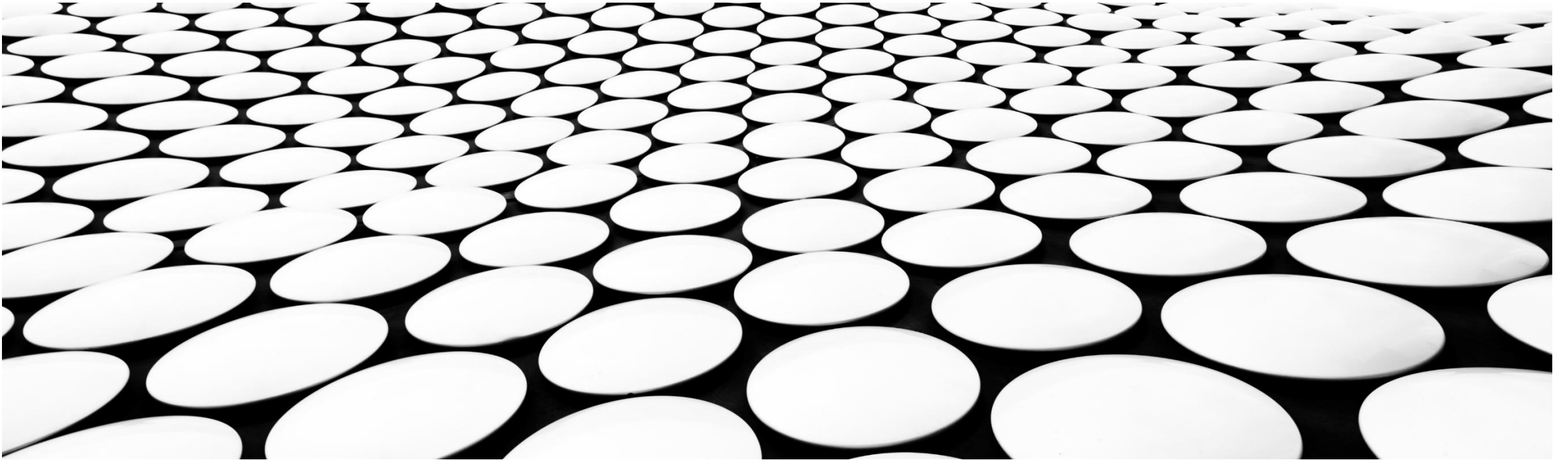


R5.C.07: *DONNEES MASSIVES*



CONTENU DU COURS

- L'objectif de cette ressource est d'écrire un scénario complet pour comprendre les étapes de la chaîne de la donnée: collecte, exploration et analyse, et visualisation.
- Savoirs de référence étudiés
 - Collecte des données (différents formats) et nettoyage
 - Exploration et analyse (fouille simple)
 - Visualisation
 - Les différents savoirs de référence pourront être approfondis

ENSEIGNANTS

Enseignants:

- Mario CATALDI

Volume:

- 15h environs

MATÉRIELS

Matériels du cours: **Slides, énoncés et solutions TD et TP** disponibles dans la **page Moodle** du cours.

<https://moodle.iut.univ-paris8.fr/enrol/index.php?id=746>

Quelques livres suggérés:

*P. Delort, « **Le Big Data** ». Presses Universitaires de France, 2015*

*P. Lemberger, M. Batty, M. Morel, J.L. Raffeëlli, « **Big Data et Machine Learning** », Dunod, 2015*

*S. Abiteboul, I. Manolescu, P. Rigaux, M-C Rousset, P. Senellart, « **Web Data Management** », Cambridge University Press 2011 (en ligne : <http://webdam.inria.fr/Jorge/?action=chapters>)*

CONTRÔLES & ABSENCES

ABSENCES au contrôle

- > *absence justifiée: me contacter par mail pour rattrapage*
- > *absence injustifiée: note de 0.*

INTRODUCTION: BESOIN DES DONNEES ET ANALYSE

Actuellement nous produisons annuellement une masse de données estimée à près de 3 trillions (3 millions de millions) d'octets de données.

On estime ainsi que 90% des données dans le monde ont été créées au cours des 2 années précédentes!

Tous les secteurs sont touchés, tant scientifiques qu'économiques, ainsi que le développement des applications Web et les réseaux sociaux.

Dans ce contexte, est apparu le terme **Big Data** .

L'origine de ce terme anglo-saxon, littéralement « *grosses données* », est controversée, et sa traduction française officielle recommandée est *mégadonnées* ou *données massives*.

www.wipro.com

BIG DATA

Big Data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze.

The right use of Big Data allows analysts to spot trends and gives niche insights that help create value and innovation much faster than conventional methods.

THE "THREE V'S", i.e the Volume, Variety and Velocity of data coming in is what creates the challenge.

VOLUME

Amount of Big Data stored across the world (in petabytes)

Region	Volume (petabytes)
NORTH AMERICA	>3,500
EUROPE	>2,000
CHINA	>250
JAPAN	>400
MIDDLE EAST	>200
INDIA	>50
LATIN AMERICA	>50

VARIETY

Category	Examples
PEOPLE TO PEOPLE	NETZETNS, VIRTUAL COMMUNITIES, SOCIAL NETWORKS, WEB LOGS...
PEOPLE TO MACHINE	ARCHIVES, MEDICAL DEVICES, DIGITAL TV, E-COMMERCE, SMART CARDS, BANK CARDS, COMPUTERS, MOBILES...
MACHINE TO MACHINE	SENSORS, GPS DEVICES, BAR CODE SCANNERS, SURVEILLANCE CAMERAS, SCIENTIFIC RESEARCH...

VELOCITY

Category	Volume/Rate
EMAILS SENT EVERY SECOND	2.9 MILLION
OF VIDEO UPLOADED EVERY MIN	20 HOURS
TWEETS PER DAY	50 MILLION

VALUE

Industry	Productivity Increase	Sales Increase
RETAIL	49%	\$9.6B
CONSULTING	39%	\$5.0B
AIR TRANSPORTATION	21%	\$4.3B
CONSTRUCTION	20%	\$4.2B
FOOD PRODUCTS	20%	\$3.4B
STEEL	20%	\$3.4B
AUTOMOBILE	19%	\$2B
INDUSTRIAL INSTRUMENTS	18%	\$1.2B
PUBLISHING	18%	\$0.8B
TELECOMMUNICATIONS	17%	\$0.4B

40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR

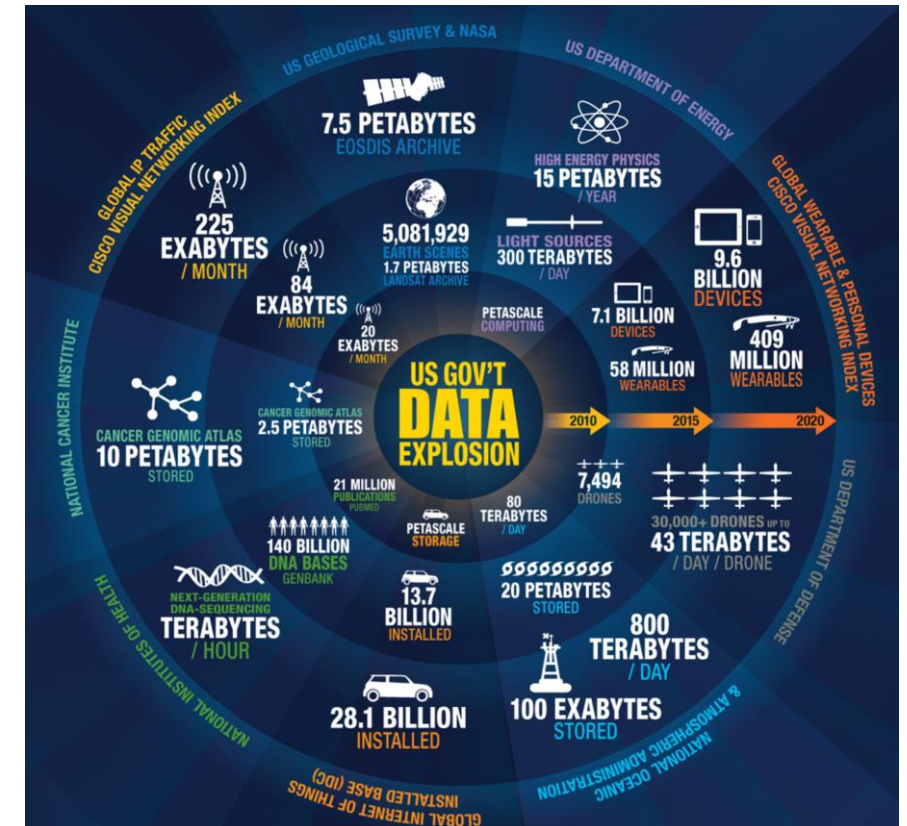
5% PROJECTED GROWTH IN GLOBAL IT SPENDING PER YEAR

The estimated size of the digital universe in 2011 was 1.8 zettabytes. It is predicted that between 2009 and 2020, this will grow 44 fold to 35 zettabytes per year. A well defined data management strategy is essential to successfully utilize Big Data.

Source: • Measuring the Rewards of Big Data • Wipro Report • Big Data: The Next Frontier for Innovation, Competition and Productivity • McKinsey Global Institute Report • Governance, Industrial Group • Measuring the Business Impacts of Effective Data - study by University of Texas, Austin • IIS Department of Labor

DO BUSINESS BETTER

WIPRO



INTRODUCTION: BESOIN DES DONNEES ET ANALYSE

Deux grandes problématiques:

- **Stockage et gestion** (Big Data Engineering):

les techniques traditionnelles de stockage de type bases de données relationnelles ne permettant pas de stocker de telles quantité de données

- **Analyse** (Big Data Analysis):

Cela repose essentiellement sur des méthodes d'apprentissage et de calcul distribué

INTRODUCTION: 5 V

On parle souvent de 3 « V » :

Volume,
Variété,
Vélocité

Mais on ajoute 2 autres « V » complémentaires :

Valeur
Validité



1 - VOLUME DES DONNEES

Fait référence à la quantité d'informations, *trop volumineuses* pour être acquises, stockées, traitées, analysées et diffusées par des outils standards,

Peut s'interpréter comme le *traitement d'objets* informationnels *de grande taille* ou de grandes collections d'objets,

Le développement de l'IoT (Internet des Objets) et la généralisation de la géolocalisation ou de l'analytique ont *engendré une explosion du volume de données collectées*,

On estime qu'en 2020, *43 trillions de gigabytes* ont été générés, soit 300 fois plus qu'en 2002.
(1 trillion = un milliard de milliards de bytes)

2 - VARIETE DES DONNEES

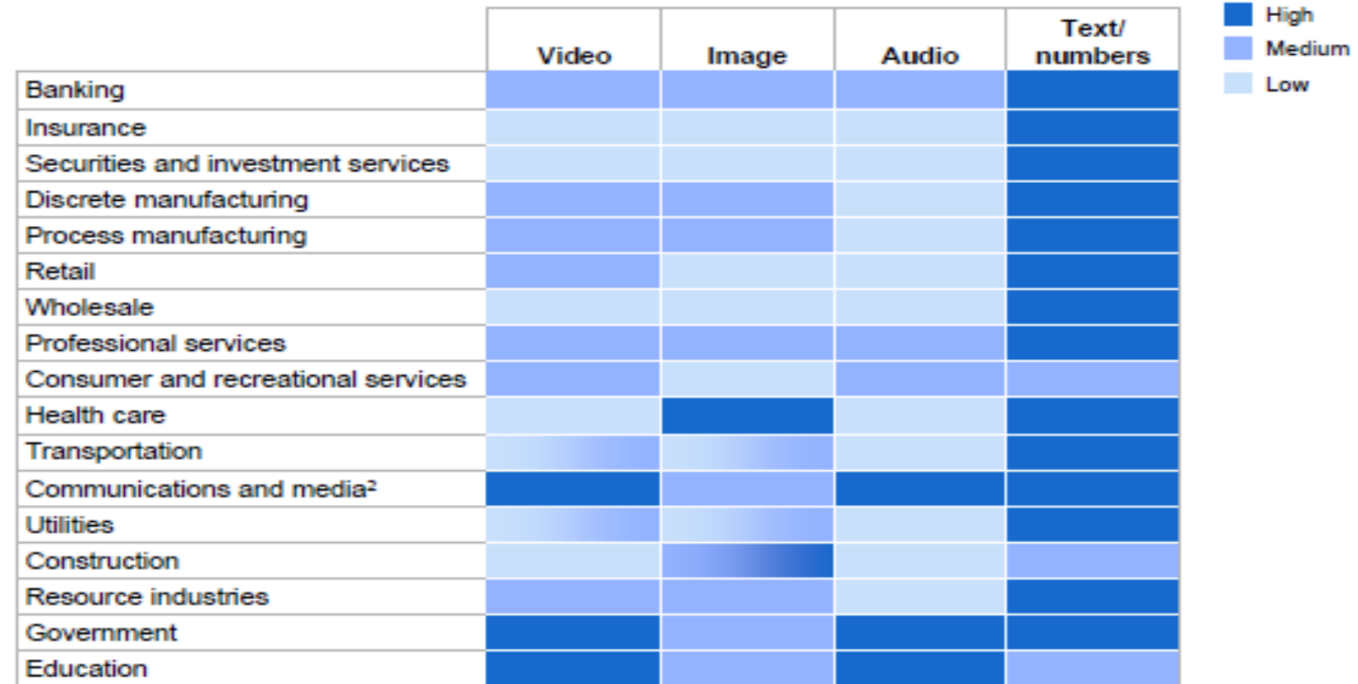
Fait référence à l'**hétérogénéité** des **formats**, **types**, et **qualité** des informations

- Est lié au fait que ces données peuvent présenter des formes complexes du fait qu'elles trouvent leurs origines dans :
 - des capteurs divers et variés
 - des messages échangés (e-mails, médias sociaux, etc),
 - des textes, des publications en ligne (bibliothèques numériques, sites web, blogs, ...),
 - images, vidéos, sons
 - enregistrements de transactions d'achats, des plans numérisés,
 - des annuaires, des informations issues des téléphones mobiles,
 - etc.
- Usage de technologies nouvelles pour analyser et recouper les données non structurées (mails, photos, conversations...) représentant au moins 80 % des informations collectées.

2 - VARIETE DES DONNEES

Exemple de type de données générées par secteur d'activité:

The type of data generated and stored varies by sector¹



¹ We compiled this heat map using units of data (in files or minutes of video) rather than bytes.

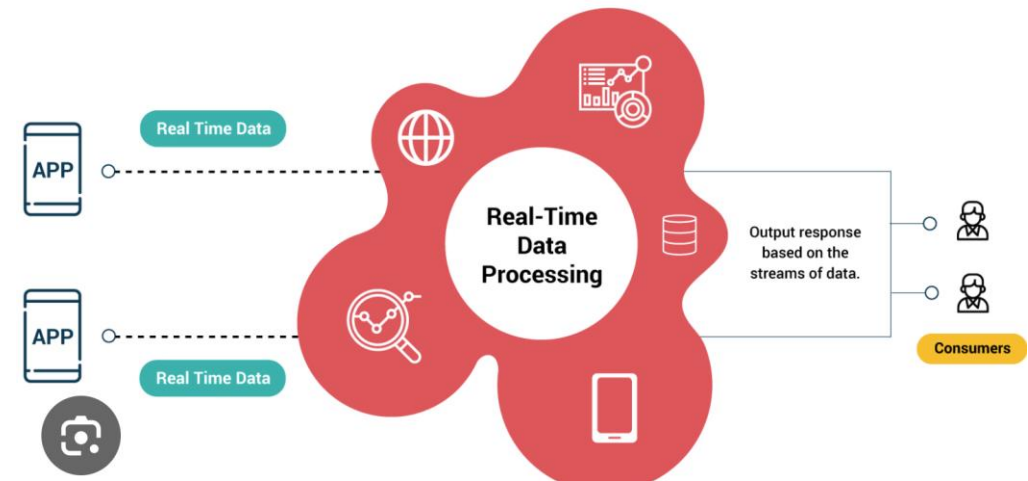
² Video and audio are high in some subsectors.

SOURCE: McKinsey Global Institute analysis

3 – VELOCITE DES DONNEES

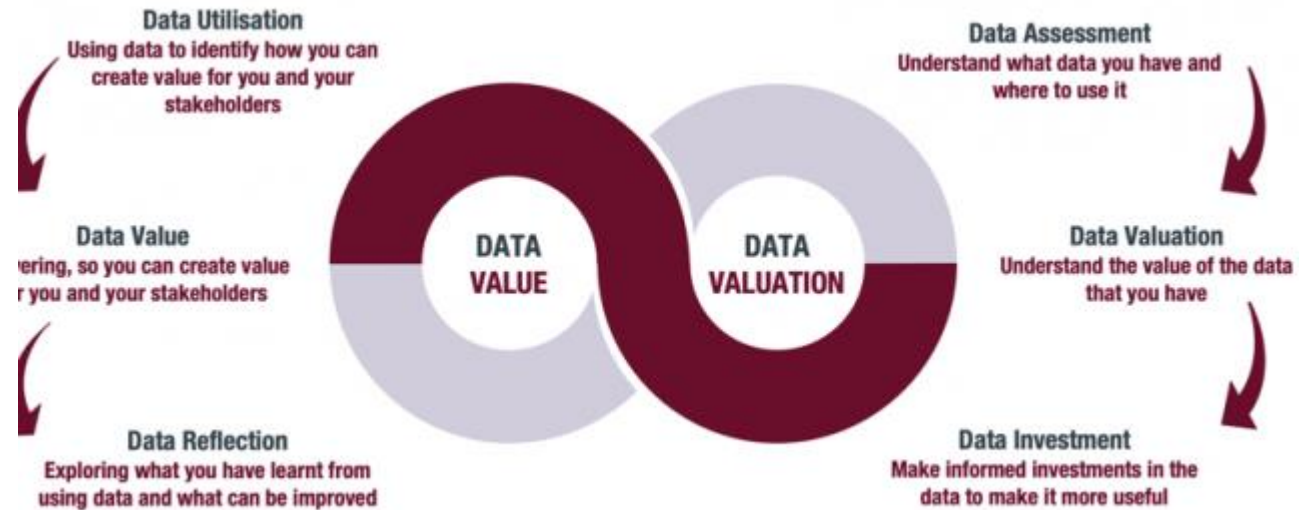
Fait référence à l'aspect **dynamique** et/ou **temporel** des données, à leur delai d'actualisation et d'analyse:

- les données ne sont plus traitées, analysées, en différé mais en **temps réel (ou quasi réel)**
- elles sont produites en flots continus
- ce sont les données notamment issues de capteurs, nécessitant un traitement rapide pour une réaction en temps réel
- dans le cas de telles données de grande vélocité engendrant des volumes très importants, il n'est plus possible de les stocker en l'état, mais seulement de les analyser en flux (streaming)



4 – VALEUR DES DONNEES

- Associé à l'usage qui peut être fait de ces données massives, de leur analyse, notamment d'un point de vue économique.
- L'analyse de ces données massives demande une certaine expertise tant liée à des méthodes et techniques en statistique, en analyse de données, que de domaine pour l'interprétation de ces analyses.



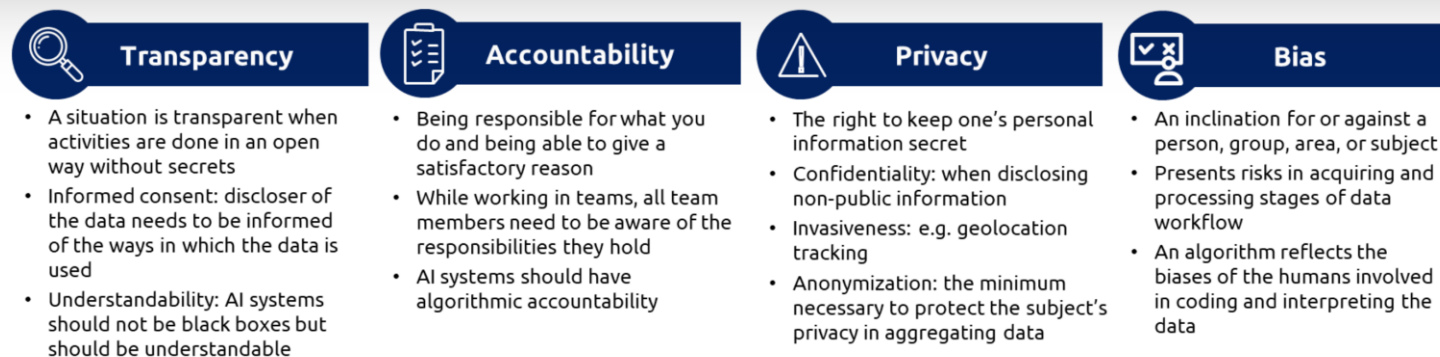
Les termes « Data Scientist » et « Data Science » sont liés à cette expertise recherchée et à cette nouvelle discipline émergente.

5 – VALIDITE DES DONNEES

Fait référence à la **qualité** des données et/ou aux **problèmes éthiques** liés à leur utilisation. Il comprend les problèmes de valeurs aberrantes ou **manquantes** (ces problèmes pouvant être résolus par le volume de données)

Exemple: Imaginons d'utiliser nos données pour des buts prédictifs. Quelle pourrait être l'évolution de la carrière de chaque employé? Pour ce faire, nous utilisons les données existantes de l'entreprise. Mais celles-là sont biaisés par des problèmes éthiques historiquement non traités (carrière des femmes, accès aux études différents, cadres presque toujours hommes, etc.).

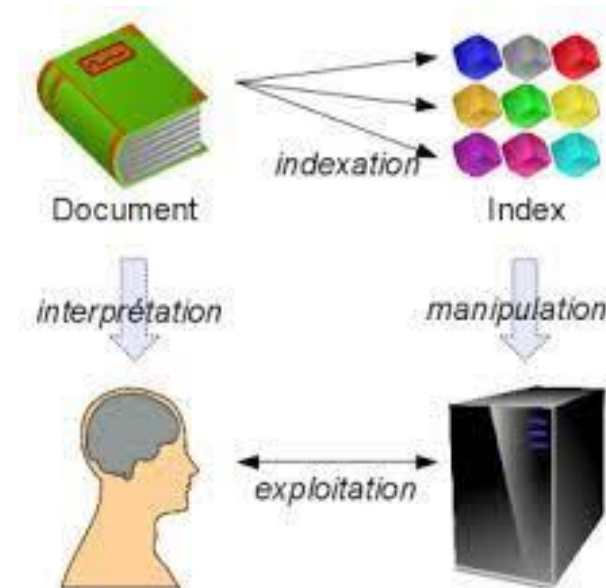
L'utilisation de ces données impliquerait alors une déduction discriminatoire due à de données biaisées.



INDEXATION: PREMIERE ETAPE DE L'ANALYSE

La première étape de l'analyse des données massives consiste en les **indexer**. Autrement dit, les analyser pour en extraire seulement l'information essentielle nécessaire (en supprimant le reste).

De plus, l'objectif est de se préparer pour en donner un accès rapide.

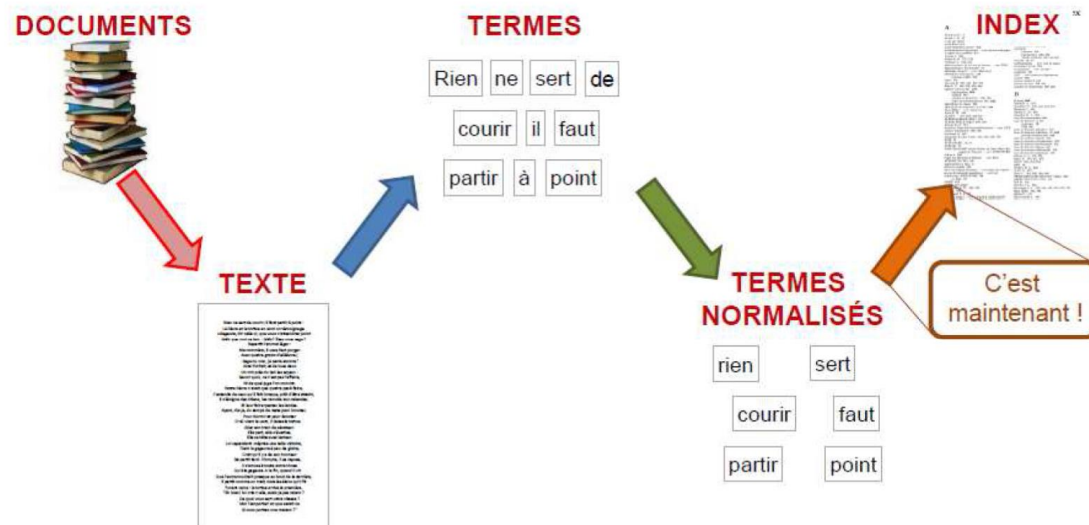


INDEXATION: PREMIERE ETAPE DE L'ANALYSE

Indexation: transformer des documents en **substituts** capables de représenter le contenu de ces documents (Salton et McGill, 1983)

Les index peuvent prendre plusieurs formes : mots simples, mots complexes, entrées de thésaurus.

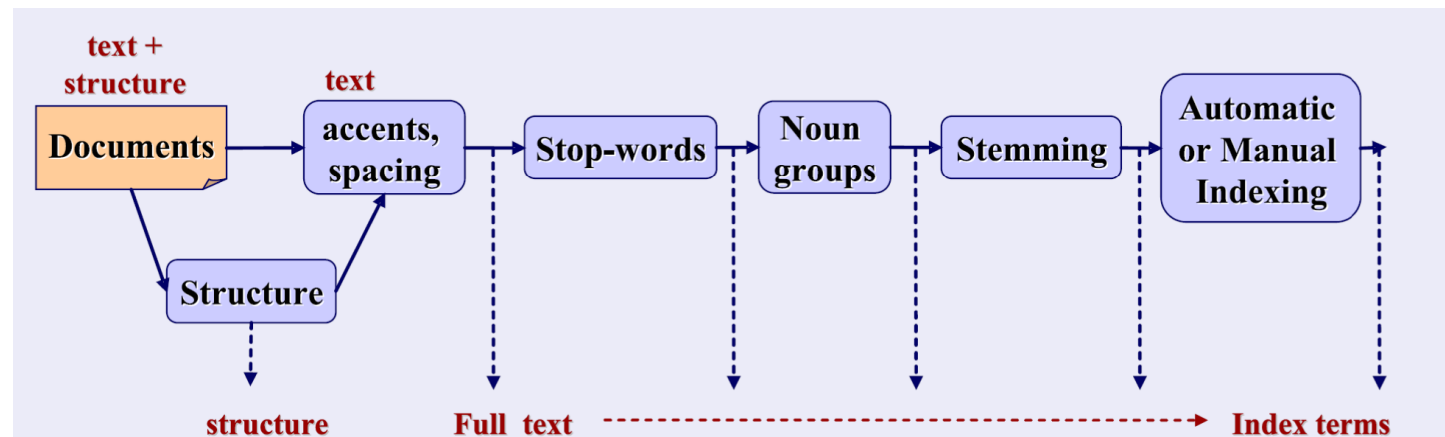
L'indexation est souvent précédée par un filtrage/transformation d'information. L'indexation doit être AUTOMATIQUE!



INDEXATION: PREMIERE ETAPE DE L'ANALYSE

Exemple: pour un document textuel; un ensemble de termes d'index ou de mots clés.

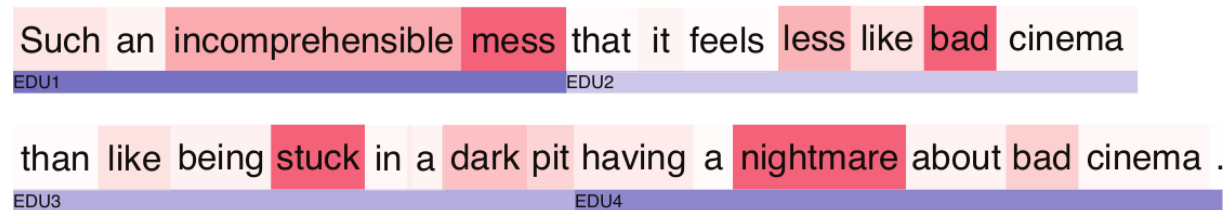
- Information inutile supprimée (**STOP-WORDS**): articles, conjonctions, ...
- **Pré-filtrage**: Analyse des groupes de noms en fonction des objectifs: élimination des adjectifs, verbes, adverbes ..
- **Stemming** (ou Lemmatisation) : transformation d'un mot en sa forme canonique (pour réduction de l'information)



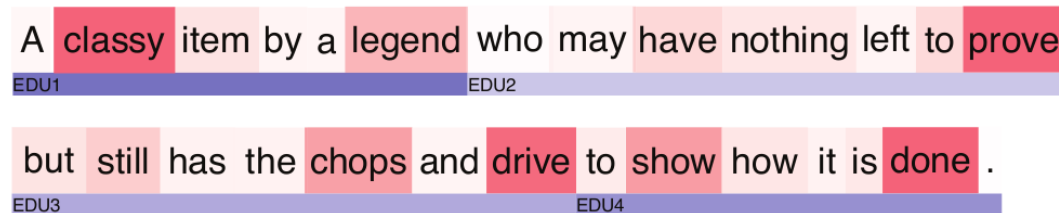
INDEXATION: BAG OF WORDS

Comment reconnaît-on un mot?

- **Segmentation** en unités (espaces, tirets, ponctuation)
- **Stop lists**: suppression mots non intéressants.
- **Stemming**: procédure de troncature pour trouver forme générale.



(a) A sentence (negative) from the MR test dataset.



(b) A sentence (very positive) from the SST-5 test dataset.

INDEXATION: MATRICE

L'objectif de l'indexation reste celui de sauvegarder **où** chaque mot est **présent**.

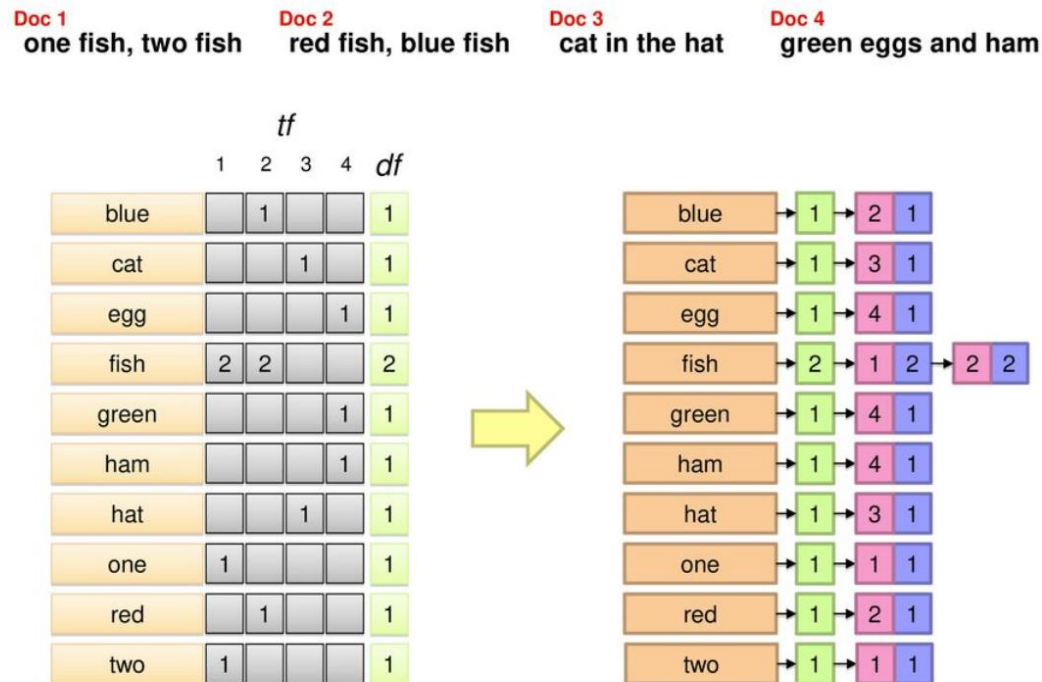
Matrice d'incidence: chaque ligne est un document (identifiant d'un document), chaque colonne un mot.
Matrice très creuse!

	↕ about ↴	↕ against ↴	↕ amazing ↴	↕ america ↴	↕ america-first ↴	↕ american ↴	↕
1	0	0	0	0	0	0	
2	0	0	0	0	0	0	
3	0	0	0	0	0	0	
4	0	0	0	0	0	0	
5	0	0	0	0	0	0	
6	0	0	0	0	0	0	
7	0	0	0	0	0	0	
8	0	0	0	0	0	0	
9	0	1	0	0	0	0	
10	0	0	0	0	0	0	

INDEXATION: INDEX INVERSE'

La matrice d'incidence est à la base de la construction de **l'index inversé**. Pour chaque mots, on sauvegarde où il est présent.

Inverted Index: TF.IDF



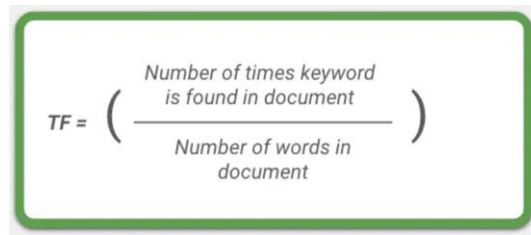
INDEXATION: FREQUENCE D'UN MOT

De plus, l'objectif est celui de stocker une valeur de fréquence normalisée par chaque document. Le TF*IDF pour « *Term Frequency * Inverse Document Frequency* » est une méthode qui permet de déterminer la pertinence d'un document par rapport à un terme, normalisée sur l'ensemble de documents.

Elle se base d'abord sur le calcul de fréquence (i.e. *Combien de fois un mot est présent dans un document?*)

TF: *Term frequency*

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$


$$TF = \left(\frac{\text{Number of times keyword is found in document}}{\text{Number of words in document}} \right)$$

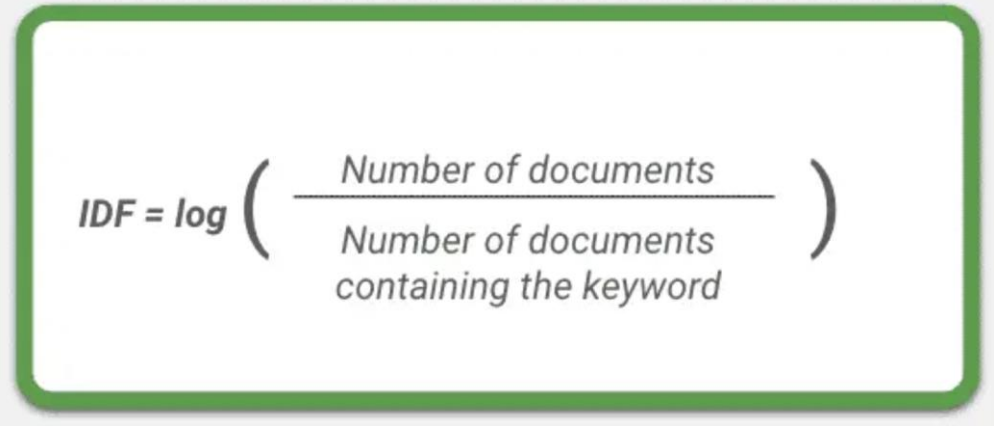
Ou « fréquence brute »

Schéma de pondération	formule du TF
binaire	0, 1
fréquence brute	$f_{t,d}$
normalisation logarithmique	$\log(1 + f_{t,d})$
normalisation « 0.5 » par le max	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
normalisation par le max	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

INDEXATION: FREQUENCE D'UN MOT

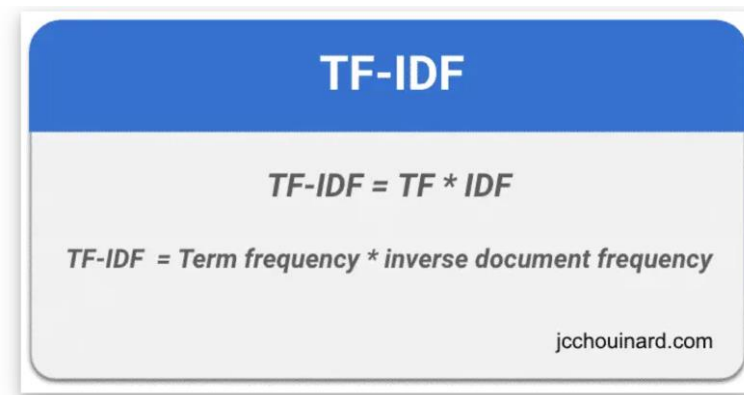
La fréquence inverse de document (*inverse document frequency*) est une mesure de l'importance du terme dans l'ensemble du corpus. Elle vise à réduire le poids des mots le plus utilisés (trop générique pour contextualiser un argument spécifique).

$$\text{idf}_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$


$$IDF = \log \left(\frac{\text{Number of documents}}{\text{Number of documents containing the keyword}} \right)$$

INDEXATION: TF-IDF

Le TF-IDF vise à combiner les deux méthodes.



INDEXATION: EXEMPLE

Corpus (tiré d'œuvres de [Friedrich Gottlieb Klopstock](#))²

Document 1	Document 2	Document 3
Son nom est célébré par le bocage qui frémit, et par le ruisseau qui murmure, les vents l'emportent jusqu'à l'arc céleste, l'arc de grâce et de consolation que sa main tendit dans les nuages.	À peine distinguait-on deux butts à l'extrémité de la carrière : des chênes ombrageaient l'un, autour de l'autre des palmiers se dessinaient dans l'éclat du soir.	Ah ! le beau temps de mes travaux poétiques ! les beaux jours que j'ai passés près de toi ! Les premiers, inépuisables de joie, de paix et de liberté ; les derniers, empreints d'une mélancolie qui eut bien aussi ses charmes.

Source: wikipedia

Considérons un terme générique: **qui**. Il est présent 2 fois dans le document 1 et une fois dans le document 3. Donc présent dans 2 documents sur 3.

$$tf_{1,1} = \frac{n_{1,1}}{\sum_k n_{k,1}} = \frac{2}{38}$$

$$idf_1 = \log \frac{|D|}{|\{d_j : t_1 \in d_j\}|} = \log \frac{3}{2}$$

$$tfidf_{1,1} = \frac{2}{38} \cdot \log \frac{3}{2} \approx 0,0092$$

$$tfidf_{1,2} = 0 \cdot \log \frac{3}{2} = 0$$

$$tfidf_{1,3} = \frac{1}{40} \cdot \log \frac{3}{2} \approx 0,0044$$

FOUILLE DE DONNEES: MOTEURS DE RECHERCHE

Un moteur de recherche est une application permettant à un utilisateur de trouver des ressources à partir d'une requête (composée de terme ou autre).

Sur le principe, ils fonctionnent généralement :

- avec une indexation des contenus dans une ou plusieurs base de données (indexation effectuée préalablement à la recherche).

FONCTIONNEMENT (3 PHASES):

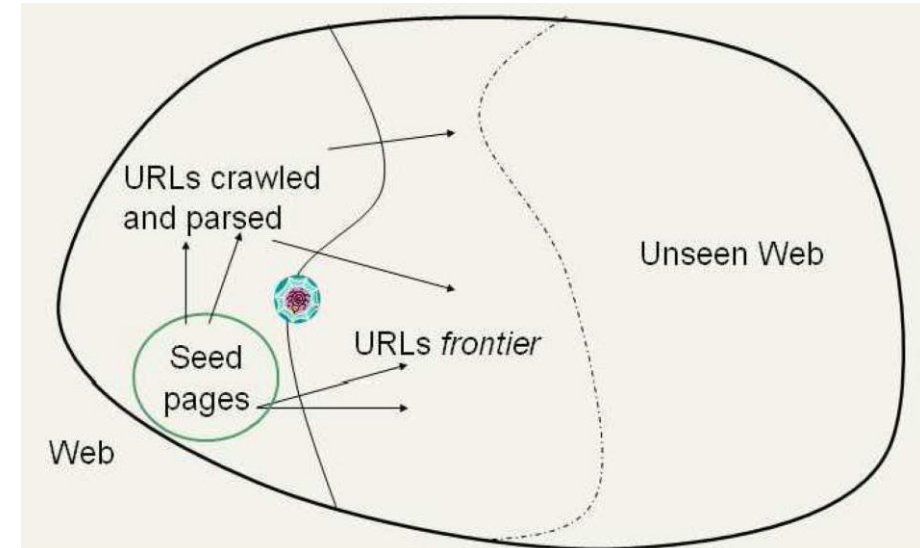
- 1) **L'exploration ou crawl** : le web est systématiquement exploré par un robot d'indexation et récupérant les ressources jugées intéressantes.
- 2) **L'indexation** des ressources récupérées consiste à extraire les mots considérés comme significatifs du corpus à explorer.
- 3) **La recherche** des documents qui correspondent le mieux aux mots contenus dans la requête, afin de présenter les résultats des recherches par ordre de pertinence supposée

EXPLORATION OU CRAWL: LE WEB

Prenons en exemple le WEB: suivre les liens à partir d'un ensemble de pages initiales (*seed pages*).

Pour chaque page retrouvée, on analyse sa structure, on va extraire les informations importantes et on procède avec l'indexation.

PRINCIPE DE ROBUSTESSE: éviter les pièges à araignées.
PRINCIPE DE POLITESSE : politiques implicites et explicites avec les serveurs web pour réguler le nombre de visites.
Ne pas indexer ceux qui ne veulent pas être indexé.

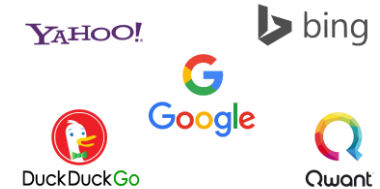


Composantes souhaitables

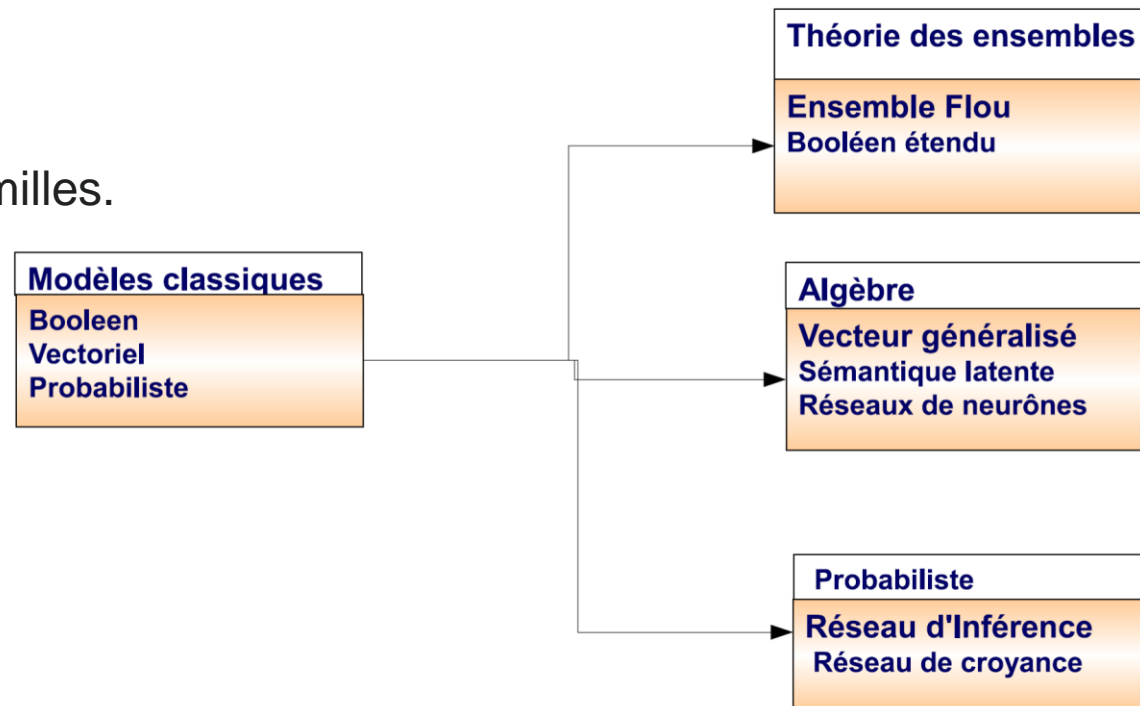
- **Principe de sélection** : choix des pages à charger
- **Principe de re-visite** : quand vérifier si il y a des changements dans les pages. Un travail continu
- **Principe de distribution** : coordination des robots d'indexations distribués
- **Principe de passage à l'échelle** : support de nouvelles machines
- **Principe de performance et d'efficacité** : utilisation efficace des différentes ressources
- **Principe d'extensibilité** : nouveaux formats, nouveaux protocoles

MOTEURS DE RECHERCHE

Les moteurs de recherche Internet précèdent les débuts du Web fin 1990.



Trois grandes familles.



MOTEURS DE RECHERCHE: MODELE THEORIQUE

Modèle

$$< D, Q, F, R(q_i, d_i) >$$

- D - ensemble des vues logiques des documents
- Q - ensemble des vues logiques des besoins de l'utilisateur : requêtes
- F - framework de modélisation des documents, des requêtes et de leurs relations
- $R(q_i, d_j)$ - fonction de ranking : associe un nombre réel à une requête $q_i \in Q$ et une représentation de document $d_j \in D$. Il permet de définir un ordre parmi les documents en rapport avec la requête q_i


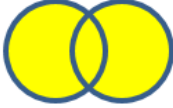
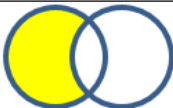
MOTEUR BOOLEEN

Modèle le plus simple basé sur la théorie des ensembles. Mise en correspondance exacte.

Modèle :

- Les documents sont trouvés si ils satisfont une expression booléenne
- Classement en fonction du choix.

How to Search Using Boolean Operators:

Concept	Search Examples	Results
AND	politics AND media children AND poverty "civil war" AND Virginia	 Results will include both terms
OR	"law enforcement" OR police labor OR <u>labour</u> 60s OR sixties	 Results will include one or both terms
NOT	"civil war" NOT American Caribbean NOT Cuba therapy NOT physical	 Excludes results with the term following NOT

MOTEUR BOOLEEN

Operateurs logiques usuels.

conjonction

\wedge	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Disjunction

\vee	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Equivalence

\Leftrightarrow	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Implication

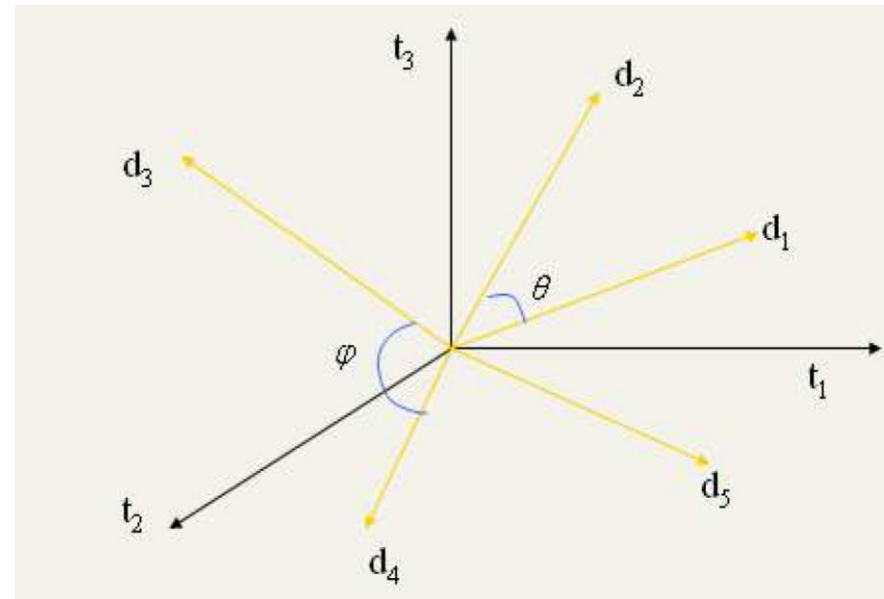
\Rightarrow	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

MOTEUR VECTORIEL

Principe de base: représentation des documents comme des vecteurs. Chaque document peut être vu comme un vecteur de valeurs. On a une composante par terme.

Les termes sont les axes. Les documents sont des vecteurs dans l'espace n-dimensionnel.

Les documents qui sont proches dans l'espace vectoriel pourraient traiter de sujets similaires (utilisent les mêmes termes)

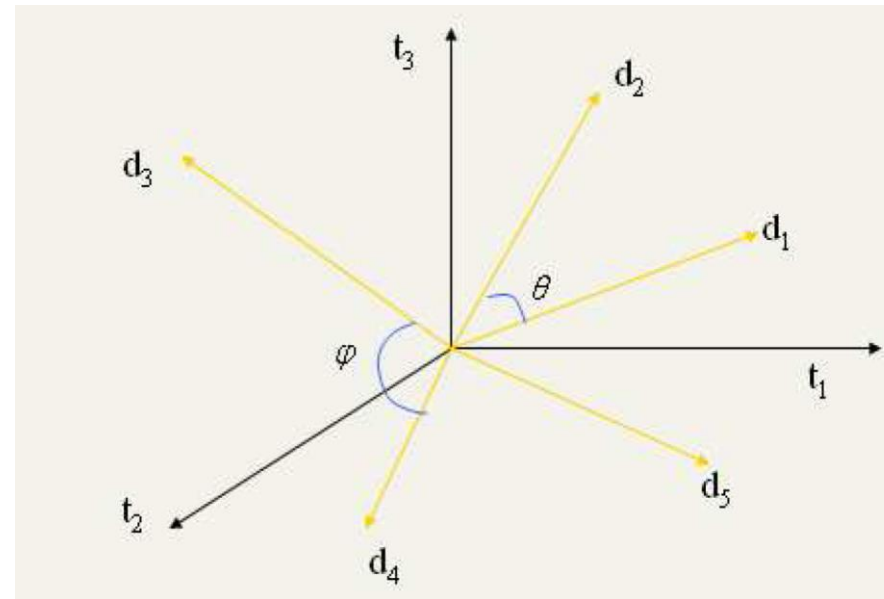


MOTEUR VECTORIEL

Principe de base: représentation des documents comme des vecteurs. Chaque document peut être vu comme un vecteur de valeurs. On a une composante par terme.

Les termes sont les axes. Les documents sont des vecteurs dans l'espace n-dimensionnel.

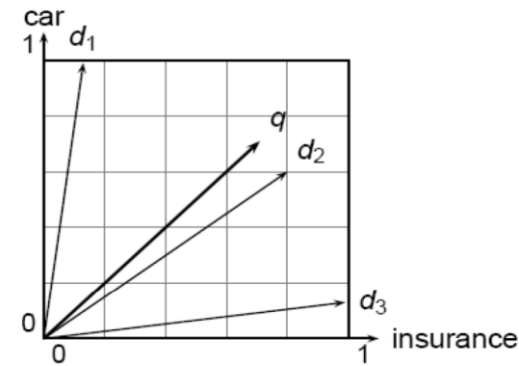
Les documents qui sont proches dans l'espace vectoriel pourraient traiter de sujets similaires (utilisent les mêmes termes)



MOTEUR VECTORIEL

Idée de base: documents et requetes sont visualisé dans cet espace comme 2 vecteurs.
Leur degré de similarité est le cosinus de l'angle entre ces 2 vecteurs.

*Imaginons un espace de 2 mots seulement (**car** et **insurance**). On utilisera le TF (ou TF-IDF) de chaque terme dans chaque document (et dans la requête) pour définir chaque vecteur dans cet espace bi-dimensionnel.*



Mesure de similarité

$$\text{sim}(q, d_j) = \cos(\Theta) = \frac{\sum_{i=1}^n d[i] * q[i]}{\sqrt{\sum_{i=1}^n (d[i])^2} * \sqrt{\sum_{i=1}^n (q[i])^2}}$$

- On normalise par la longueur du document
 - ▶ Document long : termes plus fréquents, plus de termes
- Les documents sont classés par valeur décroissante de cosinus.
 - ▶ $\text{sim}(d, q) = 1$ quand $d = q$
 - ▶ $\text{sim}(d, q) = 0$ quand d et q ne partagent aucun terme

MOTEUR PROBABILISTE

Les probabilités sont un outil naturel pour la quantification de l'incertitude.

Principe

- Modélisation de la recherche comme un **problème de classification**
- Pour chaque requête : deux classes :
 - ▶ Documents *pertinents* : C_P
 - ▶ Documents *non pertinents* : C_{NP}
- Objectif : étant donné un document D , trouver la probabilité du document d'appartenir à C_P . On n'effectue la recherche que si $P(C_P|D) > P(C_{NP}|D)$
- Rang de classement = probabilité d'appartenance : $\frac{P(C_P|D)}{P(C_{NP}|D)}$
- Autant de modèles que de manière d'estimer les probabilités

BINARY INDEPENDENCE RETRIEVAL

On suppose que :

- L'utilisateur a vu le premier classement
- L'utilisateur a labelisé plusieurs des documents comme pertinents (**relevance feedback**)

Etant donné un deux mots i et l'ensemble de documents pertinents R_i

On essaiera de trouver la corrélation avec des autre termes j avec cette formule:

- N sont les documents dans la collection
- R_i le nombre de document pertinent (qui, au moins, contiennent dans le mot i)
- n_j le nombre de documents qui contiennent le mot j
- $r_{i,j}$ le nombre de documents qui contiennent i et j

$$u_{i,j} = \log \frac{r_{i,j}/(R_i - r_{i,j})}{(n_j - r_{i,j})/(N - n_j - R_i + r_{i,j})} \times \left| \frac{r_{i,j}}{R_i} - \frac{n_j - r_{i,j}}{N - R_i} \right|$$

Relevance Feedback implicite!

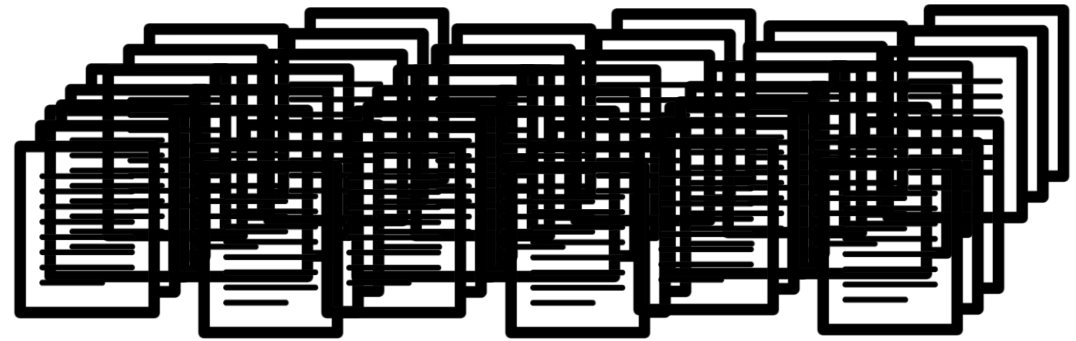
BINARY INDEPENDENCE RETRIEVAL - EXEMPLE

Imaginons d'effectuer la recherche **Foot** sur internet

Documents qui contiennent le
terme PSG



Documents qui ne contiennent pas le terme PSG



L'objectif est de retrouver les termes qui sont beaucoup présents dans l'ensemble contenant **Foot** et très peu dans l'ensemble de documents qui ne le contiennent pas (ex: « **Mbappé** »?)