# LE BIG DATA SIGNIFIE-T-IL LE " BIG BANG " DES ETUDES MARKETING ?

Philippe Jourdan, Jean-Claude Pacitto

# LE BIG DATA SIGNIFIE-T-IL LE « BIG BANG » DES ETUDES MARKETING ?

Philippe **Jourdan**
Professeur des Universités- IAE Gustave Eiffel, UPEC
CEO, Promise Consulting
Mail : philippe.jourdan5@gmail.com
http://whatsnewinmarketing.blogspirit.com

Jean-Claude **Pacitto**
Maître de Conférences – IUT Créteil, UPEC
Membre du Comité Scientifique, Promise Consulting
Mail : jean-claude.pacitto@orange.fr

*Philippe Jourdan et Jean-Claude Pacitto sont également rédacteurs en chef de la Revue Française du Marketing (RFM), la revue marketing de l'ADETEM.*

Il n'est pas une émission consacrée au marketing, une publication, un ouvrage, un séminaire ou bien encore une rencontre interprofessionnelle qui n'aborde le sujet : le Big Data signifie-t-il la fin des études marketing ? Les professionnels des études de marché se complaisent bien souvent dans une attitude frileuse, promptes à annoncer leur propre déclin, craignant qu'une révolution ne les emporte et le régime « honni » qu'ils pensent parfois incarner tant au sein des organisations marchandes que de la société civile. Or il est peut-être temps de rétablir certaines vérités : face au constat que le comportement des consommateurs est devenu difficilement prévisible, en raison de leur attitude changeante, errante et de leur grande mobilité, le Big Data n'est au fond qu'un outil supplémentaire, qui en aucun cas ne peut prétendre à lui seul remplacer une intelligence humaine indispensable. Or cette « intelligence » humaine, dans le sens américain du terme « renseignement » humain, nous pensons que les marketeurs et les hommes d'études sont les mieux placés pour l'assumer.

## 1- Le Big Data : définition et enjeux

Pour bien saisir l'enjeu du Big Data, encore convient-il de rappeler quelques définitions d'un terme trop souvent galvaudé. Le Big Data que l'on peut traduire par « données massives » en français (bien que le terme soit peu élégant) désigne des corpus de données qui deviennent à ce point volumineux que les méthodes de stockage dans des bases de données, les processus d'extraction et de codage ainsi que les algorithmes de traitement des données habituellement utilisés deviennent inopérants pour en extraire le sens. Les chiffres communiqués sur la croissance exponentielle des données accumulées sont saisissants : selon le décompte d'IBM, le monde génère quotidiennement 2,5 trillions d'octets de données, à tel point que « *90% des données dans le monde ont été créées au cours des deux dernières années seulement*[1] ».

### 1.1. Origine et sources du Big Data

Au-delà même des volumes, ce qui caractérise aujourd'hui la révolution du Big Data, c'est la grande diversité des origines et des formats des données disponibles : capteurs météorologiques, objets connectés, messages sur les réseaux sociaux, courriels, images numériques et vidéos postées depuis son téléphone mobile, enregistrements de transactions d'achats, signaux GPS permettant une géolocalisation et un enregistrement des déplacements de l'individu, données d'études, renseignements nominatifs issus de formulaires ou données de navigation sur des sites, etc., ce ne sont là que quelques-unes des sources qui alimentent quotidiennement des serveurs de données dans le monde entier. Au-delà de l'hétérogénéité des données quant à leur nature et à leur format, ce qui caractérise le Big Data – et en constitue une difficulté

[1] « *Le Big Data à l'écoute de votre business* ».- http://ibm.co/1jq0XQC

supplémentaire – c'est également la localisation de ces mêmes données sur des terminaux variés (PC, smartphones, tablettes, objets communicants, etc.) dont le nombre ne manquera pas de s'accroître dans un futur proche, en particulier en raison de l'engouement pour les terminaux mobiles (smartphones) et les objets connectés. Selon une infographie publiée sur le site des Echos[2], l'Internet des objets est une mutation profonde dans le fonctionnement même des entreprises et s'adresse à une grande diversité de secteurs : la maison avec la domotique, le transport avec la géolocalisation, la santé avec la multiplication des capteurs individuels, l'industrie et la distribution avec la systématisation des puces RFID[3] et enfin l'aménagement du territoire avec le concept de villes intelligentes. Certes, toutes ces données n'ont pas toutes un intérêt dans le cadre du champ qui nous préoccupe, celui de l'étude du comportement du consommateur, mais force est de reconnaître que les études marketing sont les premières concernées par cette avalanche de données individuelles le plus souvent au départ décentralisées et non structurées.

### 1.2. **Les enjeux posés par la loi des 3V**

Le Big Data – terme dont l'origine remonte aux premiers articles parus sur le sujet en Octobre 1997 aux Etats-Unis - est donc attaché à la croissance exponentielle des données produites et stockées et se situe à la rencontre de deux phénomènes : une accélération de la production de données de nature diverse dans des secteurs très variés et en parallèle une diminution de leur coût de stockage. Pour autant, les données aujourd'hui disponibles suivent une loi des 3V (volume, vélocité et variété) qui pose en termes d'architecture de stockage et de processus de traitement des problèmes très particuliers. Arrêtons-nous quelques instants sur la signification de ces termes (volume, vélocité, variété) pour la première fois employés par le cabinet Gartner en 2001 et auxquels se sont greffés depuis lors les autres qualificatifs de véracité, valeur et visibilité des données.

Le volume de données fait bien sûr référence au premier défi du Big Data : un accroissement sans commune mesure des données produites tant par les entreprises que par les individus. On estime généralement que l'humanité a produit depuis l'origine jusqu'en 2003 un volume de 5 exabytes de données (5 Mds de GB) alors que ce même volume a été produit en deux jours seulement en 2011 et qu'il est vraisemblable qu'il suffit aujourd'hui de 10 minutes pour atteindre cette même quantité de données[4]. Cet accroissement des volumes de données est lié bien sûr aux progrès technologiques (multiplication des objets connectés, généralisation des infrastructures d'échanges de données, etc.) mais il est aussi le fait d'une évolution sociétale majeure vers un plus grand partage des informations entre les objets, entre les applications, entre (et au sein) des organisations et au final entre les individus. Le secteur des télécommunications est un exemple éloquent de l'explosion des données archivées par les opérateurs. Au stockage des données vocales (messagerie) s'ajoutent aujourd'hui avec la généralisation des smartphones la sauvegarde des données numériques échangées sur le réseau Internet, la collecte des données de géolocalisation, la sauvegarde des historiques de recherche et enfin l'archivage des données accumulées par les applications. Dans le volume des données numériques échangées entre particuliers, une place éminente est évidemment occupée par le commerce électronique et les réseaux sociaux.

Ce volume de données s'accompagne d'une grande vitesse de production et d'échange qui forme le deuxième grand défi. La vitesse décrit en réalité la « *fréquence avec laquelle les données sont générées, capturées et partagées*[5] ». Tout se passe en réalité comme si les données échangées entre particuliers ou bien

---

2 Les Echos.fr.- « *Bienvenue dans le monde des objets connectés* ».- octobre 2014.- http://bit.ly/1OE7Skm

3 Une puce RFID (« Radio Frequency Identification ») désigne un circuit électronique d'identification automatique qui utilise le rayonnement radiofréquence pour identifier les objets porteurs d'étiquettes lorsqu'ils passent à côté d'un interrogateur (appelé station de base ou plus généralement lecteur). Les applications sont multiples : la clé pour démarrer un véhicule, le badge pour accéder à un bâtiment ou une salle, le pass des remontées mécaniques, le titre de transport dans les bus ou les métros, le suivi des marchandises ou des produits en cours de montage dans l'industrie, etc. (Source : centre national de référence CNRFID.- http://bit.ly/1ZLqT9S).

4 Source : Databusiness.fr.- « *Big Data : définition, enjeux, étude de cas* ».- Portail francophone sur le Big Data et les Analytics.- http://bit.ly/1hKgEAp.

5 Schmidt Stefan (2012).- *« Les 3V du Big Data : Volume, Vitesse et Variété »*.- Journal du Net.- http://bit.ly/1GLUuUO

entre organisations et particuliers s'inscrivaient dans un temps de plus en plus court, proche du temps réel. Or le temps réel est probablement l'obstacle majeur des systèmes classiques de gestion de la relation client (CRM) car la personnalisation exigée aujourd'hui par le consommateur de l'offre, de la relation, du point de vente, etc. s'inscrit dans un délai proche de l'instantanéité. Lorsque la marque est en mesure de décrypter les attentes du consommateur issues d'une analyse fine des données collectées, il est déjà trop tard car un autre cycle de production de données est déjà en cours.

Le troisième défi, c'est celui de la variété des données produites. Prenons l'exemple d'une enseigne de distribution classique disposant de magasins physiques, d'un site Internet et d'un programme de fidélité, soit la norme des outils disponibles a minima aujourd'hui. Elle dispose potentiellement aujourd'hui des données suivantes sur ses clients : caractéristiques démographiques, données d'achat, données de cumul de points de fidélité, données de localisation sur les magasins fréquentés, données de navigation sur son site Internet, contenus d'e-mails échangés, enregistrement des conversations téléphoniques avec le service client ou le SAV, etc., et nous nous situons ici dans le cas le plus simple ; l'ensemble de ces données peut être doublé si l'enseigne dispose d'un site de e-commerce par exemple. Si l'on se situe du côté de l'individu, la connaissance fine de son profil fait appel à un croisement de sources multiples qui forment autant de « traces » de son parcours numérique : sites Internet, blogs et réseaux sociaux, moteurs de recherche, téléchargements et écoute en streaming de vidéos ou de musique, données échangées depuis son mobile, applications téléchargées, données de géolocalisation, courriers électroniques, formulaires de réclamation, données d'enquête de satisfaction, données enregistrées par les objets connectés possédés ou portés sur soi, etc. Et nous ne prenons ici en compte que les données échangées sur les réseaux, à l'exclusion donc de toutes celles également produites dans le monde « réel ». A la variété des sources s'ajoute une grande diversité des formats, la donnée pouvant être quantitative, textuelle, issue de l'imagerie ou de la vidéo pour ne citer que les formats les plus courants. En théorie, cette grande variété de la donnée (tant dans sa nature que dans ses sources) est une opportunité permettant une micro-segmentation fine du besoin du client[6]. En pratique se pose le redoutable problème de la véracité et de la valeur de l'information puis celui de l'usage des données ainsi collectées.

### 1.3. **Quelle valeur, quelle véracité pour quels usages ?**

Si le Big Data a été initialement défini par les 3V précédemment analysés (volume, vitesse et variété), les experts s'attachant à une description en surface du phénomène, deux autres considérations n'ont pas tardé à émerger et nous semblent du point de vue des enjeux en marketing plus importants encore : le degré de véracité et de valeur de l'information d'une part, puis celui de l'usage attendu. Dès la formation des premiers entrepôts de données (les fameux « data warehouses »), s'est posée la question de la valeur intrinsèque des données, qui ne peut naturellement s'apprécier qu'en fonction de l'objectif de traitement poursuivi (autrement dit du contexte comme le soulignent avec justesse Teboul et Berthier[7]). L'importance de la valeur et de la véracité de l'information a été soulignée dans un rapport paru en 2012 « *The real-world use of Big Data* » produit par IBM et la Saïd Business School de l'université d'Oxford. Ces deux notions sont en réalité au cœur d'un enjeu essentiel : sans valeur, ni véracité, l'information n'a aucune utilité.

La véracité d'une information – c'est-à-dire sa capacité à reproduire ce qui est conforme à la réalité d'un phénomène, d'un comportement, d'une attitude ou d'une opinion sans bruit ni distorsion – est dépendante de multiples facteurs[8]. En premier lieu, la donnée peut être trop ancienne et, par là même, dépassée et donc inopérante pour expliquer les comportements présents et plus encore futurs des consommateurs. Les comportements d'achat ne sont hélas pas toujours reproductibles (et moins encore sur

---

[6] Voir à ce propos le dossier de Christophe Dumoulin (2015).- « *Le Big Data et ses 3V, un vaste océan d'opportunités* ».- blog.business.decision.com.- http://bit.ly/1PzxP5j

[7] Teboul Bruno, Thierry Berthier (2015).- « *Valeur et Véracité de la donnée : Enjeux pour l'entreprise et défis pour le Data Scientist* ».- Acte du Colloque « La donnée n'est pas donnée ».- Ecole Militaire, 23 mars 2015.

[8] Dhélin Hervé (2015).- « *Big Data et marketing, info ou intox ?* ».- Marketing Professionnel.- http://marketing-professionnel.fr, 11 Septembre 2015.- http://bit.ly/1OBVmU1

des durées longues) et les modèles prédictifs fondés sur des données anciennes aboutissent au mieux à ne rien expliquer lorsqu'ils n'induisent pas le chercheur en erreur. En second lieu, la donnée peut être non pertinente pour le phénomène étudié. Certes, des techniques statistiques existent pour éliminer la donnée non pertinente, redondante, mais c'est faire fi des ressources parfois colossales investies pour collecter et stocker une information trop ancienne et (ou) non pertinente. En outre, certaines variables peuvent avoir un lien de causalité directe avec un phénomène, en raison de la présence d'une variable médiatrice. Dans ce cas précis, la véracité de la relation causale entre le prédicteur et la variable dépendante est fonction d'une troisième variable souvent insoupçonnée et, par là même, non mesurée. Ainsi, si on observe que la valeur immobilière d'une résidence principale est fonction du niveau d'éducation du chef de famille, on pourrait conclure à tort que les personnes éduquées accordent plus d'importance au confort de leur habitat. En réalité, le salaire joue ici un rôle médiateur : les personnes les plus éduquées ont des emplois mieux rémunérés et s'achètent des maisons plus grandes qui valent donc plus chères ! La véracité des données est aussi fonction de l'intention ou de la sagacité du répondant. Celui-ci peut bien sûr ne pas vouloir dire toute la vérité, en premier lieu sur son identité comme l'atteste la multiplication des pseudonymes. Il peut également ne pas avoir toute la sagacité et la finesse de jugement nécessaires pour exprimer la vérité. Ainsi, dans le domaine particulier des essais cliniques en myologie, les patients souvent lourdement handicapés sont dans l'incapacité de percevoir une amélioration de leur état musculaire, en raison d'un déficit prononcé depuis de longues années. Cela nécessite la mise au point d'instruments de mesure d'une très grande sensibilité à l'effort. Se pose ici le problème de la fiabilité et de la sensibilité des capteurs aux phénomènes étudiés. Soulignons que la véracité de l'information peut être conditionnée par un niveau d'incertitude qu'il est parfois difficile de réduire. Ainsi, les prévisions météorologiques sont-elles entachées d'un degré de fiabilité plus ou moins élevé qui peut dépendre de la source (bulletins officiels, locaux, nationaux, services de tourisme, etc.). Or, les producteurs d'énergie renouvelable (ex. éolienne, solaire) ont besoin d'une prévision fiable et géolocalisée très finement pour décider du lieu d'implantation d'un champ d'éoliennes ou de panneaux solaires. Une fiabilité qui ne peut être obtenue que par le croisement de sources nombreuses dont le degré de véracité est pondéré[9]. Concluons sur le fait qu'il serait dangereux de penser que la véracité est un débat dont la criticité est exagérée. L'exemple du faux tweet de l'Associated Press annonçant en avril 2013 deux explosions à la Maison Blanche ayant sérieusement blessé Barack Obama est éloquent : ce faux tweet en réalité rédigé par des pirates informatiques syriens a été en l'espace de quelques minutes retweeté des milliers de fois. En deux minutes, le Dow Jones s'est effondré de 143 points perdant ainsi près de 1% de sa valeur, soit un montant de près de 136 milliards de dollars.

La valeur de l'information est bien sûr conditionnée par la véracité, mais il ne suffit pas qu'une information soit avérée pour qu'elle ait une valeur pour son utilisateur. Il s'agit bien sûr ici du caractère opérationnel de la donnée traitée pour son commanditaire. Les données prises isolément n'ont pas grande valeur car elles n'ont au fond pas grand sens. L'enjeu du Big Data est bien évidemment de faire émerger une vérité statistique par une agrégation structurée des données élémentaires[10]. Ceci s'applique tout autant aux données quantitatives qu'aux données qualitatives, mais les données qualitatives posent cependant des problèmes spécifiques. Prenons par exemple, les tweets échangés à l'occasion de la manifestation du 11 janvier 2015 faisant suite à l'attentat contre Charlie Hebdo. Difficile de leur trouver un sens d'un point de vue sociologique s'ils sont examinés individuellement tant ce qui frappe c'est la « pauvreté » de leur contenu. En revanche, ce sens s'éclaire lorsque l'ensemble est passé au crible de l'analyste, ce qui pour autant n'éteint pas la polémique sur les arrières pensées des manifestants et de non-manifestants ce jour-là[11]. En matière de valeur, on distingue généralement trois domaines d'application du Big Data pour les entreprises, ces trois domaines pouvant naturellement être impactés en même temps : une amélioration de la connaissance du client, une optimisation des processus organisationnels et de production de l'offre et du

[9] IBM (2012).- "*Analytics : The real-world use of big data. How innovative enterprises extract value from uncertain data?*".- 22 pages.- http://ibm.co/1GcIfWK

[10] Normier Bernard (2013).- « *Du Big Data au Bug Data en 5V* ».- Le Blog de Bernard Normier.- http://bit.ly/1LDiIWW

[11] Cassely Jean-Laurent (2015).- « *Manifestations du 11 janvier : qui était (vraiment) Charlie ?* ».- http://slate.fr.

service et enfin une détection des opportunités de modification du modèle économique[12]. Pour faire simple, un parallèle avec la démarche marketing consisterait à mieux connaître son client, optimiser son offre de produits et de services et enfin être à même de détecter et d'adresser de nouvelles opportunités de marché. Soulignons pour conclure sur la valeur que ces trois champs d'application du Big Data ont pour le premier une valeur marchande (les achats de données qualifiées), pour le deuxième une valeur en tant que levier (l'amélioration de la performance via une optimisation des processus de production et de distribution) et enfin pour le dernier une valeur d'actif stratégique (pour ce qui est de la détection de nouveaux marchés).

## 2- Les limites technologiques et juridique du Big Data

Si le Big Data est souvent présenté comme un nouvel « eldorado », il convient de prendre conscience des obstacles très concrets qui se posent aux scientifiques des données. A la frontière de l'informatique, des mathématiques et des statistiques, le Big Data contient en germe une promesse : aider les entreprises à bâtir leur stratégie marketing en analysant un gisement croissant d'informations. Deux types d'obstacles principaux restent à surmonter, technologique et juridique, avant d'aborder peut-être le point essentiel : le consommateur est-il au fond « réductible » à un ensemble de données aussi volumineuses soient-elles ?

### 2.1. Les limites technologiques

Dans le cadre d'un ouvrage à vocation généraliste et orienté vers les professionnels du marketing, nous n'aborderons les limites technologiques que de manière succincte laissant aux spécialistes des domaines concernés (informatiques, mathématiques et statistiques) le soin de prolonger le débat. L'obstacle qui se présente est celui posé par la volumétrie de données à traiter au-delà des problèmes de stockage.

Pour faire simple, trois solutions sont aujourd'hui plus communément adoptées. En premier lieu, une architecture de bases de données en « colonnes » se substituant aux bases de données en « lignes » traditionnelles. L'intérêt ? Un mode de compression efficace et partant d'une volumétrie moindre à extraire des disques durs lors de chaque requête, ce qui améliore les performances de restitution[13]. Mais pour accélérer les traitements, la meilleure solution est encore de remonter les données en mémoire vive depuis les disques de stockage : ces solutions sont qualifiées de traitements en mémoire ou « in memory analytics » et les disques ne sont dès lors plus sollicités que comme espaces de sauvegarde. Cette approche semble prometteuse mais elle est encore coûteuse à mettre en œuvre pour des volumes importants de données. Enfin, une approche intégrée consiste à optimiser simultanément le hardware et le software pour faciliter l'administration et l'exploitation des données (« appliances »). Cette solution haut de gamme combine des accès très rapides aux données et des logiciels exploitant les formats « colonnes » des bases de données, mais reste également très coûteuse sur une grande échelle. Face aux coûts des solutions actuellement proposées, certains acteurs de l'Internet (Google, Yahoo, Facebook, LinkedIn, etc.), parce qu'ils ont été confrontés dès l'origine aux limites technologiques ci-dessus évoquées, ont parié sur une technologie logicielle basée sur une architecture parallèle mais compatible avec le matériel existant et a priori moins onéreuse. Cette solution – ou plutôt ces ensembles de solutions – sont appelés « NoSQL » pour « Not Only SQL » et ce afin d'échapper aux trois standards dominants en matière de systèmes de gestion de bases de données, Oracle, MySQL et Microsoft SQL Server. En réalité, il existe encore des limites technologiques et de coûts aux traitements de volumes de données conséquents, et ce dans les quatre domaines qui forment ensemble le Big Data : la gestion des sources de données, la structuration des données, la construction d'insights et de modèles de décision et d'action.

---

[12] Databusiness.fr.- « *Impact et applications du Big Data en entreprise* ».- Portail francophone sur le Big Data et les Analytics.- http://bit.ly/1kgMUwF

[13] Walter Stéphanie (2014).- « *Les solutions technologiques du Big Data* ».- Big Data & Blog Digital.- http://bit.ly/1QIsaIF

## 2.2. **Les limites juridiques**

Une décision rendue par la Cour de justice de l'Union européenne le 06 Octobre 2015 va probablement faire couler beaucoup d'encre dans les mois qui viennent. Il ne s'agit ni plus ni moins que la suspension de l'accord connu sous le nom de « Safe Harbor » et qui encadre l'utilisation des données des internautes européens par les entreprises du Web dont bon nombre sont américaines[14].

De quoi s'agit-il ? Selon les termes de la Directive 95/46/CE sur la protection des données personnelles, entrée en vigueur en Octobre 1998, il est interdit à une entreprise ou une organisation de transférer des données personnelles vers un Etat non membre de l'Espace économique européen dont la législation ne garantirait pas un niveau de protection des données personnelles au moins équivalent à celui des Etats membres de ce même espace. Une première difficulté concerne l'interprétation de la protection des données personnelles entre les Etats-Unis et l'Europe : si les USA et l'Europe partagent peu ou prou le même objectif de protection de la vie privée, principes et modes d'application sont très différents. Pour garantir la conformité des entreprises américaines à la Directive européenne, le Département du commerce des Etats-Unis et la Commission européenne ont mis en place un cadre juridique appelé Sphère de sécurité[15] (« Safe harbor »). Dans ce cadre, une dérogation au non transfert des données en dehors de l'Espace économique européen est accordée aux entreprises américaines certifiées qui souscrivent aux principes européens, en particulier celui concernant l'intégrité des données : *« l'entreprise s'engage à n'utiliser les données collectées que dans le seul but pour lequel l'utilisateur a donné son accord »*. Cette dérogation concerne près de 4.000 entreprises américaines dont évidemment les plus gros acteurs informatiques et Internet : Microsoft, General Motors, Amazon, Google, Hewlett Packard, Facebook, etc. Or le 06 Octobre 2015, la Cour de justice de l'Union européenne a invalidé l'accord dit du « Safe harbor ».

Tout a commencé avec la saisie de cette même cour par un Internaute autrichien, Max Schrems, militant de la protection des données personnelles sur Internet, contestant le stockage de ses données Facebook sur un serveur américain[16]. Après avoir examiné sa demande, la Cour de justice européenne a estimé que les conditions d'application de la dérogation n'étaient effectivement plus réunies, en particulier depuis la révélation par le site Wikileaks du programme de surveillance des données numériques mis en place par l'Agence de sécurité nationale américaine (NSA) avec la collaboration active des sociétés privées américaines. Elle a donc invalidé l'accord, laissant le soin aux autorités américaines et aux institutions des états membres (dont la CNIL) chargées de la protection des données nominatives et du respect de la vie privée (G29) de trouver les solutions politiques, juridiques et techniques permettant une poursuite du transfert des données européennes vers le territoire américain[17]. A défaut, actions répressives et sanctions financières seront appliquées.

Or cette décision est loin d'être anecdotique, même si les entreprises concernées ont réagi en arguant de leur application d'autres codes de bonne conduite pour poursuivre les transferts de données personnelles vers leurs serveurs américains (clauses contractuelles types ou règles internes entre entreprises pour les transferts entre filiales). Cette invalidation crée tout de même un vide juridique sans précédent, au moins jusqu'à l'adoption d'un nouvel accord « Safe harbor 2 » dont les négociations sont par ailleurs déjà ouvertes. Au-delà, ce que condamne, usant de termes très durs, la Cour de justice européenne, ce sont les programmes de surveillance de la NSA jugés « *incompatibles avec les droits fondamentaux garantis par le droit européen* »[18]. La Cour européenne estime, selon les propos mêmes de la commissaire européenne à la

---

[14] Untersinger Martin (2015).- *« La justice européenne invalide le très controversé Safe Harbor, un accord sur les données personnelles »*.- Lemonde.fr.- 06 Octobre 2015.- http://bit.ly/1MOe8WE

[15] Pour plus de détails, voir le site export.gov des autorités fédérales américaines destinées à aider les entreprises américaines à exporter.- http://1.usa.gov/1PBnHsT

[16] LesEchos.fr.- « *Safe harbor : Européens et Américains ont trois mois pour conclure* ».- 16 Octobre 2015.- http://bit.ly/1ZQ7p3I

[17] Voir le site de la CNIL pour mieux appréhender les conséquences de l'invalidation du Safe Harbor.- http://bit.ly/1OObuQR

[18] LeMonde.fr (2015).- « *Les conséquences de l'invalidation de l'accord Safe harbor sur les données personnelles* ».- http://bit.ly/1GlP8ET

justice, Vera Jourova, qu'il est plus que jamais nécessaire d'avoir des « *garde-fous solides* » en matière de protection de la vie privée des Internautes.

Plus largement, les Internautes se montrent de plus en plus préoccupés par toutes les formes d'atteinte au respect de leur vie privée sur la toile. Rappelons que la protection de la vie privée des citoyens a été affirmée par la Déclaration universelle des droits de l'homme votée par les Nations unies en 1948 (article 2) et qu'en France, ce même droit est régi par l'article 9 du Code civil depuis la loi du 17 juillet 1970. Quatre sphères de la vie privée sont plus spécifiquement protégées : la protection du domicile, le secret professionnel et médical, la protection de l'image et enfin la protection de l'intimité[19].

Certes, il subsiste bel et bien un paradoxe dénoncé par le magazine Time[20] qui parlait de « *Génération Moi, Moi, Moi* » pour désigner les pulsions narcissiques de la génération du Millénaire qui s'expose ouvertement sur le Web tout en dénonçant les atteintes à leur vie privée. Il est vrai que la surveillance par son entourage proche (collègues, employeurs, relations, etc.) est davantage ressentie comme une menace que la surveillance au fond encore très abstraite des Etats. Mais entre les deux s'intercale l'utilisation des données nominatives par les organisations marchandes. Trois défis propres à Internet sont ainsi soulignés par un rapport récent de l'Unesco[21] : les progrès technologiques permettent de collecter de nouveaux types d'informations dont le caractère intime est de plus en plus avéré, les capteurs de santé ou de forme physique étant l'exemple le plus probant ; la collecte et la localisation des informations personnelles sont grandement facilitées : chaque ordinateur, téléphone portable ou autre appareil connecté à Internet possède une adresse IP unique qui en fait un identificateur aisément localisable ; l'accroissement de la puissance des ordinateurs et des architectures de réseaux et de SGBD offre aux acteurs publics et privés de nouvelles capacités de stockage et d'analyse des données personnelles. C'est, nous l'avons vu, tout l'enjeu du Big Data. A ces trois défis s'ajoutent deux conséquences propres aux nouveaux modèles économiques issus de l'économie numérique : la création de nouvelles opportunités d'usage commercial des données personnelles collectées via des services, des offres ou des applications gratuites et la nature transnationale des échanges sur Internet qui pose de nouveaux défis à la réglementation (nous avons abordé ce point avec le contentieux autour de la Sphère de sécurité).

Il ne fait donc aucun doute que l'Internaute va être de plus en plus sensible à la protection de sa vie privée et ne manquera pas dans un avenir proche de sanctionner les entreprises jugées permissives : il dispose pour cela de deux types de recours, la sanction judiciaire avec le renforcement des réglementations nationales ou l'arsenal de dispositifs permettant de protéger son identité réelle en la maquillant, l'effaçant ou l'altérant pour mieux brouiller les pistes[22].

### 3- <u>Résumer le consommateur à un ensemble de données est un leurre</u>

Avec la naissance du Big Data, nous assistons depuis quelques années à une abondante activité de publications, de colloques, de salons et de manifestations professionnelles, toutes marquées par un regain d'optimisme : la révolution qui s'annonce doit permettre aux organisations de mieux comprendre leur marché, d'optimiser leur processus et de s'inscrire dans une démarche d'innovation permanente. Si nous pensons que le Big Data ouvre de nombreuses perspectives, il convient toutefois de ne pas céder à l'illusion du Saint Graal.

---

19 Source : viepublique.fr.- « *Chaque citoyen a-t-il droit au respect de sa vie privée ?* ».- http://bit.ly/1NjXk8z

20 Time Magazine (2013).- « *The Me Me Me Generation* ».- Joël Stein, 9 mai 2013.- http://ti.me/1RRevzU

21 Mendel Toby et al. (2013).- Etude mondiale sur le respect de la vie privée sur l'Internet et la liberté d'expression.- Collection Unesco sur la Liberté de l'Internet.- 132 pages.

22 Il existe aujourd'hui des technologies qui permettent une navigation anonyme sur Internet. Il suffit de rejoindre un réseau appelé TOR, un acronyme pour « Tor Onion Routing », un ensemble de routeurs organisés en couches (appelées nœuds de l'oignon) qui transmettent de manière anonyme les flux TCP. Il est donc possible de naviguer sur Internet sans laisser de trace, une pratique qui pourrait être amenée à se généraliser.

## 3.1. Les limites d'une approche scientiste du Big Data

Le Big Data pourrait par bien des côtés être au marketing ce que le positivisme puis le scientisme ont été à la science : de vraies avancées mâtinées d'illusions non moins réelles. On retrouve dans le Big Data, les mêmes excès que ceux qui ont présidé au scientisme : la même volonté de tout contrôler, de tout comprendre en faisant définitivement entrer l'humanité dans l'âge positif.

Or les aspects « totalitaires » du scientisme ne peuvent être écartés, en même temps que le caractère parfois naïf des croyances qu'il véhicule, ainsi qu'en atteste la profession de foi du biologiste Félix Le Dantec, l'un des premiers à avoir employé le terme de scientisme dans un article de la Grande Revue en 1911 : « *Je crois à l'avenir de la Science : je crois que la Science et la Science seule résoudra toutes les questions qui ont un sens ; je crois qu'elle pénétrera jusqu'aux arcanes de notre vie sentimentale et qu'elle m'expliquera même l'origine et la structure du mysticisme héréditaire anti-scientifique qui cohabite chez moi avec le scientisme le plus absolu. Mais je suis convaincu aussi que les hommes se posent bien des questions qui ne signifient rien. Ces questions, la Science montrera leur absurdité en n'y répondant pas, ce qui prouvera qu'elles ne comportent pas de réponse.* »[23]. Enfin des outils qui permettront de dominer la complexité humaine, de tout savoir sur le consommateur et encore mieux de prédire ses comportements futurs. Plus d'aléas, plus de mauvais choix, l'algorithme Nouvelle Pierre Philosophale travaille pour nous. Il est la nouvelle boîte noire celle qui transforme des masses de données en informations utiles directement exploitables. C'est la nouvelle martingale des temps modernes. Mais quittons ce monde enchanté et féérique pour revenir aux réalités.

Retour d'abord sur un postulat de base de l'individualisme méthodologique : les données agrégées ne sont par construction que la résultante de l'agrégation des données individuelles, et ce qui intéresse le praticien en marketing, c'est de comprendre les motivations à l'origine des comportements individuels. Par définition, une démarche compréhensive s'intéresse en tour premier lieu au sens que les individus (ou les organisations) donnent à leurs pratiques et à leurs représentations[24]. Elle ne saurait donc se contenter de tests statistiques aussi sophistiqués soient-ils, car derrière les comportements il y a des rationalités à l'œuvre, multiples, qu'il convient précisément de révéler. L'achat d'un même produit peut au fond être la résultante de motivations fort différentes et s'inscrire dans des rationalités tout aussi diverses sans compter l'influence du contexte et de la situation d'achat.

## 3.2. Comment résoudre la contradiction inhérente aux nouveaux consommateurs ?

De surcroît, les consommateurs assument aujourd'hui des postures parfaitement contradictoires : ils plébiscitent la consommation collaborative mais restent arc-boutés sur des postures très individualistes ; ils disent vouloir se préserver des espaces d'intimité mais n'ont jamais autant affiché leur vie privée sur les réseaux sociaux ; ils affichent une conscience écologique aiguë mais refusent de payer plus pour rouler propre, etc. Prétendre que le Big Data, en raison des volumes de données individuelles qu'il permet de traiter, est une façon de résoudre la contradiction inhérente aux nouveaux comportements d'achat, c'est sans doute s'illusionner à peu de frais. Car lorsque le portrait statistique du nouveau consommateur commence à se dessiner, il a déjà changé, il n'est plus le même. Il échappe de plus en plus aux caractérisations sociodémographiques habituelles et relève de ce fait d'un déterminisme de plus en plus flou. C'est un consommateur qui brouille à dessein les cartes et contre lequel la grosse artillerie du Big Data apparaît vaine.

---

[23] « *Qu'est-ce que le scientisme ? Tentatives de définition* ».- Et vous n'avez encore rien vu : critique de la science et du scientisme ordinaire.- http://bit.ly/1hPHirx

[24] Pour plus de détails sur ce courant de la pensée sociologique, voir Max Weber.- « *De la sociologie compréhensive* ».- Les cahier de la Psychologie politique.- En ligne, n°19, Août 2011.- http://bit.ly/1M4xUhO

Puisque l'on est dans l'analogie militaire, deux exemples nous permettront de mieux illustrer notre propos. Lors de la guerre du Vietnam les américains utilisèrent pour la première fois des programmes informatiques sur une grande échelle pour estimer le nombre de morts nord-vietnamiens et en fonction d'estimations pointues sur les quantités d'armes disponibles (ou de bombes déversées) estimer l'échéance de fin des combats après l'anéantissement de l'armée nord-vietnamienne. Tout était calculé et tout échoua[25]. En premier lieu, parce que quantifier la victoire est un exercice autrement plus complexe que de dénombrer un nombre de victimes supposées ou de décompter des tonnes de bombes ou d'armement utilisés. Mais aussi parce que l'armée US avait en face d'elle une armée nord-vietnamienne d'une très grande agilité d'esprit, capable de changer de stratégie et de tactique, déjouant tous les calculs des stratèges du Pentagone engoncés dans les seuls schémas qu'ils connaissaient vraiment, ceux éprouvés dans d'autres conflits. Si vous alimentez des modèles de prévisions en vous inspirant de scénarii figés et peu pertinents, peu importe le volume de données que vous utilisez et les milliers d'informations que vous traitez ! On pourrait en dire de même du renseignement américain qui croule sous les informations mais qui est parfois incapable de les traiter au point de remettre au goût du jour les techniques classiques de l'espionnage au moyen des relations interpersonnelles (HUMINT pour « Human Intelligence Gathering ») par opposition aux techniques de collecte du signal (SIGINT), de l'image (IMINT) ou bien de la mesure et de la signature (MASINT)[26].

Le Big Data, c'est au fond la conception d'une guerre classique menée avec des armées nombreuses mais ces armées doivent aujourd'hui affronter un consommateur qui mène des combats qui relèvent de la technique de guérilla : guerre de la surprise permanente et guerre de mouvement. Lorsqu'on croit l'avoir repéré, il a déjà changé de tactique, car il est avant tout un tacticien. Pour lui le champ économique n'est un champ d'opportunités où il cherche à conserver sa liberté de manœuvre. Ainsi, d'années en années, nos enquêtes sur les nouvelles pratiques de consommation n'en finissent pas de nous désarçonner, tant dans un laps de temps très court le consommateur se redéfinit en permanence : il expérimente sans cesse d'autres tactiques, contribuant à faire émerger des pratiques nouvelles que rien dans les données du Big Data ne permettront d'anticiper. Pour conclure sur ce point, disons qu'on ne mène pas une guerre de mouvement, voire de guérilla, avec des armées qui sont programmées pour des affrontements classiques. Autant comme le disait TE Lawrence « *manger une soupe avec un couteau* » !

### 3.3. **Big Data et agilité stratégique et organisationnelle**

Au fond, Big Data ou non, difficile de comprendre les nouveaux comportements d'achat si on n'admet pas d'entrée que le nouveau consommateur se définit précisément par sa capacité… à se redéfinir en permanence et à adopter les nouveaux outils que la technologie met à sa disposition pour varier les modalités d'affrontement. Une enseigne emblématique comme Tesco, pourtant pionnière dans le traitement des données, a-t-elle su prévenir la désertion de nombre de ses clients ? Non. D'une part, parce que les enseignes de distribution, tout comme les marques, surestiment toujours leur degré de connaissance du client, et d'autre part, parce que trop focalisées sur les seules données relatives aux clients, elles sous-estiment les évolutions de la concurrence. L'émergence de nouveaux concepts de distribution, la rapidité de transformation du paysage commercial, l'apparition de nouveaux modèles économiques sont d'abord la conséquence de nouvelles attentes pas toujours explicitement formulées du consommateur, mais que certains acteurs savent saisir dans leur diversité avant d'autres. N'est-ce pas là la seule explication à l'ubérisation d'un certain nombre de services (transport, énergie, banque, assurance, services à la personne, etc.) ?

La stratégie a toujours eu besoin de l'information, elle se nourrit même de l'information mais elle ne s'y réduit pas et c'est là parfois l'illusion du Big Data. La montée en puissance des informations stockées et disponibles doit être concomitante avec une réflexion nouvelle sur les liens entre informations et stratégies, en se plaçant du point de vue du client. Clausewitz aimait à dire que « l'accidentel », la chance et le hasard

---

[25] Summers Harry G. Jr (1991).- « Body Count Proved to Be a False Prophet (..) ».- Los Angeles Times.- 09 Février 1991.- http://lat.ms/1LlEmxu

[26] Voir pour plus de détails, les informations données sur le site même de la CIA : http://1.usa.gov/1GQFrJw

étaient des données fondamentales de la guerre. Ceux qui imaginent que le Big data permettra d'éliminer ces trois aléas se trompent. Il ne fera que plus ou moins les circonscrire et ce n'est pas la même chose. Ce qu'il faut désormais concilier c'est l'agilité stratégique et organisationnelle avec le Big data en permettant aux « data scientists » d'alimenter les décideurs en « small data » au service d'objectifs peut-être plus limités mais sûrement plus efficaces.

## 4- Le Big Data : un outil supplémentaire pour les études marketing

La révolution du Big Data prépare-t-elle le « Big Bang » des études marketing ? Nous ne prétendons pas avoir épuisé le sujet, mais nous avons souhaité mettre l'accent en toute transparence et sans céder à l'hyperbole technologique sur quelques-uns des enjeux, des défis mais également des obstacles qui pourraient demain s'opposer à la généralisation du Big Data dans un très grand nombre d'activités humaines.

### 4.1. L'enjeu : gagner la confiance des consommateurs

Nous avons dans un premier temps insisté sur les défis technologiques posés par l'accroissement du volume, de la vélocité et de la variété des données aujourd'hui disponibles. En tant que chercheurs en analyse de données, nous n'avons pas manqué d'insister sur deux critères qui, de notre point de vue, dominent les autres, la véracité de l'information et sa valeur, les deux dimensions étant évidemment étroitement liées.

La véracité des données n'est pas qu'une affaire de finesse des capteurs ou de fiabilité psychométrique des instruments de mesure. La véracité est fortement dépendante de la volonté du consommateur de « coopérer » à une transmission des données privées, des informations et des transactions le concernant. L'enjeu n'est donc pas uniquement juridique, même si celui-ci est déterminant comme le prouve la décision récente de la Cour de justice européenne d'invalider l'accord de Sphère de sécurité entre les pays européens et les Etats-Unis. L'enjeu est de gagner la confiance des consommateurs. Sans confiance, les tentatives de dissimuler son identité, son opinion ou bien encore d'effacer toute trace de son comportement ne manqueront pas de se multiplier et aucune technologie ne pourra l'empêcher. A défaut, les scientifiques de la donnée seront nombreux à adhérer à cette maxime : « *chacun dénomme vérité le fief qu'il se taille au royaume de l'incertitude* » (Maurice Chapelan, Main courante, 1957).

### 4.2. L'enjeu : positionner la qualité de l'analyse au centre de la valeur

Lorsqu'on parcourt l'abondante littérature sur le Big Data, on ne peut qu'être surpris de la place accordée aux technologies, aux plateformes et aux choix matériels et logiciels. Certes, nous n'avons pas manqué de souligner les défis posés aujourd'hui par l'accroissement exponentiel des données disponibles. Mais au final, la place de l'humain dans la chaîne de traitement et d'analyse est déterminante et force est de constater que bien des outils du Big Analytics sont en réalité « *des extensions des outils classiques de datamining* »[27], quand bien même les problèmes spécifiques posés par les volumes en jeu nécessitent le développement d'applications spécifiques. Nous sommes profondément convaincus que les équipes en pointe dans le domaine de la Big Data seront nécessairement multidisciplinaires et feront appel à des informaticiens, des programmeurs, des statisticiens, des psychologues, des mathématiciens. Mais ce qui sera déterminant au final, c'est la capacité de ces équipes à dégager des tendances de consommation, des schémas de comportement propres à stimuler l'innovation et la croissance : l'enjeu est bien là et nulle part ailleurs !

Pour se faire, trois compétences distinctives nous semblent nécessaires. Une première concerne la capacité du « data scientist » à bien cerner les enjeux de la représentativité qui conditionnent la capacité à

[27] ZDNet (2015).- « *La valeur n'est pas dans le Big Data : la donnée est muette* ».- Symposium ITxpo, 05 Octobre 2015.

généraliser sans biais les résultats observés à l'ensemble d'une population. Les analyses porteront certes sur des échantillons sans commune mesure en termes de taille avec ceux aujourd'hui manipulés, mais se poseront toujours les mêmes interrogations sur la représentativité des sujets, la qualité des données, leur provenance, l'information contextuelle disponible sur la donnée, etc. Le traitement, fût-il exhaustif, des données d'un fichier clients, posera toujours le redoutable problème de la généralisation des enseignements ainsi retirés à l'ensemble des non clients pour comprendre les logiques de marché dans leur ensemble, en raison des biais évoqués ci-dessus. Une deuxième compétence porte sur une « ouïe » fine, capable de capter les signaux faibles (et porteurs de sens) au milieu d'un bruit de données parfois assourdissant. Les attitudes, les opinions, les comportements sont complexes, ambigus, déroutants parfois. Nous avons tenté de le démontrer. Dès lors, la recherche des déterminants (au sens causal), des facteurs explicatifs ou prédictifs n'en est que plus complexe. L'enjeu n'est pas ici dans la capacité à traiter les volumes de données, mais plutôt dans la capacité à mobiliser la ou les théories pertinentes pour confronter sans cesse la théorie à la réalité, par une suite d'itérations successives qui permettent d'approcher au plus près le réel. Enfin, la troisième compétence consiste à savoir dépasser les enseignements tirés des données du passé pour situer recommandations et plans d'action dans une anticipation constante des changements de contexte et d'environnement.

La principale faiblesse de la donnée issue du Big Data, c'est qu'elle est tournée vers le passé. S'arrêter aux seuls enseignements qui en découlent, c'est pour reprendre l'expression de Pierre Servent, s'inscrire dans « *l'esprit de défaite* »[28], une autre forme du complexe de l'autruche. En d'autres termes, c'est théoriser la victoire de 14-18 pour mieux préparer la défaite de 1940 ! Trop de commanditaires sont aujourd'hui persuadés que l'historique de données dont ils disposent leur permettra de comprendre l'environnement de demain sinon d'après-demain. Grave illusion, lorsqu'on songe aux écarts d'opinions, d'attentes, de valeurs, de comportements qui opposent aujourd'hui les préadolescents aux adolescents, aux jeunes adultes, aux adultes et plus encore aux seniors. Or, c'est bien la mobilisation des compétences pluridisciplinaires et le développement des champs d'analyse qualitative en sciences sociales (psychologie, sociologie, ethnologie, prospective, etc.) qui forment la meilleure réponse au défi ainsi posé.

### 4.3. **Le Big Data ne signifie pas le fin des études marketing, bien au contraire !**

Vous l'avez compris, nous ne partageons pas l'opinion de ceux qui pensent que le Big Data signifie la fin des études marketing. Le métier sera-t-il appelé à se transformer ? Oui certainement. A disparaître à tout jamais ? Non probablement pas sous réserve que les professionnels du secteur sachent prendre l'ampleur de la mutation en cours de leur métier.

Rappelons que les études marketing ont déjà affronté bien des (r)évolutions dont elles sont sorties sinon grandies, en tout cas renforcées : la généralisation des données de panels, la croissance des études longitudinales, l'intégration des données d'achats aux données d'usage, la modélisation prédictive, etc. Dans les années 1990, la révolution des PC et le remplacement des progiciels d'analyses de données par des logiciels de statistiques ont accompagné une première révolution dans le traitement de masse des données. Les acteurs majeurs de la profession (en tous cas, les meilleurs d'entre eux) ont à chaque fois su prendre les bonnes options stratégiques. Pourquoi ne seraient-ils pas de nouveau les acteurs les mieux placés pour gérer le changement qui se profile ? En réalité, l'expertise des Instituts d'études est bien appelée à muter mais elle est loin d'être obsolète dans le nouveau contexte dessiné par l'explosion des données. Si les besoins d'analyse qui se posent nécessitent des solutions technologiques qui restent encore pour partie à inventer, donner du sens et faciliter l'appropriation par les managers des enseignements que dessine la donnée consommateur fait appel aux qualités et à un savoir-faire que les Instituts d'études mettent en œuvre quotidiennement pour leurs clients. Car les instituts sont les mieux placés pour savoir que le consommateur est difficilement « réductible » à un ensemble de données quelles qu'elles soient.

[28] Servent Pierre (2013).- *Le complexe de l'autruche : pour en finir avec les défaites françaises*.- Editions Perrin.