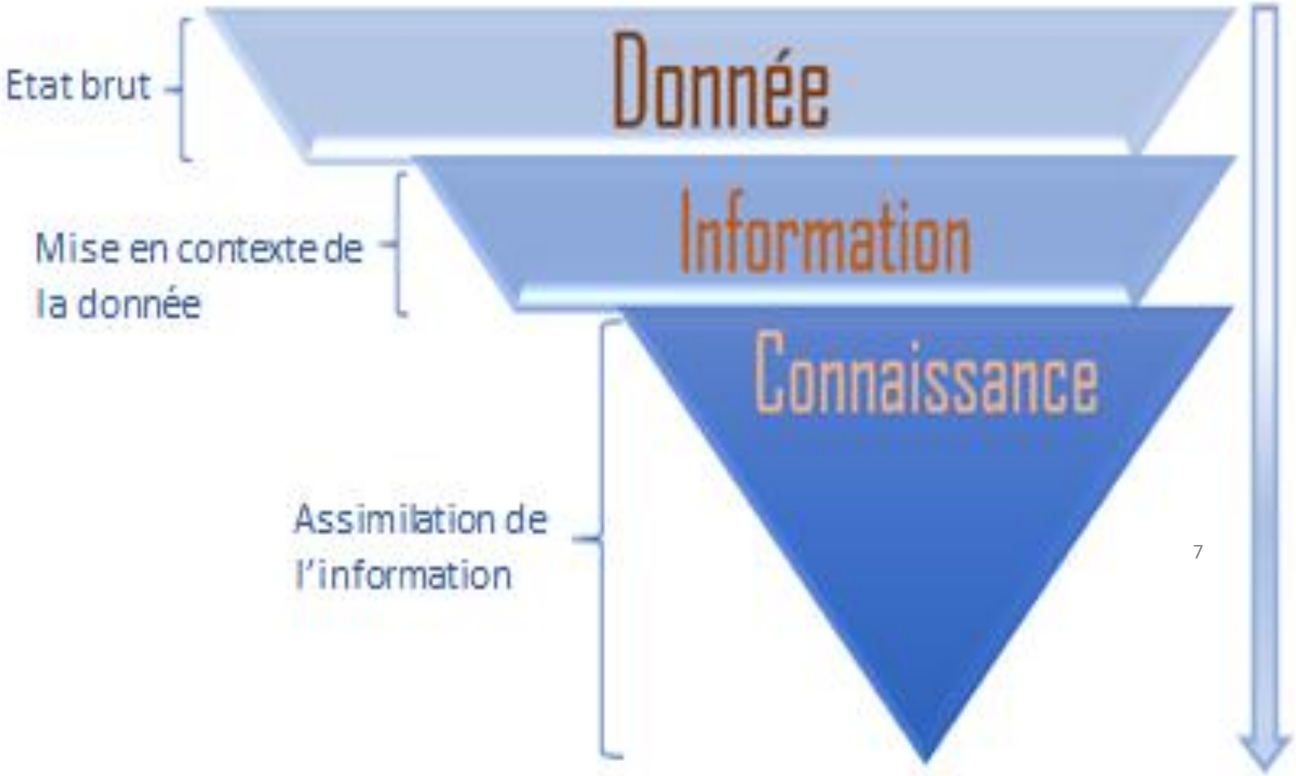
- Gestion documentaire et archivage (SharePoint -Microsoft)
- Systèmes de gestion électronique de documents (GED)
- Veille stratégique et extraction de connaissances à partir de données massives



Ordinateurs personnels (PC / Smartphones)

- Recherche de fichiers, emails et documents stockés localement
- Organisation automatique des contenus multimédias
- Systèmes intelligents de recommandation (musique, films, apps, etc.)

IV. DONNEE-INFO-CONNAISSANCE



SYSTÈME DE GESTION DE BASE DE DONNÉES

Donnée stocké = 0,87

SYSTÈME DE RECHERCHE D'INFORMATION

0,87 est le taux de reconnaissance donné par un modèle de classification qu'une image représente un chat

DATA MINING

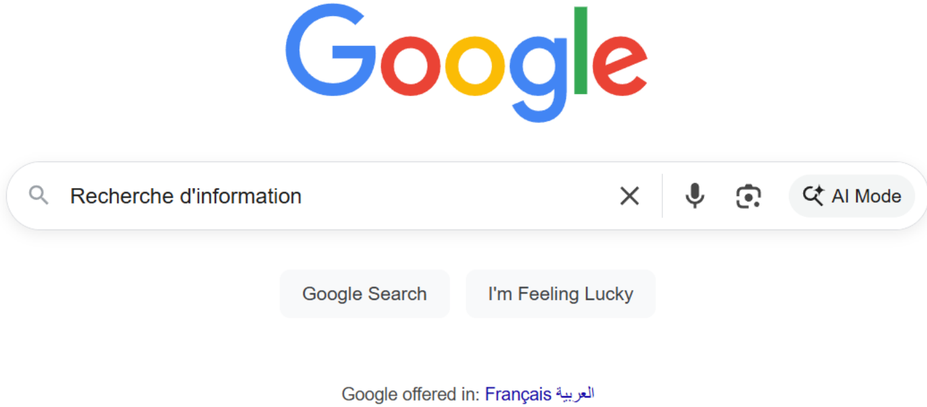
Avec 0.87, l'image est très probablement un chat, cela indique qu'il convient de la classer dans la catégorie ''chat''

LE SGBD GÈRE LA DONNÉE BRUTE, LE SRI LUI DONNE DU SENS ET LE DATA MINING EN DÉDUIT UNE CONNAISSANCE.

V. TACHE DE LA RI

- 🔎 **RECHERCHE ADHOC**

- Recherche d'informations sur un sujet donné à partir d'une requête.
- La requête est soumise au Système de Recherche d'Information (SRI) → retourne une liste de documents pertinents.



- 📌 **PLUSIEURS TYPES DE RI ADHOC**

- **Domaines spécialisés** : médical, légal, chimie, etc.
- **Recherche d'opinions** (*opinion retrieval*, analyse de sentiments)
- **Recherche d'événements**
- **Recherche de personnes** (experts)

V. TACHE DE LA RI

- 🗂️ **CLASSIFICATION / CATÉGORISATION**
  – Regrouper les informations (documents) selon un ou plusieurs critères
- ❓ **QUESTION-RÉPONSES (QUERY ANSWERING)**
  - Chercher des réponses à des questions
  – par exemple
    « Quelle est la version actuelle du Python ? »
    « Comment réinitialiser le Mot de Passe sur Win 11? »
- 🎯 **FILTRAGE D'INFORMATION / RECOMMANDATION (FILTERING/RECOMMENDATION)**
  - **Recommandation** : proposer des contenus adaptés (**Netflix / YouTube**).
  - **Dissémination sélective d'information** : diffusion ciblée selon critères.
  - **Systèmes d'alerte** : notifier en temps réel (ex. alerte météo, sécurité).
  - **Push** : envoi automatique d'informations à l'utilisateur.
  - **Profilage (profiling)** : personnalisation en fonction des préférences et de l'historique.

V. TACHE DE LA RI

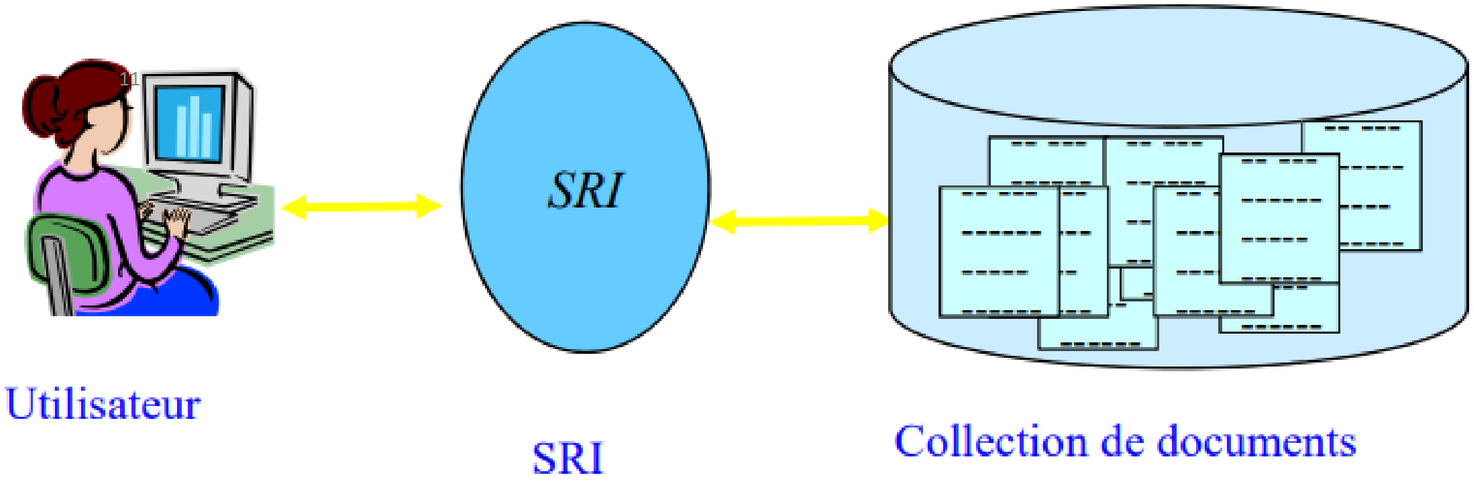
- 📑 **Résumé automatique (document summarization)**
  - Produire automatiquement un résumé concis à partir d'un ou plusieurs documents.
- 🌐 **Recherche agrégée (Aggregated search)**
  - Agréger des moteurs : interroger les résultats de plusieurs moteurs (ex. Google + Bing).
  - Agréger des résultats : interroger plusieurs sources (vertical search)
  - Agréger des contenus : fusionner plusieurs documents en un seul résultat pertinent.

VI. SYSTEME DE RECHERCHE D'INFORMATION –SRI

Un Système de Recherche d'Information (SRI) est un programme(ensemble de programmes) informatique qui a pour but de sélectionner des **informations pertinentes** répondant à des **besoins des utilisateurs.**

Dans cette définition il y a 3 notions clés à savoir :

- **L'information**
- **Le besoin de l'utilisateur**
- **la pertinence**



VI. SYSTEME DE RECHERCHE D'INFORMATION –SRI

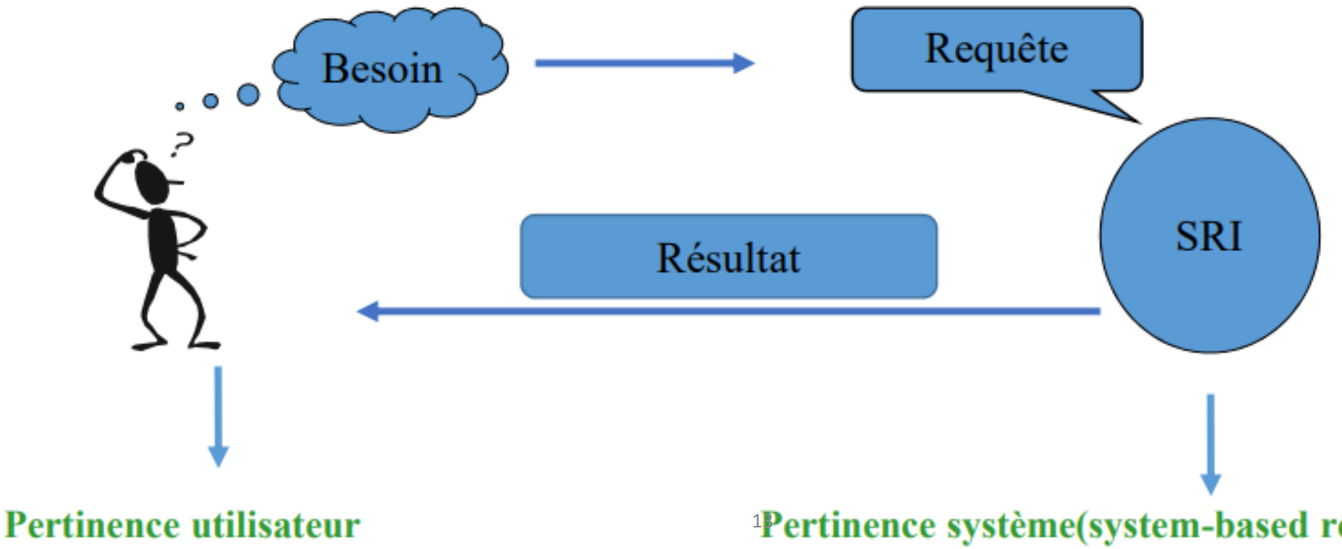
L'information -Information : Peut-être un texte libre, texte structuré, document, une page web, une image, une vidéo …etc. Dans ce cours nous traitons seulement les documents textuels.

Le besoin de l'utilisateur : Connu généralement par le mot « requête- query », qui exprime le besoin d'information d'un utilisateur. Une requête peut avoir différentes formes selon le modèle utilisé. Souvent elle est exprimée par une liste de mots-clés

La pertinence - Relevance: C'est une relation de correspondance entre un document et une requête (besoin en information), selon l'utilisateur ou le système. On distingue deux types de pertinences.

- Pertinence système (similarité calculée par le système entre une document et une requête)
- Pertinence utilisateur (satisfaction de l'utilisateur par le document)

VI.1. PERTINENCE UTILISATEUR VS. PERTINENCE SYSTÈME



✓ **PERTINENCE SUBJECTIVE : dépend de la perception de l'utilisateur.**

✓ **Ex. un article peut être pertinent pour un étudiant en droit, mais pas pour un étudiant en informatique.**

✓ **PERTINENCE OBJECTIVE ET ALGORITHMIQUE :**

- **Calcul automatique du SRI.**
- **Le système compare la requête avec les documents → attribue un score de pertinence.**
- **Plus le score est élevé, plus le document est jugé pertinent par le système.**

VI.2. PROBLÉMATIQUE DE LA PERTINENCE

🌐 **La pertinence est multidimensionnelle**

- dépend de plusieurs paramètres : profil de l'utilisateur, besoin en information, situations des utilisateurs, …

📈 **La pertinence est graduelle**

- un document A peut être plus pertinent que B

⏳ **La pertinence est dynamique**

- peut changer dans le temps, selon l'état de connaissance de l'utilisateur au moment de la recherche

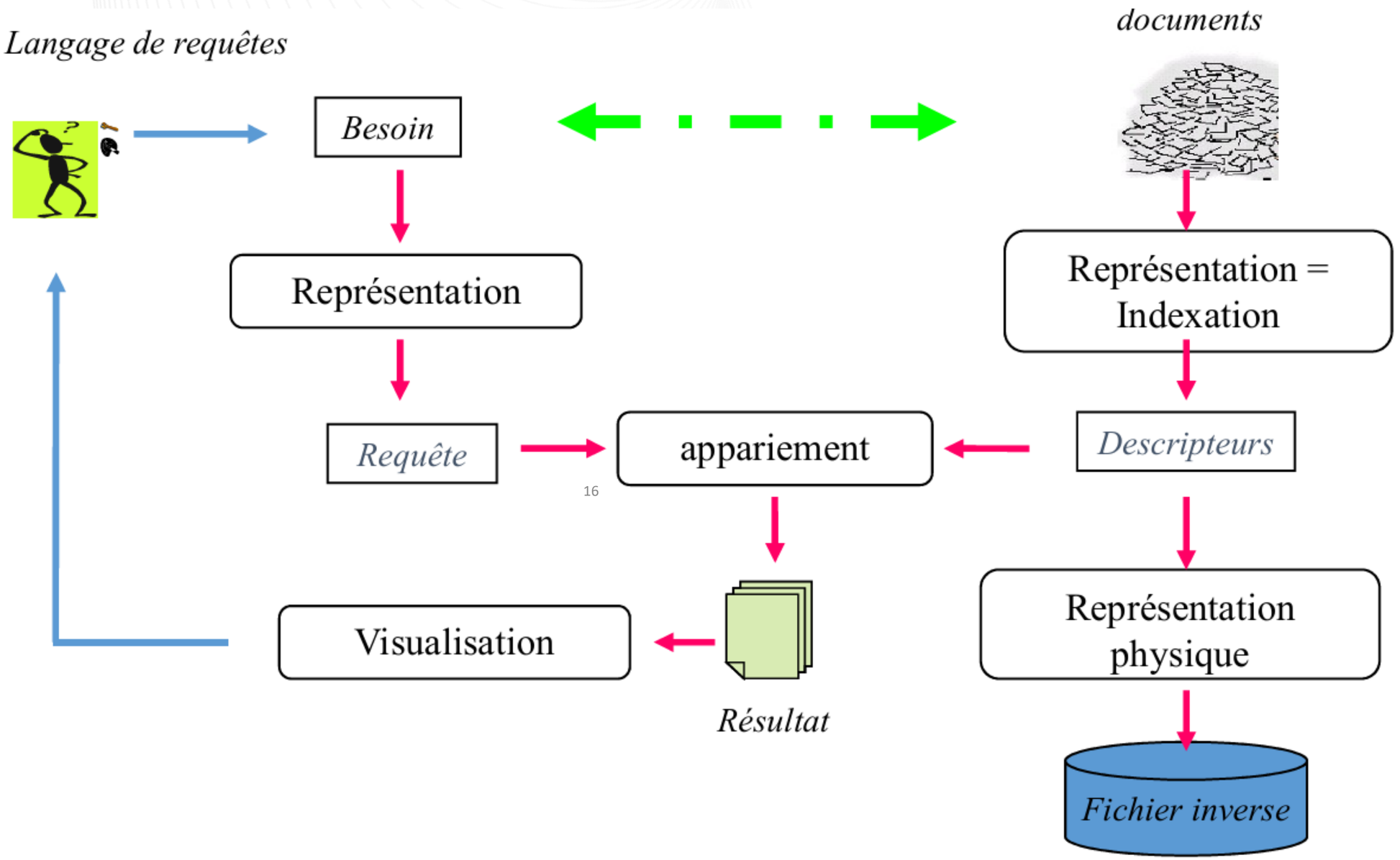
💻 **La pertinence est difficile à automatiser, un système calcule des scores, mais il ne peut pas totalement remplacer le jugement humain. C'est un des grands défis de la recherche d'information.**

VI.3. APPROCHE GÉNÉRALE D'UN SRI

La vision simple de l'approche de la RI textuelle est de trouver les documents ayant les mêmes mots que la requête :

- La requête est une liste de mots clés
- Le document est une liste de mots clés
- Comparer les mots de chaque document à ceux de la requête
- Sélectionner les documents qui contiennent les mots de la requête.

VII. PROCESSUS DE RI



VII.1. DEFIS FONDAMENTAUX EN RI

📝 **Comment représenter (indexer) un document ?**

🔎 **Comment représenter (indexer) une requête ?**

⚖️ **Comment mesurer la pertinence (similarité ou appariement) entre un document et une requête ?**

VII.2. TRAVAUX DE RECHERCHE EN RI

- ✓ **Proposer des solutions** : nouveaux modèles, techniques, approches, outils pour répondre à ces problèmes pour améliorer la pertinence des systèmes
- ✓ **Quels supports théoriques ?** Souvent basés sur des théories mathématiques : probabilités, statistiques, ensembles, algèbre, logique floue, analyse de données, …
- ✓ **Quel processus pour la validation ?** Pour évaluer une approche ou un système de RI, il faut y avoir des données d'évaluation (environnement de tests, datatsets, benchmarks).

CONCLUSION

La RI est un domaine en pleine expansion de plus en plus important car :

- Les masses d'information n'arrêtent pas d'augmenter.
- Les demandes d'information (utilisateurs) n'arrêtent pas d'augmenter