FONDATION
DE L'ACADEMIE DES TECHNOLOGIES

Trimestriel de l'intelligence technologique #1

Le Big data – entre faits et fantasmes
Qu'en faire ? – Décryptage (*)

Le « Big Data » par l'exemple :

- La société « **Uber** » choisit ses nouvelles implantations grâce à des analyses utilisant les informations ouvertes présentes sur le Net.
- Tout le monde connaît le système « **Coyote** » qui renseigne les automobilistes grâce à leur connexion téléphonique et leur localisation GPS.
- L'observation des habitudes d'usage des cartes bancaires peut permettre d'identifier des comportements illicites par la simple observation macroscopique des pratiques d'usage sans s'intéresser au détail des achats.
- L'analyse des pôles d'intérêt d'une jeune femme sur Internet a permis à un organisme de vente par correspondance de détecter sa probable grossesse avant son environnement.
- Une société de promotion immobilière s'appuie sur des informations obtenues grâce au réseau « **linked in** » (offres d'embauche, changements d'adresse, …) pour déterminer les zones les plus favorables à des investissements immobiliers.

Ces quelques exemples illustrent **la richesse** qui peut se cacher dans la masse énorme de données amassée chez soi ou dans le monde, directement accessibles grâce au « Net » ? ou cachées mais pouvant être rendues disponibles ? et **la variété** de ce qui peut en être tiré.

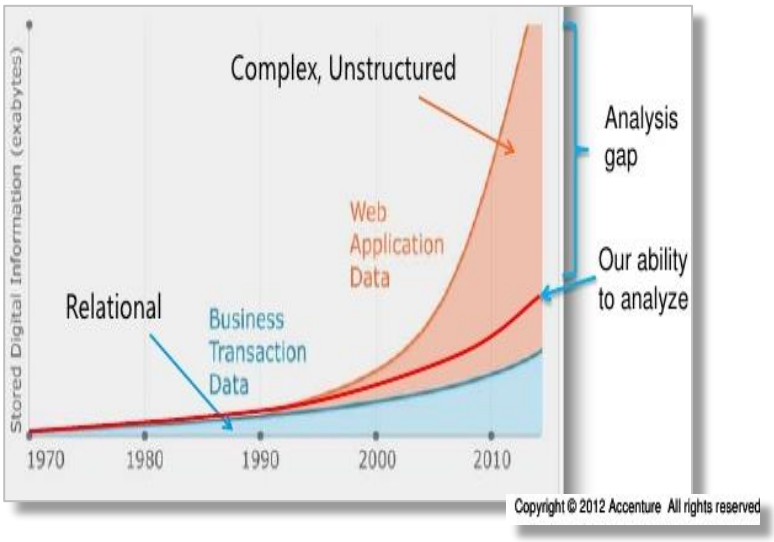
Le « Big Data » - Concept et perspectives :

Dès la constitution des premières civilisations organisées, l'intérêt de mesurer et d'enregistrer les résultats a été reconnu et mis en œuvre (mesure des crues du Nil, tablettes cunéiformes …). Dans le même temps, l'écriture a complété la tradition orale dans son rôle de transmission de l'information. Au fil du temps, l'apparition de l'imprimerie et du télégraphe ont considérablement augmenté capacité et rapidité de transmission de l'information recueillie.

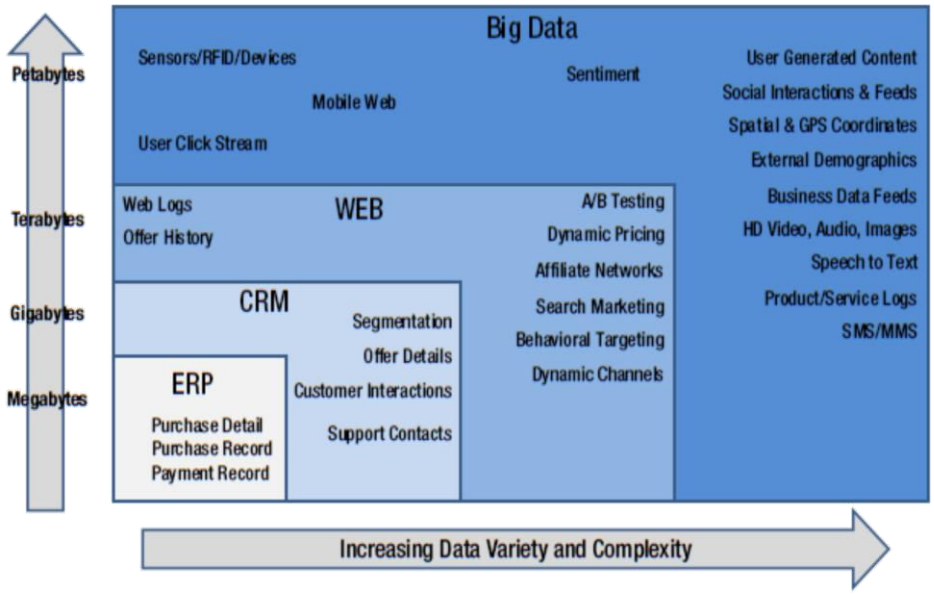
L'évolution à laquelle nous assistons actuellement est d'une autre ampleur.

(*) fiche réalisée en collaboration avec la société « Altran Research »

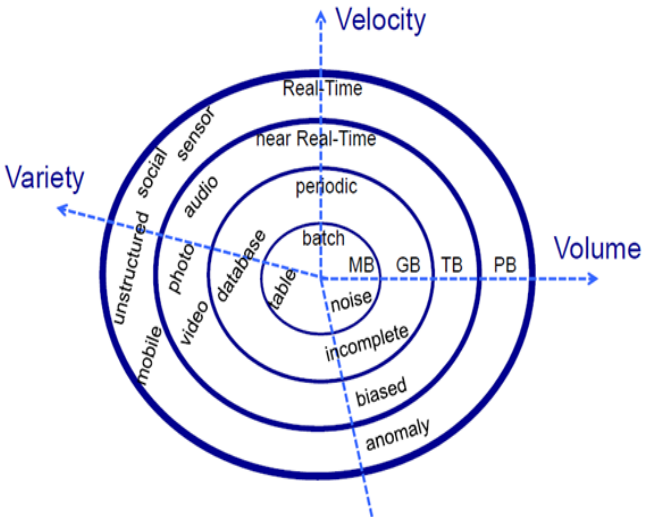
Grâce à la **dématérialisation des échanges, à une digitalisation du monde** et, plus généralement, aux progrès accomplis dans les Nouvelles Technologies de l'Information et de la Communication (capacité, rapidité, convivialité), la quantité d'informations disponibles immédiatement, ou accessibles plus ou moins facilement a augmenté dans des proportions considérables, et non envisagées il y a seulement 15 ans.



Notre monde crée aujourd'hui en 2 jours l'équivalent de ce qui a été accumulé entre l'aube de l'humanité et l'année 2003.



Source : collaboration Hertonworks / Teradata Inc



Source : //whatis.techtarget.com

Au cours de la période récente, la nature des données recueillies a considérablement évolué. En ajoutant aux données traditionnelles structurées des informations plus complexes, mais surtout **non structurées** telles que **des images ou des messages** écrits ou vocaux, ce changement a énormément enrichi l'information disponible mais, dans le même temps, en a rendu **l'analyse beaucoup plus ardue.**

Au cours de cette même période volume et variété n'ont pas été les seules caractéristiques importantes de l'information à évoluer. **Le temps s'est accéléré et la rapidité de mise à disposition de cette information permet désormais d'en réaliser l'usage en temps réel** (trading haute fréquence par exemple).

- 450 millions de tweet par jour
- 30 milliards d'étiquettes RFID
- Plusieurs centaines de millions d'objet GPS vendus chaque année
- 200 millions de compteurs connectés en 2014
- 4,6 milliards de téléphone photo
- 3 milliards de personnes sur internet
- Facebook : 25TB de données chaque jour

Dans ce nouveau monde digitalisé et connecté, **toute action, qu'elle soit professionnelle ou privée, personnelle ou collective, raisonnée ou automatique, matérielle ou intellectuelle, dans le monde de l'industrie ou celui des services, laisse des traces numériques indélébiles**.
Ces « **empreintes** » ne sont rien individuellement, mais **collectivement** elles peuvent être considérées comme une projection dans le monde digital de la réalité du monde et de nos comportements.

Source : trivadis

Le domaine du commerce est certainement un des premiers domaines à l'avoir compris et exploité intensément grâce à la possibilité offerte d'observer à 360° les consommateurs, statistiquement et individuellement.

Comme pour un plan (projection d'un objet tridimensionnel sur une feuille de papier) ou pour des empreintes digitales, l'information ne suffit pas, il faut pouvoir l'interpréter. Pour un plan, c'est la formation et le capacité individuelles à imaginer la troisième dimension qui permettent la reconstitution d'une information utile à partir de celles disséminées sur le plan. Pour des empreintes digitales l'homme, initialement utilisé, a été progressivement remplacé, ou fortement épaulé, par des outils de recherche et de reconnaissance (des algorithmes) seuls capables d'explorer la grande quantité d'empreintes désormais disponibles. Par extension de ces principes,

on aboutit à l'idée que l'on peut se construire **« une idée du monde réel »** à partir de sa projection dans le monde numérique **si l'on dispose des bons outils de reconstruction**.

Certains vont même beaucoup plus loin. C'est ainsi que dans un article-manifeste, Chris Anderson va même jusqu'à proclamer que l'abondance des données peut remplacer la méthode scientifique. Pourquoi chercher à comprendre le réel à coup d'hypothèses forcément hasardeuses puisque, désormais, **le réel peut nous être livré dans sa totalité par les data** ? Selon lui, l'abondance de données nous permet d'accéder à une connaissance complète du réel sans avoir à nous embarrasser de la science.

Sans aller jusqu'à cette vision, il est certain que les informations contenues dans le « Big Data » peuvent informer sur de nombreux sujets ou répondre à de nombreuses questions **pour qui sait les « faire parler »**. Et d'autres évolutions, au-delà de la masse des données, ont rendu possible l'avènement **d'un « Big Data utilisable »,** non réservé aux grosses structures :

- L'**arrivée des grandes mémoires centrales** dans les processeurs de traitement de données pour des prix très abordables permet d'y effectuer directement les traitements analytiques **(« in-memory computing »)**, sans passer par des allers/retours avec un disque, ce qui génère un gain jusqu'à 100 000 dans la vitesse de ces traitements.
- Les **nouvelles architectures de serveurs** banalisés bon marché (technologie dite « Hadoop ») permettent le stockage d'un nombre beaucoup plus important de données, voire sans limites, avec une garantie accrue de résilience ; il faut également y associer la capacité à décentraliser un certain niveau d'analyses des données directement sur les serveurs de stockage (technologie dite « Map/Reduce »))
- L'utilisation de **nouveaux types d'organisation du stockage des données**, au-delà de la très dominante base de données relationnelle, permet de prendre en compte la diversité de représentation des données : colonnes, clé-valeur, document ….
- La nouvelle dynamique des **développeurs de logiciels libres** qui sont apparus dans ce domaine, celui-ci ayant été accaparé jusqu'à présent par des éditeurs classiques de logiciels payants, permet l'irruption de nouvelles stratégies de traitement et de visualisation des données en s'appuyant sur les divers éléments décrits ci-dessus.
- On dispose désormais d'une **panoplie très large d'outils d'analyse** qui ont été développés et progressivement améliorés, tels que « Apprentissage automatique (Machine learning) », « Intelligence artificielle », « Exploration des données (Data mining) ».

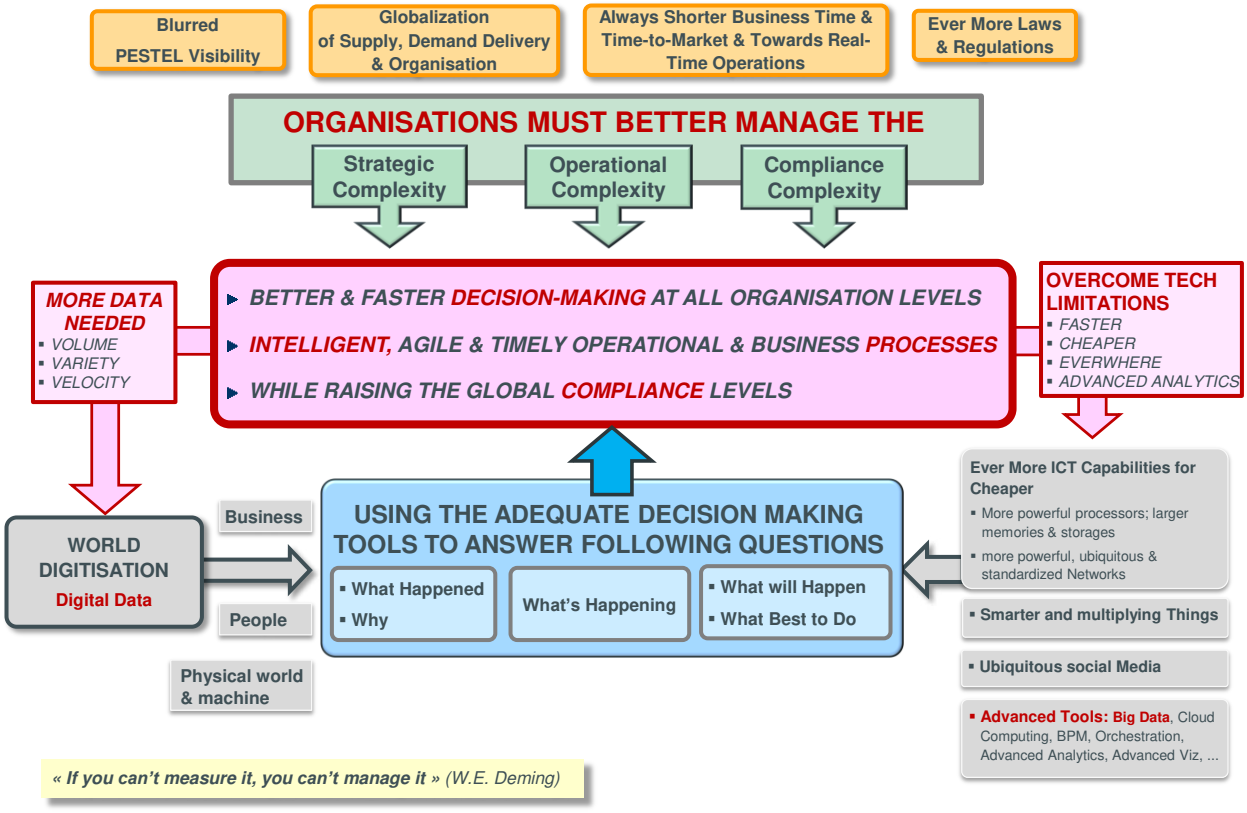
Les usages du « Big Data », ce qu'ils peuvent apporter :

Pour une personne ou une organisation, la richesse et l'intérêt du « Big Data » résident donc dans l'association de la masse énorme des données disponibles avec des outils de recherche et d'analyse de plus en plus performants et adaptés à ses propres objectifs.

Dans les entreprises l'analyse des données a généralement débuté dans le secteur de la comptabilité et du contrôle de gestion. Les utilisateurs de tous les secteurs de l'entreprise se sont progressivement accaparé les techniques de traitement de leurs données pour mieux gérer leurs divers besoins, que ce soit dans la gestion globale de l'entreprise ou de ses diverses subdivisions comme la comptabilité, la finance, le marketing, la vente, la conception (retour d'expérience), les chaînes d'approvisionnement et de production, etc.., et de fait le contrôle de toutes les activités.

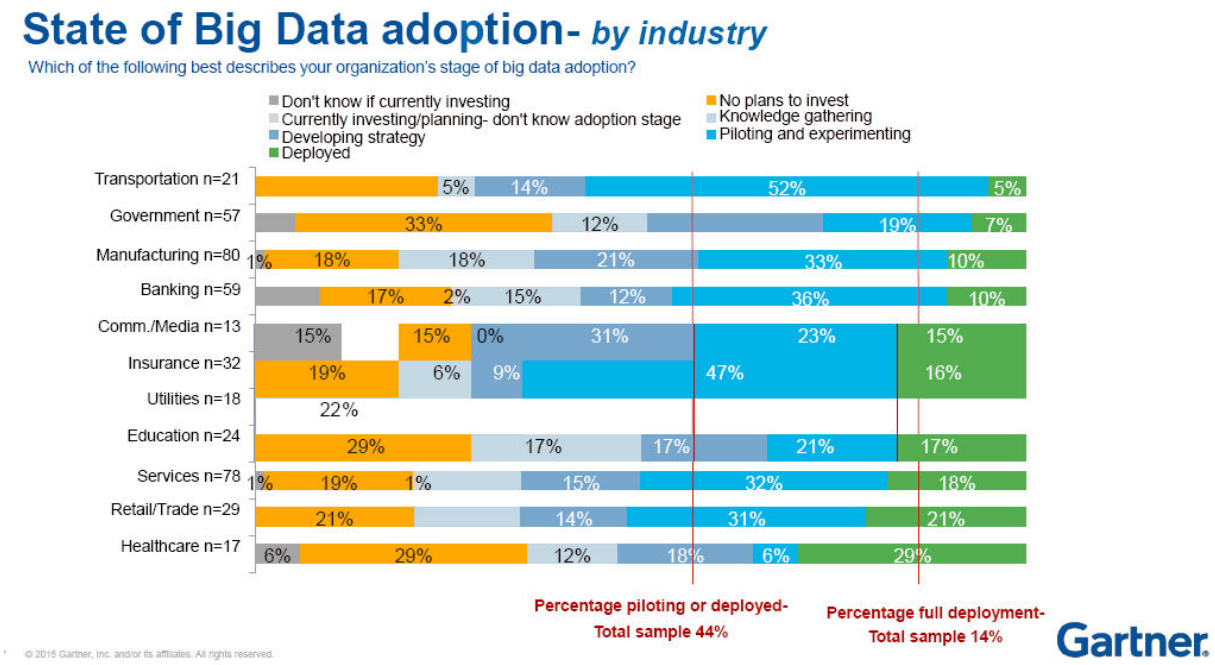
En permettant un élargissement des bases de ces analyses, le **Big Data est bien une réelle percée dans le domaine de l'aide à la décision** sous toutes ses formes. Cette percée devrait permettre à (presque) **chaque employé de l'entreprise de trouver et d'utiliser les outils décisionnels nécessaires à son niveau**, une forme de démocratisation de l'aide à la décision s'appuyant sur des outils puissants, dépassant les limites bien connues du logiciel décisionnel le plus utilisé, Excel, et ceci dans les trois domaines majeurs de création de valeur que sont :

1. **La maîtrise stratégique des enjeux de l'entreprise** par les capacités à prendre des décisions plus rapidement, mieux argumentées et avec un bien meilleur degré d'anticipation et ceci à tout niveau de l'entreprise,
2. **L'amélioration de la performance opérationnelle** en s'appuyant sur une connaissance plus étayée, de plus en plus en temps réel, des paramètres opérationnels des processus (qualité, événements perturbateurs, …), de l'état des machines (développement de la maintenance prédictive au-delà de la maintenance préventive), par l'automatisation de plus de tâches, par une plus grande capacité d'adaptation des processus,
3. **L'amélioration de la maîtrise des risques** de l'entreprise par un suivi constant et précis de tous les risques y compris la détection de la fraude et la lutte contre les cyber-attaques grâce à une capacité d'observation démultipliée.



Les usages du « Big Data », les difficultés rencontrées :

La simple lecture des paragraphes précédents devrait entraîner une adhésion massive des entreprises de toutes tailles à la culture et à l'usage du « **Big Data** », mais le constat de son usage montre une progression lente. De fait, même si des réticences existent, ce sont plutôt **les difficultés rencontrées pour réussir un projet « Big Data »** qui expliquent la situation présentée ci-dessous.



Plus de données ne signifie pas toujours de meilleures données aboutissant à de meilleures décisions.

- **Les données externes** sont, en général, mal ou pas structurées, d'une qualité non établie et moins fiable que celle des données structurées internes. Elles sont souvent difficiles à épurer, sans garantie statistique et ne respectant pas le référentiel de l'entreprise.
- **Leur provenance** n'est pas établie avec certitude, réduisant d'autant la confiance que l'on tire d'une connaissance approfondie des processus de recueil et de traitement initial des données, par exemple traçabilité des étalonnages, type de capteur utilisé, protocole de mesure ….
- **Plus de données** améliorent les corrélations sans forcément faire apparaître les causalités.
- **L'hétérogénéité des données**, associée à une méconnaissance des intentions des sources de ces données peut entraîner des raisonnements faussés et des connaissances inutilisables /
- **Le futur ne s'identifie que rarement au prolongement du passé.**

L'approche « Big Data » est encore récente et souvent immature.

- **Moins de 1% des données** « atteignables » a été analysé à ce jour.
- **Confidentialité et propriété des données** sont des sujets encore essentiellement en friche.
- **Les nouvelles technologies d'encodage et les outils d'usage associés** sont encore balbutiants.
- **Un gros déficit de savoir-faire** subsiste dans le domaine du développement des outils et de la connaissance des données.
- **Des interfaces « userfriendly » sont encore inexistantes.**

L'expertise nécessaire n'existe encore que dans la petite communauté des « data scientists », qui seuls comprennent le contexte des données et disposent des outils pouvant en tirer un savoir utilisable.

Les usages du « Big Data », les précautions à prendre :

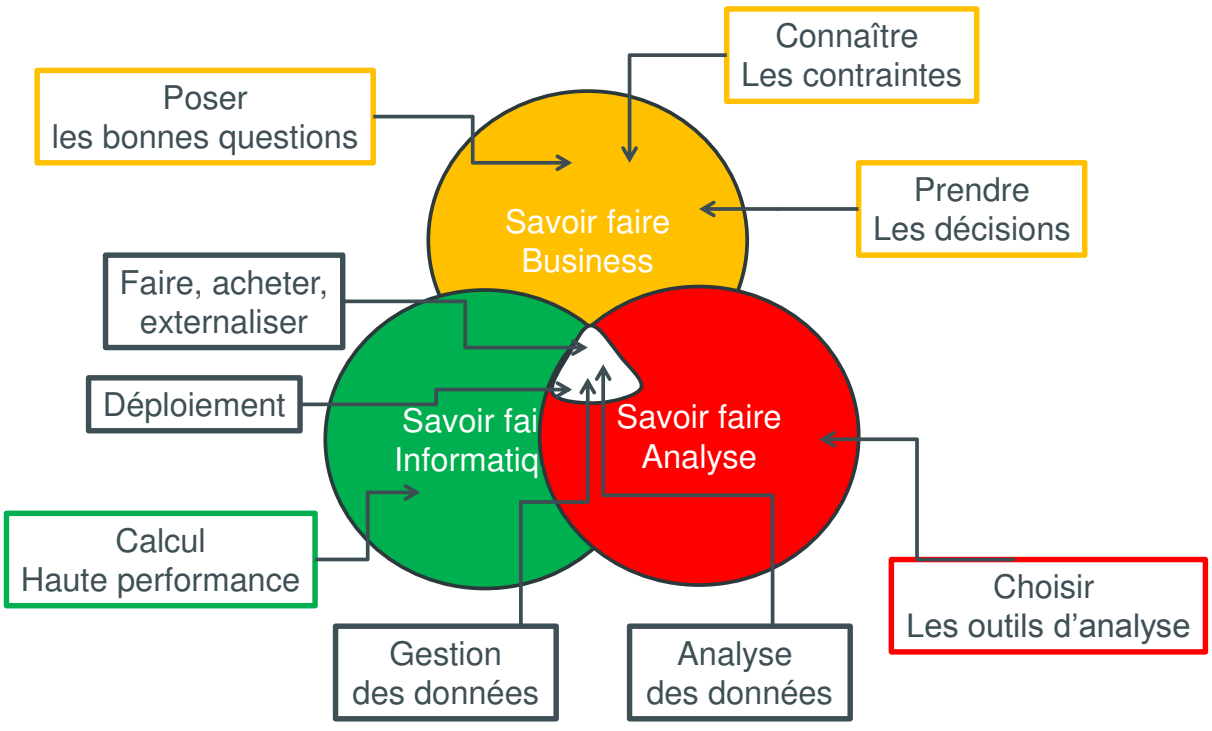
D'abord y gouter

Comme pour tout nouveau développement, il est judicieux d'aborder le thème du « Big Data » **par la mise en place d'une première application** qui devra se dérouler de façon structurée, par étapes, et bénéficier de l'apport d'expérience de ceux qui ont déjà entamé ce parcours.

1. **Comprendre** l'apport de ces technologies pour l'atteinte des objectifs stratégiques et **choisir un premier projet**,
2. **Identifier un véritable besoin business**, exprimer de façon claire une question à laquelle il faut répondre,
3. **Identifier les données nécessaires** à la résolution du besoin (explorer en priorité les données disponibles dans l'environnement immédiat de l'entreprise),
4. **Prétraiter ces données** afin d'en augmenter leurs niveaux de qualité et de confiance,
5. **Choisir les approches et outils** de traitement et d'analyse des données,
6. **Analyser ces Informations** pour les transformer en Connaissances utilisables dans le cadre du besoin exprimé,
7. **Evaluer** les apports et difficultés du projet.

La réussite d'un projet « Big Data » impose une **bonne maîtrise de compétences dans les 3 domaines essentiels** que sont :

- Le domaine thématique du projet
- La capacité d'analyse des données
- L'expertise informatique



Pour un premier projet ces conditions peuvent être remplies dans le cadre d'une **organisation de type projet solidement armée dans les domaines de compétences nécessaires identifiées ci-dessus**.

Avant de généraliser

La réalisation d'un premier projet doit permettre de faire l'état des lieux sur les capacités disponibles dans les différents domaines critiques et d'engager, si nécessaire, les renforcements indispensables avant d'**Installer de façon stratégique l'approche Big Data au niveau global de l'entreprise.**

Ce bilan et ces renforcements sont des éléments indispensables avant d'engager des projets de plus grande ampleur mais **d'autres écueils** peuvent surgir, dont il faut avoir conscience afin d'en prévenir les risques. Sans être exhaustif on peut citer :

- La **culture et l'organisation des entreprises**, en particulier dans les grandes, permettent difficilement l'introduction d'une démarche big data. La valeur des données n'y est pas souvent reconnue et quand elle l'est, il y a une résistance à les partager (« l'information c'est le pouvoir »), alors que le partage est une des clefs d'un traitement big data.
- Le **mode de raisonnement**, qui vise à décrire le réel et prévoir l'avenir avec des modèles statistiques des données construits par la machine, plutôt qu'en appliquant une méthode et un raisonnement logique et scientifique est culturellement difficile à accepter dans des entreprises traditionnellement rationalistes.
- Les **limitations réglementaires ou juridiques** sur l'utilisation des données personnelles (CNIL en France). Le droit en la matière est très évolutif en France comme à l'étranger.
- La **volatilité des ressources humaines** dans les compétences clefs que sont la connaissance des sources et celle des moyens d'analyse de ces masses de données. Il y a en France un véritable problème de nombre de data scientists et une grande difficulté à les embaucher puis à les conserver (pour des raisons financières mais aussi culturelles).
- La **disponibilité des moyens de calcul**, car pour traiter les quantités de données évoquées on ne peut se contenter d'un simple PC il faut des grosses capacités d'archivage, des grosses machines avec beaucoup de mémoire centrale

Enfin il faut insister sur la nécessaire **transformation de l'approche managériale**, nécessaire en préalable au lancement d'une démarche big data si on veut avoir une chance raisonnable de succès.

Quand le Big Data impose de nouveau outils …
et débouche sur de nouveaux usages
Heuritech et l'analyse automatique d'image

Fondée en 2013 par deux docteurs en Intelligence Artificielle, **Heuritech est une startup** experte dans l'analyse et l'extraction automatique de nombreux types d'informations à partir de quantités massives d'images.

Utilisant les technologies de pointe du **Machine Learning et du Deep Learning**, nous avons développé un **moteur d'Intelligence Artificielle dédié à l'analyse des images** et dont le fonctionnement est similaire à celui du cerveau humain : détectant des éléments basiques dans toute image comme des niveaux de couleur, des formes simples, puis les associant progressivement en formes de plus en plus complexes et abstraites, notre moteur identifie tous types d'objets, de couleurs, de textures, de personnages, de contextes… et analyse à l'heure actuelle plus de 100 millions de pages web par jour.

Cette performance est basée sur un **processus d'apprentissage** développé en étroite collaboration entre notre équipe de 17 experts techniques dont 8 docteurs en intelligence artificielle et nos partenaires leaders dans leur secteur (dont la maison Louis Vuitton, n°1 mondial du luxe), pour assurer une combinaison optimale de diversité, de précision et de robustesse des caractéristiques détectables sur tous types d'images, notamment celles prises dans les contexte « naturels » les plus variés.

Le **cas d'usage principal** de notre technologie consiste en la reconnaissance automatique des formes et des caractéristiques (couleur, matière, motif, style…) des vêtements et des accessoires de mode sur les images des sites de vente en ligne (e-tailers), et des réseaux sociaux.

L'enjeu pour nos clients, plateformes commerciales et marques de mode, est celui de l'indexation de leurs catalogues produits : ce processus est crucial pour assurer le bon référencement des produits, et faciliter leur recherche par les internautes. Manuel, lent et laborieux, ce processus est inadapté au regard de la taille croissante des collections et de leur vitesse de renouvellement et, surtout, il ne permet pas d'assurer le bon référencement des produits lors des recherches des internautes, avec un impact direct sur le taux de conversion des e-tailers.

Notre solution permet de **multiplier le nombre de descriptifs** attachés à chaque produit, pour fournir aux consommateurs des options de recherche et de filtres plus précis et plus détaillés, ce qui facilite et optimise l'expérience d'achat, et augmente les taux de conversion.

De plus, notre technologie est particulièrement adaptée aux nouveaux usages mobiles des consommateurs : à **l'instar d'un Shazam de l'image**, notre moteur propose une fonction de recherche visuelle qui détecte et caractérise les vêtements présents sur toute photo prise par un internaute, et les compare au catalogue du e-tailer pour trouver la référence exacte recherchée, ou les produits visuellement les plus similaires.

Un **autre domaine d'application de notre technologie concerne l'agriculture** : en collaboration avec les experts botanistes de notre partenaire, nous avons entraîné notre moteur à reconnaître l'espèce et le stade de croissance des mauvaises herbes les plus dangereuses pour les cultures, directement à partir des photos prise en temps réel sur le terrain par les agriculteurs, pour aider au diagnostic optimal des traitements herbicides à épandre.

Ainsi, notre technologie **accompagne aussi la révolution de la smart agriculture**, en assurant l'efficacité maximale des traitements tout en minimisant leur impact écologique.

Trimestriel de l'intelligence technologique
Directeur de la rédaction : Patrick Ledermann
Auteurs : Michel Laroche, Jean-Luc Strauss