Les Big Data

Introduction

Le texte qui suit est la translation d'un exposé fait à la FAL le 16 novembre 2017, celui-ci a été possible par la publication de : *Les Big Data à découvert, Mokrane Bouzeghoub et Rémy Mosseri, CNRS édition*.

La science des données est basée sur 3 V :

- Volume des données, exprimé en octet, par exemple la NSA peut stocker 1000 zetta octets (10^{24}),
- variété des données : nombre, texte, image, son,
- vitesse d'acquisition et de transmission.

Un bref historique

Les documents les plus anciens sont des peintures rupestres retrouvées dans des grottes en Espagne et ont été réalisées avant l'arrivé d'*homo sapiens* en Europe. Elles ont donc été réalisées par *homo néandertalien*. D'autres peintures rupestres, plus récentes, sont l'œuvre d'*Homo sapiens*, comme celles de la grotte Chauvet (30 000 ans BP). Un très grand saut dans le temps, c'est en Mésopotamie que naît les premières formes d'écriture, il y a environ 5 500 ans. Celles-ci sont réalisées sur des tablettes en argile, cuites ou seulement séchées et donc recyclables. L'étape suivante est l'utilisation du papyrus, en Égypte. Puis c'est l'invention du papier.

Une étape importante sera franchise lors de l'invention de l'imprimerie par Gutenberg en 1455. Ce qui permet la reproduction d'un texte en plusieurs exemplaires et non seulement la copie en un seul et nouveau exemplaire. Puis ce sera la machine à écrire, dont le premier brevet a été déposé en 1714 par Henry Mill, l'idée des touches sera breveté en 1833 par Xavier Prongan et la première véritable machine à écrire sera réalisée en 1870 et commercialisée par Remington. La première machine à écrire permettant de visualiser et de corriger d'une à trois lignes a été commercialisée par Canon en 1986, l'ancêtre des traitements de texte.

En parallèle au texte, les images ont suivie un itinéraire semblable, à la suite des peintures rupestres de très grands progrès ont été réalisés dans les colorants, extraits de minéraux et de plantes puis de synthèse, mais la vrai rupture technologique est l'invention de la photo par Joseph Niépce et Louis-Jacques Daguerre en 1829, par l'utilisation des propriétés photosensibles du chlorure d'argent. L'étape suivante sera celle du cinéma par les frères Lumière en 1895.

Les premiers enregistrements sonores auront lieu en 1860.

Ces trois modes de communication utiliseront bientôt un même support de stockage grâce à la numérisation. La boîte à musique a été inventée en 1796 et la carte perforée, déjà utilisée dans les métiers à tisser Jacquart, sera proposée en 1834 par Charles Babbage pour effectuer des calculs. Il est considéré comme le précurseur de l'informatique.

Les prés requis

La numérisation.

La numérisation ou codage a nécessité le développement de norme comme le code ASCII pour la typographie. L'*American Standard Code for Information Interchange* (Code américain normalisé pour l'échange d'information), connu sous l'acronyme ASCII, est une norme informatique de codage de caractères apparue dans les années 60. C'est la norme de codage de caractères la plus influente à ce jour. Le code ASCII définit 128 caractères numérotés de 0 à 127 et codés en binaire de 0000000 à 1111111. Sept bits suffisent donc. Toutefois, les ordinateurs travaillant presque tous sur un multiple de huit bits (un octet), chaque caractère d'un texte en ASCII est souvent stocké dans un octet dont le 8e bit est 0.

Pour les images et le son, il y a une première transformation en différence de potentiel, celle-ci est ensuite numérisée.

Acquisition des données.

L'acquisition manuelle des données laisse de plus en plus la place à la capture automatisée par le développement des capteurs et des neurones artificiels ayant une capacité d'auto-apprentissage. Il faut noter des progrès dans le passage du continu au discret et aussi dans les problèmes d'échantillonnage.

Le traitement des données.

Les traitements de données reposent sur des modèles théoriques et utilisent des algorithmes. Ce terme provient du nom du mathématicien persan Al-Khwarizni (IX^{ème} siècle). Un algorithme est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre un problème ou d'obtenir un résultat en particulier par des opérations de tri de données (insertion, fusion, éclatement). Le traitement des données profite du développement des ordinateurs. La capacité de calcul des microprocesseurs augmente d'un facteur 2 tous les 2 ans mais les données le sont d'un facteur 3.

La transmission des données.

La transmission des données entre les différents utilisateurs est l'un des frein, en particulier par la consommation d'énergie. Elle est proportionnelle au carré de la distance pour la transmission hertzienne, comme pour la téléphonie mobile. Elle dépend linéairement de la distance lors de transfert filaire (cuivre), c'est le cas de l'ADSL, la déperdition se fait principalement par effet Joule. Elle très peu dépendante de la distance dans le cas des fibres optiques.

Le stockage des données.

Le stockage des données est problème important, en particulier pour leur conservation dans la durée. Mémoires SRAM, DRAM, CD, DVD, mémoire flash.

Des exemples de big datas.

Du marchand de glaces au supermarché.

Comment un marchand de glace peut-il optimiser ses profits lors d'un long week-end. Il doit tenir compte de ses résultats précédents, avoir une idée précise des différentes festivités dans sa région mais aussi obtenir des prévisions météorologiques fiables car sa marge de profits dépend de la vente et de l'absence de stocks invendus à la fin du week-end. Il peut éventuellement prendre une assurance météo.

Les tickets de caisse des supermarchés non pas un grand intérêt pour les clients mais l'ensemble de ceux-ci l'est pour les gérants. Ils peuvent ainsi analyser les habitudes de consommation des clients en fonction par exemple de la météo pour rester dans le même esprit. Par exemple, ils peuvent constater que lors de week-end ensoleillé les clients qui achètent de la viande à griller achètent aussi du charbon de bois et du vin rosée. Pour augmenter le chiffre d'affaire ils ont la possibilité de faire un prix d'appel sur l'un des produits et d'augmenter un peu les autres et aussi de disperser les articles dans le magasin pour susciter des achats supplémentaires.

Les requêtes en hypertexte.

Étymologiquement, le préfixe « hyper » suivi de la base « texte » renvoie au dépassement des contraintes de la linéarité du texte écrit. Un hypertexte est un document ou un ensemble de documents contenant des unités d'information liées entre elles par des hyperliens. Ce système permet à l'utilisateur d'aller directement à l'unité qui l'intéresse, à son gré, d'une façon non linéaire. Un document hypertexte est caractérisé par : les nœuds, les liens et les ancres. Les nœuds sont les unités de base par exemple un paragraphe. Les liens permettent de naviguer entre les nœuds de manière non linéaire. Les ancres sont les cibles des liens dans un document. Les recherche sur le

web sont basées sur les liens hypertextes. L'annotation des documents est réalisée premièrement par les déposants et ensuite par les différents utilisateurs. L'ordonnancement des documents est de fait réalisé par les différents utilisateurs. Si un document est choisi par plusieurs utilisateurs son ordre d'apparition sera amélioré. Les liens sont enrichis par les précédentes recherches.

Wikipédia, une encyclopédie en commun.

La première forme d'encyclopédie est l'œuvre d'Isidore de Séville (~560 - 636) dénommée *Étymologies* (*Etymologiæ*). Ce texte est constituée de vingt livres, qui propose une analyse étymologique des mots divisée en 448 chapitres. Par cette œuvre, il essaie de rendre compte de l'ensemble du savoir antique et de transmettre à ses lecteurs une culture classique en voie de disparition. Son livre a une immense renommée et connaît plus de dix éditions entre 1470 et 1530. Il contribue à la survivance durant le Moyen Age de nombreuses œuvres antiques par sa technique de citation. C'est l'organisation particulière de ce livre, indexée par première, puis deuxième lettre (début d'une classification arborescente par lettres) qui lui vaudra d'être au XXe siècle nommé par le Vatican saint patron des informaticiens.

La plus célèbre des encyclopédies est celle de Diderot et d'Alembert. En 1745, l'Académie des sciences propose à d'Alembert de traduire de l'anglais en français le *Cyclopaedia* d'Ephraim Chambers. Le 16 octobre 1747, ils sont désignés à la tête d'un projet de rédaction d'une encyclopédie originale. Le projet se transforme en la rédaction d'une œuvre originale et unique en son genre, l'*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, avec un désir de synthèse et de vulgarisation des connaissances de l'époque. Le premier volume paraît en 1751 et le projet s'achève en juillet 1765. Diderot gardera cette charge jusqu'à son achèvement et verra l'*Encyclopédie* achevée.

Un projet d'une autre nature est apparue à la fin du 20ème siècle, c'est Wikipédia. L'idée des wiki date de 1994. Un wiki est une application web qui permet la création, la modification et l'illustration collaboratives de pages à l'intérieur d"un site web. C'est un outil de gestion de contenu dont la structure de départ est minimale et qui évolue en fonction des besoins des utilisateurs. Le plus célèbre des wiki est Wikipédia. C'est un projet d'encyclopédie universelle crée par Jimmy Wales et Larry Sanger le 15 janvier 2001 sous le nom de domaine wikipedia.org. L'encyclopédie est hébergée sur internet grâce aux serveurs financés par la Fondation Wikimedia. L'encyclopédie est en libre accès, en lecture comme en écriture, c'est-à-dire que n'importe qui peut, en accédant au site, modifier la quasi-totalité des articles.

La médecine personnalisée.

L'idée de médecine personnalisée est apparue avec le développement des techniques de

séquençage des génomes humains. La technique de séquençage de l'ADN, actuellement utilisée, est due à Frederick Sanger en 1970 et la première séquence obtenue est celle du bactériophage ΦX174. Cette technique a connue par la suite un développement exponentiel. En 2003, est publié la séquence d'un génome humain par un consortium international après un travail de plus de 3 ans. En 2011, le coût du séquençage d'un génome de 3.10^9 paires de bases est d'environ 3 milliard € ; en 2017 le même résultat est obtenu pour 1 000 € et en une semaine. En parallèle a ce développement, l'analyse de l'expression des gènes, aussi bien sous la forme d'ARN que des protéines est devenu aussi possible à haut débit.

Les génomes humains sont homologues à plus de 99,9%. Il serait donc possible de rechercher les différences dans la séquence des gènes et d'établir des corrélations entre ces différences et les diverses pathologies qui affectent les hommes. Pour l'instant cette approche est utilisée en cancérologie pour adapter les traitements plus ciblés, en comparant les séquences de l'ADN tumorale vs normal. Les résultats actuels ne sont pas encore à la hauteur des attentes.

Les études épidémiologiques.

La première tentative de représentation graphique dans une étude épidémiologique est due à John Snow, en 1854. Il a réalisé une carte montrant la corrélation entre un point d'eau et les victimes lors d'une épidémie de choléra à Londres.

Les études épidémiologiques se sont ensuite devenues plus courantes notamment grâce à l'utilisation de tableau mais il est difficile d'introduire beaucoup de données ceux-ci. Cette contrainte a favorisé les enquêtes ciblés, donc avec une pré-idée de résultat. Les études actuelles peuvent être réalisées sur un effectif beaucoup plus important (10^5) et plus de paramètres pouvant inclure des données issues de capteurs. Pour ces études se posent la question de l'anonymisation des données personnelles, car même si celles de l'étude sont bien sécurisées, il est possible de contourner la sécurité par l'utilisation des données personnelles publiées sur les réseaux sociaux comme Facebook ou Twitter. Ces mêmes réseaux sociaux peuvent aussi être utilisés pour révéler une épidémie, comme la grippe, plus rapidement que les systèmes de veilles sanitaires.

Autres exemples.

En astronomie le recueil automatisé des données par les caméras CCD donne un flux très important de l'ordre de 4,3 Moctets/s. Ces masses de données ne peuvent être traitées que par des programmes informatisés. Il en est de même pour celles concernant le climat, comme par exemple les relevés du niveau des océans et celles des températures.

Même l'agriculture, industrielle, est dépendante des BD par une cartographie fine des

parcelles pour un apport optimisé des engrais et autres intrants.

En sociologie, deux exemples :

- les études sur la mobilité et les déplacements peuvent se faire très facilement à l'aide du suivi des téléphones portables ;
- les sondages d'opinion ? Est-il préférable d'augmenter les effectifs pour s'approcher de la population ou de définir précisément un échantillon représentatif de la population.

Les réseaux sociaux.

Le modèle économique des réseaux sociaux et autre GAFA (Google, Apple, Facebook, Amazon) est de proposer des services gratuits en échange de données personnelles revendues à des publicitaires. L'idée d'utiliser de la publicité, comme ressource financière n'est pas nouvelle. Dès 1920, RCA (une radio américaine) a eu recours à ce type de financement. La presse aussi trouve une partie de ses ressources dans la publicité. En France, l'éclosion des radios en bande FM après leur libération, en 1981, est aussi basée sur un financement exclusivement lié à la publicité. Il en est de même des chaînes de télévision de la TNT et des offres en ligne. Radio, presse et télévision proposent de la publicité générale de type top/down.

La grande différence pour Google et Facebook, c'est que la publicité est ciblée ; à partir des données personnelles laissées sur les réseaux, les algorithmes de ces sociétés revendent à des annonceurs publicitaires, les adresses de futurs clients. Nous avons tous remarquer qu'après une recherche avec un mot clé comme tente de camping, de recevoir des offres d'Inter-sport et/ou Décathlon. Cette publicité ciblée est sans doute plus efficace qu'une généraliste. Mais les données ainsi relevées et stockées peuvent servir à beaucoup d'autres choses que de la simple publicité comme la revente à des compagnies d'assurances ou même d'influencer le choix des électeurs comme dans le cas de Cambridge Analatyca.

L'origine étasunienne de ces réseaux n'est pas une surprise, elle remonte à la prise de contrôle des câbles sous-marins, des réseaux hertziens et le développement de compagnies mondiales comme ITT et IBM. Plus récemment, ce même pays contrôle aussi le mécanisme d''attribution des adresses IP.

Finalement la richesse de ces sociétés est le résultat d'une imposition très faible et des mécanismes mis en place pour échapper à l'impôt grâce à la complicité de certains états comme l'Irlande, le Luxembourg et les Pays-bas.

Conclusion.

La plus grande des révolutions des Big Data est la numérisation de l'ensemble des donnée et de l'augmentation des capacités de calcul et de stockage de ces données. Si les États-Unis dominent actuellement ce secteur c'est principalement à cause d'une législation plus *libérale*, comme dans le cas du séquençage des génomes, tant qu'elle est favorable aux affaires.