Systèmes de RI et Big Data

Pr. Bouzid

Génie informatique – 5ème année

Année universitaire 2024 - 2025

Objectifs du module

- **Comprendre le fonctionnement des SRI (système de recherche d'information):**
  - Comprendre la problématique de la RI et découvrir les techniques de bases pour créer un système de recherche d'information pratique et efficace (représentation des documents, modèles de recherche, …)
  - Comprendre le fonctionnement d'un moteur de recherche (documentaires, web,…), et évaluer ses résultats
- **Découvrir le domaine du Big Data**
  - Se familiariser avec les notions autour du Big Data (le quoi et le pourquoi du Big Data, ses domaines d'application, comment sont gérés les données qualifiées de Big Data,…)
  - Découvrir les technologies et plateformes utilisées pour le Big Data (technique du Map Reduce, plateformes hadoop, spark…)

Mode d'évaluation

- Partie Cours:
  - Contrôle sous forme de QCM sur le cours (SRI et Big Data)
  - Échéance: Début novembre (semaine des contrôles)

- Partie Projet:
  - Projet en groupes
  - Échéance: mi-décembre
  - Livrables: Application + Rapport + Démo

- Note du module:
  - 50% note contrôle + 50% note projet

Plan

- Introduction générale
- Systèmes de recherche d'informations
- Les fondements du Big Data

Introduction Générale

Information VS Donnée

- Une donnée est l'enregistrement d'une observation, d'un objet, d'un fait destiné à être interprété, traité par l'homme. La donnée est généralement objective.
  - *Exemples :*
    - *température = 35°*
    - *âge = 2 mois*

- Une information est le sens (*l'interprétation*) attaché à la donnée ou à un ensemble de données par association. L'information est généralement subjective, définie selon un contexte.
  - *Exemples:*
    - *[température=35°] : temps chaud*
    - *[âge=2 mois] : nourrisson*

Information

- Une information est donc subjective et peut demander un ensemble de données pour la rechercher
  - *Exemple* : seconde guerre mondiale
- Une information peut se trouver dans différents supports:
  - Fichiers (page web, fichiers xml, txt…), documents textuel (.doc, .docx, pdf,…), images, vidéo, etc.
- Quand on recherche une information dans un moteur de recherche pour obtenir le support qui contient l'information: On parle de la **Recherche d'Information (RI)**

  En anglais: Information Retrieval (IR)
- Un système (tel un moteur de recherche) qui permet de rechercher des informations est ce qu'on appelle: **un SRI (Système de Recherche d'Information)**

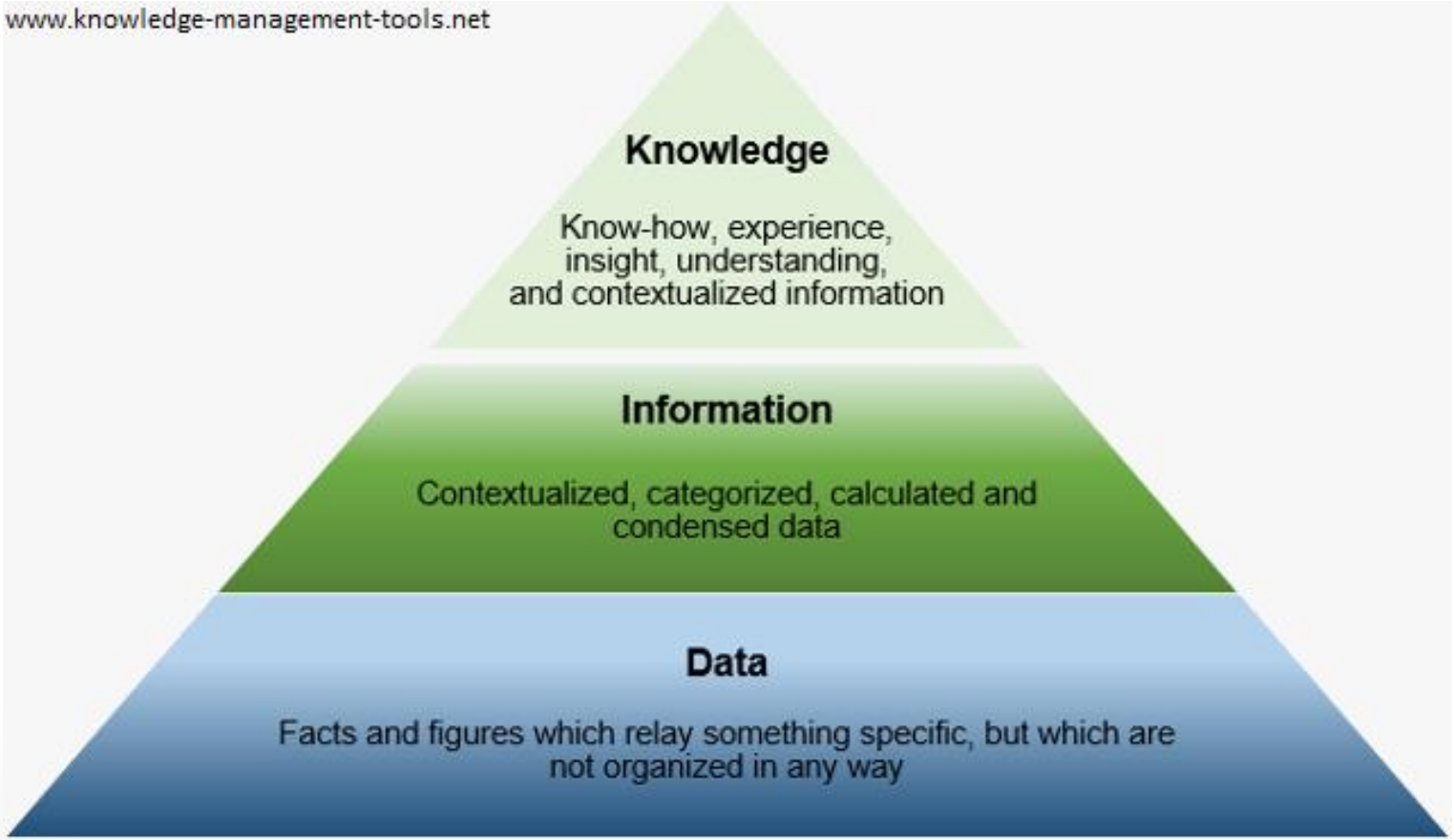
Information VS Connaissance

- Quand une information est nouvelle, créée par association d'informations de base, de règles logiques et de règles de décisions: on parle de **connaissance**
- Ces règles peuvent être issues d'expérience, d'expertise, de faits / réalités connu(e)s, etc.
  - *Exemple* :
    - *temps chaud et enfant nourrisson => risque de déshydratation*
- Il s'agit du domaine de la gestion des connaissances : **Knowledge Management (KM)**

Connaissance

- Dans le domaine du **KM** on trouve:
  - Knowledge bases : pour le stockage des connaissances acquises et/ou créées (BD, Datawarehouse,…)
  - Knowledge management systems : pour l'acquisition des connaissances (*à partir d'informations*), l'organisation de ces dernières, le partage et la recherche
  - Decision support systems (DSS): applications pour la prise de décisions dans un domaine précis (calculs statistiques, modèles d'analyse, génération de graphiques, reportings, …)
  - AI (Artificial Intelligence) : outils de raisonnement et inférence, machine learning et autres techniques pour la création de la connaissance

Data VS Information VS Connaissance



Données massives

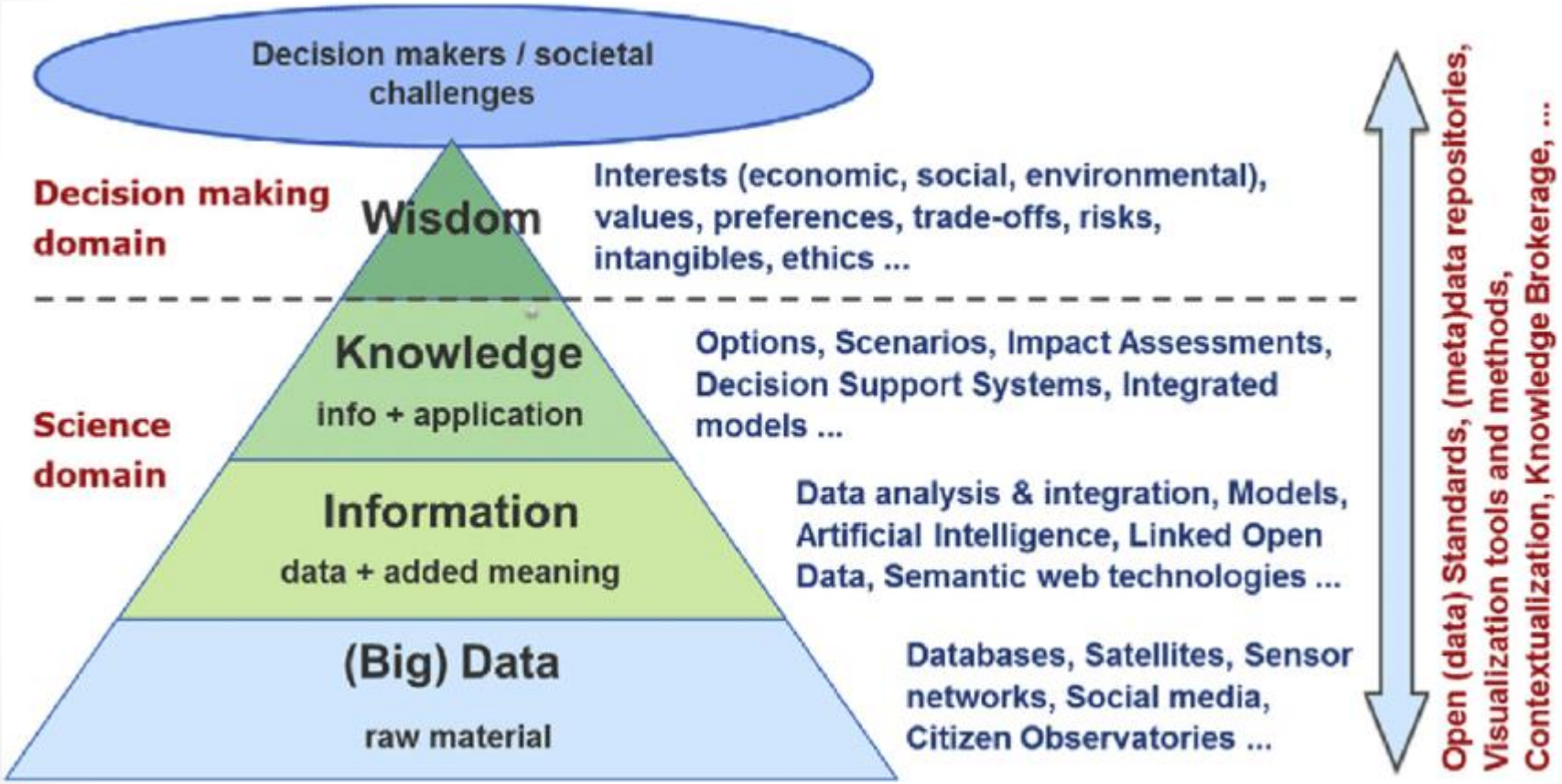
- Une donnée est précise (fait brut), sa recherche dans une BD (*relationnelle par exemple*) relève du requêtage (*select * from … where …*)
- Quand les données deviennent :
  - trop importantes (quantité énorme, dépassant les teraoctets),
  - Hétérogènes : de différents types/formats (textuelles, numéraire, date, image, symbôles, graphiques,…)
  - distribuées et de différentes sources (BD, fichiers semi-structurés ou non structurés, fichiers d'applications, documents,…)

=> On parle de **données massives**

Données massives

- **Problème de ces données massives** :
  - On ne peut pas stocker ces données dans des BD relationnelles car elles ne supporterons pas la quantité énorme des données
  - L'hétérogéneité des données et la diversité de leurs sources compliquent leur traitement (sélection, calcul, filtrage,…)
  - Le temps de réponse pour rechercher une donnée est incalculable
- La gestion de ces données massives (stockage, recherche / utilisation) demande des architectures et technologies particulières pour supporter cette quantité énorme
- => cela relève du domaine du **Big Data**

Big Data VS Information VS Connaissance



Problématiques du module

La RI

- **Problématique de la RI:**
  - On produit chaque jour énormément de quantité d'information sous différentes formes (texte, images, vidéos,…)
  - Les sources d'information n'arrêtent pas de s'accroître autant chez les professionnels (sociétés, administrations, industries,…) que chez les particuliers
  - On trouve encore plus d'informations sur le web
  - Des moteurs de recherche importants existent pour la recherche d'information sur le web…

La RI

- mais quand est-il de la RI dans le milieu des entreprises? Des administrations? En industrie? Pour les fichiers et documents issues d'applications métier professionnelles?
- **Deux techniques sont possibles:**
  1) Rechercher directement dans les fichiers/documents jusqu'à trouver ceux répondant à nos besoins
     - **Problème:**
       - pas réalisable dans le cas où les documents ne sont pas accessibles entièrement en lecture à l'utilisateur
       - la recherche manuelle peut prendre un temps interminable…
       - Il est possible d'utiliser la barre de recherche qu'on trouve dans certains documents (.docx,.pdf,…) mais les documents peuvent être longs et chargés en informations… le temps de recherche est incalculable

La RI

2) Utiliser un système de recherche d'information
   - **Problème:**
     - Si le système recherche l'information (*par mot clé par exemple*) dans le contenu de chaque document, le temps de réponse sera très long (inacceptable)
     - On recherche une information noyées dans d'autres informations => comment retrouver l'information pertinente recherchée par l'utilisateur

La RI

- **Solution:**
  - Il existe des techniques standards utilisées dans la RI pour qu'un moteur fonctionne de manière rapide et efficace
  - Parmi ces techniques:
    - Créer une représentation de chaque document (indexation)
    - Rechercher l'information dans les représentations des documents (index) et pas dans le contenu de chaque document
    - Utiliser des modèles spécifiques de RI pour obtenir la meilleure correspondance possible entre la requête de l'utilisateur et les résultats possibles
    - Classer les résultats et prendre en compte le contexte de l'utilisateur

RI et GED

- **Remarque:**
  - Il existe ce qu'on appelle la **Gestion Electronique de Documents (GED)** ou *Document Management System (DMS)*
  - Il s'agit de la gestion du cycle de vie des documents (de leur création à leur classification, diffusion et archivage).
  - Le processus de gestion du cycle de vie comporte des étapes qui peuvent être automatisée grâce à des logiciels de GED
  - Parmi les fonctionnalités qu'on trouve dans les logiciels de GED :
    - Acquisition des documents (ajout/import manuel ou automatique, création de documents)
    - Organisation (classification dans une arborescence, description)
    - Indexation (automatique ou manuelle) pour la recherche
    - Recherche d'information dans les documents

    } **RI**

    - Circuit de validation (pour automatiser une partie du cycle de vie du document)
    - Diffusion / partage de documents

RI et GED

- Système de workflow (pour automatiser une partie du cycle de vie du document)
- Création d'espace de travail, co-rédaction
- Gestion de version
- Gestion des droits d'accès
- Numérisation
- …

- Parmi les outils connus: Microsoft SharePoint, Zeendoc, Oodrive, DocuWare, M-Files, Alfresco, Nuxeo,…

- **Remarque:** Avant de créer un moteur de recherche pour des besoins spécifiques en entreprise, il est indispensable de vérifier si le besoin concerne uniquement la RI ou s'il s'agit de gestion de documents de manière plus globale (dans ce dernier cas, choisir un outil de GED)

La RI

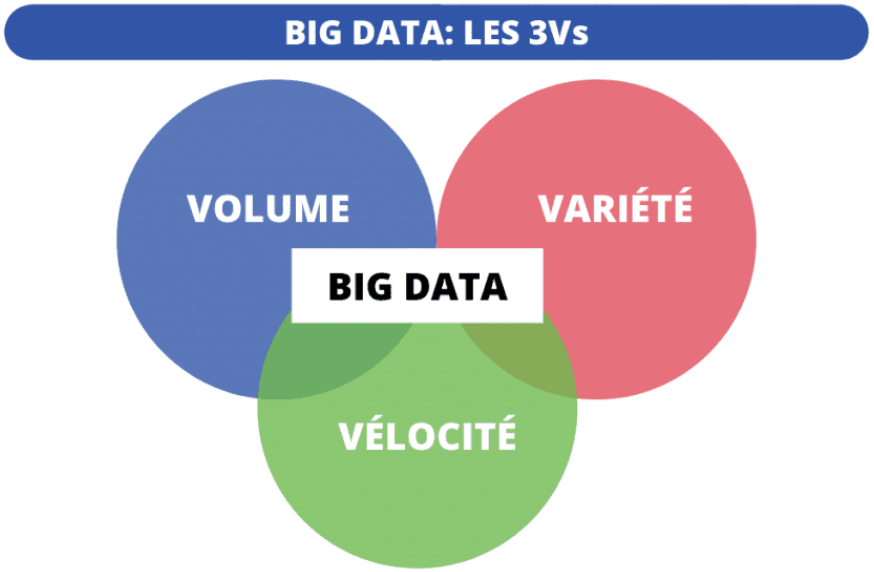
- **Conclusion sur la RI:**
  - L'objectif de cette partie du cours est de savoir identifier une problématique de RI pour un contexte métier précis et savoir quelles solutions de SRI est adéquate
  - Les techniques de bases pour créer un SRI seront étudiées

Le Big Data

- **Problématique du Big Data:**
  - Génération de quantité énorme de données chaque minute, chaque heure, chaque jour, chaque mois, chaque année
  - Ces données proviennent de différentes sources:
    - Sites web d'information, site web commerciaux, …
    - Réseaux sociaux
    - Applications mobiles
    - Caméra, Capteurs, objets connectés
    - Systèmes d'informations des entreprises, ERP, CRM
    - Machines industrielles, automates, appareils médicaux
    - …
  - => on parle de **données massives** ou de **mégadonnées (Big Data)**

Le Big Data

- Les données du Big Data sont caractérisées par :
  - Leur volumétrie : quantité énorme
  - Leur variété : différents formats et types
  - Leur vélocité: besoin de les traiter rapidement



Le Big Data

- Ces caractéristiques soulèvent deux problématiques majeurs en Big Data:
  - **Le stockage** : Comment gérer le stockage en continue de ces données (sachant que le volume est de plus en plus exponentiel)

    => on parle du **Big Data Engineering**
  - **L'analyse** : Comment analyser et traiter ces données (faire des corrélations entre elles) pour en tirer des informations et du sens

    => on parle du **Big Data Analytics** (requêtage specifique, BI, méthodes statistiques, data mining,…). On peut aussi utiliser la **Data Science** (modèles mathématiques et statistiques, techniques d'AI, machine learning, …)

N.B** : **On peut appliquer la data science sur des données qui ne sont pas qualifiées de Big Data

Le Big Data

- **Problème:**
  - Avec ces exigences, on se rend compte que les BD traditionnelles sont limitées et ne peuvent pas évoluer avec le volume exponentiel du Big Data
  - Les applications et leur architecture conçues il y a des années ne peuvent pas supporter le traitement de cette quantité énorme de données
- **Solutions:**
  - Platefomes capables de traiter le Big Data sur une architecture distribuées (Hadoop, Spark, …)
  - Systèmes de stockage adaptés aux données massives: BD NoSQL (Cassandra, Redis, Hbase,…), Système de fichiers distribués (*HDFS*), privilégier le stockage à distance (cloud storage & computing)
  - Traitement parallèle et distribué des données : Modèle du MapReduce de Google

Conclusion

- Ce cours est un tour d'Horizon sur :
  - **La RI et les SRI :**
    - Les techniques standards utilisées pour créer des moteurs de recherche d'information efficaces et savoir évaluer la pertinence de leurs résultats
  - **Le Big Data:**
    - L'intérêt du Big Data et les technologies utilisées aujourd'hui pour gérer les données massives (stockage et traitement)