Haute Ecole

Groupe ICHEC - ISC St-Louis - ISFSC

Enseignement supérieur de type long de niveau universitaire

Mesure d'impact du Big Data sur l'industrie du sondage traditionnel.

Mémoire présenté par
Karim Bouchdak

Pour l'obtention du diplôme de
Master en Sciences Commerciales

Année académique 2017 - 2018

Promoteur :
Monsieur Étienne Cuvelier

Boulevard Brand Whitlock 2 - 1150 Bruxelles

Remerciements

Nous ne comptons plus les jours, les heures et les minutes passés à la rédaction de ce mémoire. Un travail de longue haleine qui n'aurait pas pu voir le jour sans l'aide de bon nombre de personnes. C'est la raison pour laquelle nous souhaitons profiter de ces quelques lignes pour remercier celles et ceux qui nous ont permis d'atteindre notre objectif.

Dans un premier, nous souhaitons remercier tous les membres de notre famille qui d'une manière ou d'une autre nous ont aidé, soutenu et mis dans une situation de travail plus que confortable. Par la même occasion, nous remercions tous nos amis, collègues et autres connaissances qui ont répondu présents à nos appels.

Plus particulièrement, nous aimerions exprimer nos plus chaleureux remerciements à Linda K., pour la relecture, Assia A. et Hanane. B., pour le recrutement des experts, Marie N. et Abdelmounaim D., pour la traduction, Eugénia D., pour son travail de mise en page, ainsi qu'Alexandre L. pour le soutien informatique.

Nous n'oublions pas pour autant Dedicated, les membres de la direction et collègues, pour l'aide logistique apportée dans le cadre de nos enquêtes en ligne.

Aussi, nous remercions l'ICHEC et ses professeurs pour tous leurs apports en matière de connaissances et de savoirs, et un remerciement particulier à M. Étienne Cuvelier pour ses conseils et sa relecture.

Table des matières

Listes des figures et tableaux

Figures

Tableaux

Introduction générale

Plusieurs années se sont écoulées depuis notre arrivée au sein de l'institut de sondages et d'opinion, Dedicated. Cette société indépendante qui, évolue dans un environnement très concurrentiel, nous a permis d'accumuler plus de quatre années d'expérience au sein de l'univers des sondages traditionnels en tant que chargé d'études. Une formation pratique qui nous a fortement aidé lors de notre parcours académique à l'EPHEC, en marketing, dans un premier temps, et à l'ICHEC, en sciences commerciales, dans un second temps.

Passionné par cet univers et les éléments qui gravitent autour, nous avons voulu réaliser un mémoire qui puisse nous permettre de rester en contact avec l'objet de notre affection, les sondages traditionnels. Ces derniers sont par ailleurs intrinsèquement liés au caractère qui est le nôtre, à savoir : la curiosité. Celle-ci peut, par moments, être mal perçue par l'entourage qui, nous le comprenons, ne souhaite pas être le sujet d'un interrogatoire. Dès lors, les sondages traditionnels représentent, pour ce qui nous concerne, un remède, une forme de thérapie douce qui nous autoriserait à laisser s'exprimer notre curiosité.

Dans le cadre de notre questionnement sur notre sujet de mémoire, nous avons assez rapidement été confronté à la thématique du Big Data. Une technologie qui était décrite, selon nos souvenirs, comme supérieure et infaillible, pouvant être utilisée dans divers domaines dont celui qui nous concernait directement : l'univers du sondage. Cet outil semblait faire appel aux dernières trouvailles informatiques pour venir répondre à des questionnements de manière très précise et bien plus fiable que ne l'étaient les réponses fournies par l'industrie du sondage traditionnel. Autant l'avouer, nous avions aussi tôt effectué des liens avec notre emploi que nous pensions devenir obsolète en raison des apports colossaux que pourrait procurer le Big Data à nos clients.

C'est avec quelque peu d'appréhension que nous avions décidé d'aborder la thématique du Big Data en reliant celle-ci à l'univers du sondage dans lequel nous avions évolué jusqu'alors (et dans lequel nous exerçons toujours). Notre question de départ fut dès lors la suivante : « Quel est l'impact du Big Data sur l'industrie du sondage traditionnel ? ». Au moment de l'énoncé de ce thème de recherche, nous pensions assez naïvement, il faut le reconnaître, que si cette technologie n'était pas encore venue remplacer le travail réalisé par les instituts de sondages traditionnels, cela ne saurait tarder !

Une supposition que nous avions très vite écartée. En effet, l'image hégémonique que nous avions du Big Data n'était en définitive que le fruit de notre imagination. Le monde dans lequel nous évoluons est complexe, et pour traiter cette complexité, les algorithmes ne suffisent pas. Nos premières recherches ont d'ailleurs très vite soulevé pléthore de questions. La première s'intéressant à la définition même de cette récente technologie qui se fait appeler Big Data. Celle-ci ne semblait pas avoir de définition reconnue par tous les spécialistes qui en faisaient mention (Ollion et Boelaert, 2015). Par conséquent, notre raisonnement nous a fait prendre une distance critique vis-à-vis de cette appellation. Que désigne-t-elle concrètement ? Que représente-t-elle dans l'esprit des spécialistes ? Pourquoi ces derniers ne se mettent-ils pas d'accord sur les éléments composant cette définition ? Aussi, nous avions émis la déduction suivante : si l'on ne s'accorde pas avec tous les éléments de définition du Big Data et que ceux-ci font référence à des capacités

de cette technologie, il se pourrait bien que cet outil, dont on ne cesse de vanter les mérites, ne soit pas si performant qu'on souhaite nous le faire croire.

D'autres questions, qui découlent de la précédente, nous ont également traversé l'esprit. Nous nous interrogions sur les apports concrets du Big Data au sein des domaines de prédilection de l'industrie du sondage, comme le secteur privé, le secteur public, le monde politique et le secteur médical. Ensuite, tout naturellement, nous nous questionnions sur l'efficacité de cette technologie. Était-elle plus performante, plus précise que les modes de sondages traditionnels ? Quelles différences pouvions-nous observer entre les deux méthodologies ? (Etc.). Pour tenter de répondre à l'ensemble de nos questionnements, nous avons décidé de procéder minutieusement, pas à pas, en suivant la démarche que nous vous décrivons brièvement ci-dessous.

Dans un premier temps, nous aborderons les éléments de contexte qui nous ont semblé les plus pertinents. Ceux-ci auront pour objectif d'aider le lecteur à disposer de tous les outils nécessaires à la compréhension de notre étude. Nous passerons en revue au cours de cette section les définitions et modes de fonctionnement de l'industrie du sondage d'une part, et d'autre part, du Big Data. Nous parcourrons par la même occasion les forces et faiblesses de ces deux outils. Enfin, nous terminerons cette première partie en présentant de manière théorique les différents apports de la technologie Big Data dans les domaines de prédilection du sondage traditionnel.

Une fois les éléments de contexte appropriés, nous arriverons sur la partie pratique de notre étude. Ce sera l'occasion pour nous de vous présenter notre question de recherche qui s'inscrit dans la continuité de notre question de départ. Par la suite, nous établirons la méthodologie de recherche que nous avons adoptée pour répondre à notre question de recherche, et vous présenterons au moyen d'une desk research et d'une étude qualitative les résultats de notre étude.

Cette étude se terminera par une synthèse concise recoupant l'ensemble des informations pertinentes que nous avons pu récolter lors de notre partie théorique et lors de notre partie pratique. Enfin, les dernières lignes de ce travail aborderont les limites que nous avons rencontrées dans le cadre de ce projet d'étude et proposent les pistes de recherche futures qui gravitent autour de cette thématique.

Première partie : approche théorique

I. Introduction

C'est en mode miroir que nous avons décidé de vous présenter les deux outils qui font l'objet de notre étude. Cette mise en perspective permettra au lecteur dans un premier temps de se mettre à jour concernant l'industrie du sondage traditionnel ; une industrie qui a considérablement changé depuis son apparition, il y a un peu moins d'une centaine d'années sous une forme assez comparable à celle que nous lui connaissons aujourd'hui (Groves, 2011). Nous établirons ensuite le mode de fonctionnement des sondages (que nous utilisons ici et dans la majeure partie de ce document dans un sens plus large, regroupant les études de marchés et les sondages (cf. infra p.14)) et nous achèverons cette moitié de section en mettant en exergue les avantages et faiblesses que possède cette industrie.

La seconde moitié de cette section abordera, quant à elle, la technologie Big Data. Nous introduirons brièvement cette partie par l'historique de cet outil très récent. Par la suite, nous définirons ce que représentent les données avant de décrire ce qui se cache derrière l'appellation Big Data. Nous poursuivrons avec le mode d'emploi de cet outil qui sera plus étayé que celui du sondage traditionnel (en raison de la complexité et des zones d'ombre qui planent autour de cette technologie). Ensuite, nous aboutirons aux différents domaines d'application de cet outil qui entrent en concurrence avec l'industrie du sondage traditionnel (les autres utilisations ne nous intéressant pas dans le cadre de cette étude). Nous conclurons cette partie (qui terminera par la même occasion cette section) par les forces et faiblesses de la technologie Big Data, ce qui nous autorisera à infirmer ou à confirmer certains de nos a priori.

II. Les sondages traditionnels

1. Origine des sondages traditionnels

Comme il est coutume de l'entendre, il convient à l'Homme de se renseigner sur son passé pour mieux entrevoir la voie vers laquelle il se dirige. Il en est de même pour le travail qui nous concerne. C'est la raison pour laquelle il nous a semblé opportun d'établir une exhaustive, mais nécessaire rétrospection de l'univers des sondages d'une part, et du Big Data d'autre part (que nous aborderons dans un prochain chapitre). Notre regard sur ces évolutions devrait nous permettre de mieux appréhender la complexité de ces univers qui, pour subsister, n'ont cessé de s'accoutumer à leur nouvel environnement.

1.1. Genèse des sondages traditionnels

Alors que la discipline, qui consiste à analyser un marché selon une démarche codifiée afin que des décisions en découlent, est relativement récente sous la forme que nous lui connaissons, il nous faut plonger quelque peu dans l'histoire afin de trouver les esquisses de ce qui deviendra plus tard les études de marchés contemporaines (Groves, 2011). En effet, près de deux siècles séparent la première des trois ères des sondages (Groves, 2011)

et l'ancêtre du sondage qui remonterait à la moitié du dix-huitième siècle (Meynaud et Duclos, 2007).

En considérant que le recensement de la population et le dénombrement régulier de faits sociaux sont les précurseurs des enquêtes d'opinion, l'histoire des sondages est aussi ancienne que celles des Administrations responsables de la gestion d'un pays ou d'un territoire. En observant de plus près les temps anciens, on peut effectivement constater que bon nombre de citoyens français (fonctionnaires, prêtres…) étaient sollicités pour quantifier les habitants de leur commune, en précisant de manière concise les biens et revenus possédés par ces derniers (Meynaud et Duclos, 2007).

Toutefois, c'est en 1745, sous Louis XV, qu'une directive toute particulière sera émise par l'Administration publique française. L'objectif de cette dernière était de dresser un état des lieux de la population en matière de richesse, de niveau de pauvreté, d'effectif masculin potentiellement apte à rejoindre l'armée (et autres), mais également de propager de fausses rumeurs. Des rumeurs assez courantes pour l'époque mais qui, à la suite de leur dissémination au sein du territoire, devaient faire l'objet de constatation systématique par les intendants avec pour finalité la mesure des effets de leur diffusion. Une demande assez singulière et qui serait vraisemblablement une première dans l'histoire (Meynaud et Duclos, 2007).

Fin du dix-huitième siècle, les tensions sociales poussent le Régime français à laisser s'exprimer la voix du peuple tout en prenant bien soin de faire remonter ces informations au sein de l'Administration. La masse d'informations récoltées était bien entendu destinée à générer des actions adéquates de l'Autorité publique pour répondre efficacement aux souhaits de sa population (Meynaud et Duclos, 2007). Une démarche qui semble assez similaire à celles que nous constatons aujourd'hui, lorsque des partis politiques invitent la population à faire part de leur opinion sur certains sujets, mais bien entendu sous une forme beaucoup plus moderne.

Malgré l'échec cuisant de l'Ancien Régime pour canaliser l'opinion publique et éviter a posteriori la révolution de 1789, l'Administration publique tire les enseignements de cette période de l'histoire. Les Autorités sont désormais conscientes de l'utilité des données sociologiques et économiques du pays, alors qu'autrefois la finalité des « sondages » n'était centrée que sur la fiscalité et le recrutement pour le service militaire. Dès lors apparaît toute une multitude d'études, notamment des forces de l'ordre qui mesurent à présent l'état d'esprit de la population afin de se préparer et de contrecarrer toute nouvelle forme d'insurrection. Les possibles liens de causalité entre les conditions sociales précaires de certaines populations et le niveau de criminalité suscitent l'attrait pour les techniques dites d'observation, des méthodes assez similaires encore une fois à celles que nous utilisons aujourd'hui. On observe également de la part d'intellectuels de tous bords une envie de mieux comprendre l'esprit des masses pour, si ce n'est la réformer, l'apaiser. La presse se trouve également être un vecteur de communication en matière d'études sociologiques : type et niveau de criminalité, taux d'alphabétisation et autres. Enfin, d'autres études plus perfectionnées dans leur construction verront le jour, notamment par

l'intermédiaire d'Émile Durkheim ou encore Max Weber pour leurs études sur le comportement humain. Cependant, malgré l'engouement de cette période pour les connaissances sociétales, il faudra encore patienter quelque peu et traverser l'Atlantique, pour trouver une forme plus évoluée et plus proche des sondages d'opinion actuels, les « votes de paille » (Meynaud et Duclos, 2007).

1.2. Apparition des votes de paille

Les votes de paille, une simulation de vote électoral se déroulant peu de temps avant les élections officielles, voient le jour au début du dix-neuvième siècle (Giacometti, 2001). C'est en 1824 plus exactement que ce type de pratique est observé avec pour objet l'anticipation des résultats électoraux, qui tout naturellement encourageait le politique à réviser sa stratégie électoraliste (Meynaud et Duclos, 2007). Ces invitations à participer aux sondages étaient émises par les journaux qui, selon Jean Stœtzel (cité par Meynaud et Duclos, 2007, p.10), avaient pour fonction première la hausse des ventes de leur quotidien et la promotion des courants de pensée politique de leur propriétaire respectif.

Néanmoins, l'enthousiasme pour les « votes de paille » s'estompe très rapidement. Cause première de cette baisse d'attractivité, la représentativité de l'échantillon. Une représentativité qui, mise à l'épreuve, ne peut rivaliser avec l'arrivée des nouvelles méthodes scientifiques qui font montre d'une plus grande objectivité et d'une plus grande précision, au vu des résultats préélectoraux publiés en 1896. En effet, à cette époque, alors que les votes de paille ne prenaient en compte que les lecteurs d'une revue en particulier, The Record (journal de Chicago) mit en place un nouveau mode de sondage basé sur le principe de l'échantillon aléatoire. Ainsi, un électeur sur huit fut sélectionné de manière totalement aléatoire. Les résultats furent sans appel, une infime différence de 0,4% dans l'Illinois par rapport aux résultats officiels de cet État (Meynaud et Duclos, 2007, p.10).

À la suite de cet événement, la tendance fut de s'orienter de plus en plus vers une forme de scientificité des résultats par l'intermédiaire d'experts en mathématiques, en statistiques et de coopérations pluridisciplinaires afin que les résultats obtenus soient les plus proches de la réalité (Meynaud et Duclos, 2007). Ce n'est que quelques années plus tard, durant la période d'entre-deux-guerres, qu'apparaîtra la première ère des sondages tels que nous les concevons aujourd'hui (Groves, 2011).

1.3. Première ère des sondages : 1930-1960

Bien que la première ère des sondages ne débute (aux États-Unis) qu'en 1930 selon Groves (2011), un psychologue américain du nom de Daniel Starch se fait remarquer dès le début des années 1920 (Campbell, 1979). Ces travaux en études de marchés lui permirent de développer une méthode de mesure de l'efficacité de campagnes publicitaires. Dans la pratique, cette méthode consistait à attacher des coupons à des toutes-boîtes, et à demander aux individus de renvoyer ces coupons par voie postale en échange d'un cadeau (échantillon, livret…). Chaque coupon contenu dans un toutes-boîtes étant attaché à une publicité spécifique, il suffisait à Starch d'additionner les coupons pour mesurer l'efficacité publicitaire (AdAge, 2003). Par la suite, Starch tentera à l'aide d'une méthode probabiliste de mesurer l'audience totale d'une radio. Des

recherches qui porteront leurs fruits. En 1930, seuls 4% séparent l'estimation des auditeurs selon la méthodologie de Starch de celle, officielle, du bureau de recensement (Campbell, 1979).

L'année 1930 marquera « The Era of Invention » (Groves, 2011, p.862), l'ère de l'invention. Les statisticiens développeront des outils d'échantillonnage probabiliste, et Neyman depuis Londres, avant son arrivée aux États-Unis (Encyclopædia Universalis, 2018), publiera un article démontrant que les échantillons probabilistes ne présentent aucun biais et autorisent la mesure de l'erreur d'échantillonnage (Groves, 2011).

Parallèlement au développement des outils statistiques, des procédés de mesure de la population américaine sur divers sujets furent développés. Concrètement, trois groupes intéressés par l'opinion publique feront leur apparition. Le premier se formera autour de l'interview en face à face. Des journalistes constatant qu'il était bien plus intéressant pour leurs lecteurs d'avoir une idée globale de ce que pense l'ensemble de la population plutôt que de lire l'opinion de quelques individus sélectionnés aléatoirement. Le second se composera de sommités tels que Gallup (fondateur de l'institut éponyme) qui prennent conscience de l'utilité des études comportementales. Par leur truchement, les enquêtes qualitatives non-structurées se voient remplacées par des enquêtes quantitatives structurées. Enfin, le troisième parti intéressé par l'opinion publique n'était autre que le Gouvernement des États-Unis souhaitant, entre autres, plus d'informations sur la population en période de guerre (Groves, 2011).

Cette ère permettra à ce nouveau mode de sondages d'officialiser son efficacité et de se démocratiser. En effet, Gallup, à l'aide d'une méthode d'échantillonnage bien spécifique, identifiera avec une assez grande justesse (pour l'époque) le futur président des États-Unis, en n'interrogeant que quelques milliers d'individus sélectionnés de manière plus raisonnée. Les votes de paille indiquaient quant à eux le candidat perdant comme vainqueur, alors que plusieurs millions d'individus, ou plutôt de lecteurs de la revue Literary Digest avaient été interrogés (Meynaud et Duclos, 2007). Soit une première défaite, dans l'histoire, d'une ancienne version du « Big Data » (masse de données) face aux sondages d'opinion se basant sur un échantillonnage raisonné. À la suite de cet épisode, Jean Stœtzel (fondateur d'IPSOS) s'intéressera et importera en France la méthode Gallup qu'il nommera « sondage » (Meynaud et Duclos, 2007).

Nous noterons que le mode de collecte de données résultait principalement d'enquêtes en face à face, de courriers ou bien encore d'enquêtes par téléphone (pour les personnes qui pouvaient se le permettre à cette époque). Le taux de participation était très élevé, ce qui reste une difficulté majeure pour les études réalisées aujourd'hui (Groves, 2011).

1.4. Seconde ère des sondages : 1960-1990

Les trente années, qui suivront la première ère, seront appelées « The Era of Expansion » (Groves, 2011, p.864), l'ère de l'expansion. Cette période est avant tout marquée par le développement des lignes téléphoniques à travers les États-Unis. On les retrouve dans un premier temps auprès des ménages aisés, ensuite auprès de la population urbaine.

Combiné aux avancements technologiques de l'informatique, les lignes téléphoniques, qui serviront de base d'échantillonnage pour sonder la population, permettront la création d'un tout nouveau système d'enquêtes. Ce système, que l'on nomme Computer Assisted Telephone Interviewing (CATI), est encore utilisé aujourd'hui dans le cadre d'études en tous genres, lorsque la population à interroger est difficilement accessible via d'autres méthodes de recrutement. Pour résumer, cette technique de sondage consistait à contacter par téléphone des individus tirés au hasard, à les inviter à participer au sondage téléphonique et à lire les questions qui s'affichaient les unes après les autres sur l'écran des télé-enquêteurs (Groves, 2011). D'après Freeman, ces enquêtes par téléphone étaient surtout utilisées par le secteur privé (cité par Groves, 2011).

Cette période de l'histoire marque également l'accroissement de l'intérêt pour les études sociologiques. Des fonds sont levés pour investir dans la recherche, des formations viennent encadrer le travail des enquêteurs, le Gouvernement des États-Unis ainsi que le secteur privé augmentent quant à eux considérablement le nombre d'études réalisées, celles-ci étant moins contraignantes techniquement qu'auparavant (Groves, 2011).

Dans le même temps, d'autres méthodes de sélection de l'échantillon apparaissent afin de limiter les coûts de production ou faire face à des difficultés techniques (échantillon en grappes, échantillon par quotas). Au niveau de la création des questionnaires, les sciences sociales apporteront également de nouvelles connaissances en matière de psychologie humaine. L'objectif étant d'atténuer les biais liés à la formulation même des questions. Enfin, nous noterons également que des débats entre experts ont lieu pour juger de la pertinence de certains aspects des études, comme les biais liés aux non-réponses qui fausseraient les résultats obtenus (Groves, 2011).

1.5. Troisième ère des sondages : 1990 à nos jours

a. Apport d'Internet

Si la seconde ère se matérialise par le téléphone fixe, la troisième ère se distingue, elle, par l'avènement d'Internet. Cette nouvelle technologie apparue au début des années 1990 est venue bousculer le mode de fonctionnement des instituts de sondages. En outre, Internet est venu revigorer ce secteur d'activité en lui proposant un outil incomparable en termes de sélection et de recrutement d'échantillons (Groves, 2011). Mais avant d'aborder les apports d'Internet à l'industrie du sondage, il convient de resituer le contexte dans lequel se trouvait celle-ci au début de cette ère.

Lors des deux premières ères, la curiosité a permis aux instituts de sondages de récolter assez aisément, outre les difficultés techniques, le nombre d'enquêtes qu'ils souhaitaient obtenir dans le cadre de leurs études. Les personnes interrogées, soit en rue, soit par téléphone, étaient tellement étrangères au phénomène qu'elles ne voyaient aucune raison de s'y soustraire. La première ère représentait l'âge d'or pour les enquêtes en face à face avec un taux de participation atteignant les 90%. Quant à la seconde ère, tout appel téléphonique était sans doute une raison comme une autre de faire usage de ce nouvel appareil fraîchement acquis, ce qui octroyait aux sondeurs un taux élevé de réponse (Groves, 2011).

La troisième ère témoigne quant à elle d'une double détérioration. D'une part, celle des enquêtes réalisées en face à face qui affichent un taux de participation de plus en plus faible, et qui, par conséquent, sont moins souvent adoptées en tant qu'outil de sondages. D'autre part, celle des enquêtes réalisées par téléphone dont la chute du taux de réponse (l'engouement n'étant plus au rendez-vous) entraîne indubitablement une hausse des coûts pour ce mode de sondages. De plus, les nouvelles technologies en matière de téléphonie sont venues perturber les sondages par téléphone. Désormais, les enquêtes par téléphone mobile se doivent d'être bien plus courtes afin de s'adapter au nouveau comportement des individus qui effectuent certes plus d'appels téléphoniques, mais qui réduisent la durée de leurs appels. Cette réduction du temps d'interview est, en outre, associée à un taux de réponse encore plus faible de la part des utilisateurs de téléphone mobile en comparaison des utilisateurs de téléphone fixe (Groves, 2011).

L'industrie du sondage voyait son secteur d'activité peu à peu se dégrader, probablement en raison d'une certaine lassitude. Heureusement, le développement du réseau internet et des technologies informatiques est venu restimuler l'univers des sondages. En utilisant Internet comme vecteur de communication, les instituts de sondages peuvent désormais contacter tout un panel d'internautes prédisposé à participer à tous sondages en échange d'une rémunération (pratique assez courante). Dans un premier temps, le réseau internet et la disponibilité technique (ordinateur) n'autorisaient pas la seule utilisation d'enquêtes par Internet. De ce fait, ce mode de sondage était combiné à d'autres méthodes afin de compléter l'échantillon recherché pour l'étude (Groves, 2011).

Très rapidement, les enquêtes par Internet deviennent une référence en matière de sondages d'opinion, et ce, pour plusieurs raisons. Premièrement, l'usage de cette pratique a engendré la création de panels d'internautes. Autrement dit, des individus d'âge, de genre, de situation familiale et professionnelle (et autres) différents qui sont à disposition des instituts de sondages pour répondre à leurs enquêtes. Cette tendance a toutefois posé question, car elle a amené avec elle une forme de professionnalisation du métier de « répondant » qui n'existait pas auparavant. Deuxièmement, les répondants, ou plutôt les internautes, peuvent être contactés à tous moments sans restriction. Là où, les méthodes plus traditionnelles se sont de tout temps vu imposer une période fixe durant laquelle les personnes pouvaient être interrogées. Troisième raison, le coût des enquêtes par Internet reste de loin le moins élevé. Les coûts de production via les autres méthodes étant plus onéreux (et les budgets débloqués par les clients étant plus faibles), l'enquête par Internet devient l'instrument le plus envisagé. Enfin, la technologie a permis d'ouvrir de nouveaux horizons au niveau des questionnaires administrés. À présent, il devient techniquement possible de présenter aux répondants des visuels à l'écran tels des vidéos ou des images, ce qui permet dès lors de tester des campagnes publicitaires (Groves, 2011).

Cette ère caractérisée par les nombreuses percées technologiques donnera également naissance à une nouvelle forme de récolte de données. Celles-ci seront dénommées « Organic Data » et viendront peu à peu remplacer les précédentes, à savoir les « Designed Data » (Groves, 2011, p.866). Concrètement, les « Organic Data » sont des données issues de l'environnement digital. Des données qui sont automatiquement

enregistrées sur des serveurs après avoir été produites par des utilisateurs de réseaux sociaux (Facebook, Twitter…), le scanning des produits achetés par les consommateurs, par l'utilisation de cartes bancaires, etc. En d'autres termes, des données générées par des individus lambda qui ne sont pas toujours conscients des données personnelles de type comportemental qu'ils communiquent aux entreprises. À la différence, les « Designed Data » sont des données issues d'enquêtes réalisées. Les données récoltées proviennent dès lors de questionnaires qui ont été transmis à des répondants avec un but spécifique, comme par exemple, connaître leur intention de vote aux prochaines élections (Groves, 2011).

Groves (2011) nous explique à juste titre qu'une donnée ne représente pas forcément une information. L'information doit être en quelque sorte extraite d'une donnée, ou plutôt d'une série de données. Les « Organic Data », soit des enregistrements de clics ou autres encodages digitaux, proposent un ratio [nombre d'informations / nombre de données] bien moindre que celui des « Designed Data », issues d'enquêtes par questionnaire. Néanmoins, le coût de production de ces « Organic Data » étant moins élevé, la tendance se dirigerait de plus en plus vers ce type de données pour produire de l'information pertinente. Alors que selon Groves (2011), il serait plus bénéfique de tirer parti de ces deux modes de récoltes de données.

b. Dernières tendances dans l'industrie du sondage

b.1. Attrait pour les études quantitatives

Comme nous avons pu le constater, l'industrie du sondage à travers les âges n'a cessé d'évoluer et de se renouveler pour mieux correspondre à son temps. De la première à la troisième ère, ce sont deux difficultés majeures qui sont venues modifier les pratiques en matière de sondages d'opinion. La première difficulté résidait surtout dans la chute du taux de réponse que ce soit via les enquêtes en face à face ou les enquêtes par téléphone. La seconde difficulté rencontrée était de type pécunier. Le taux de réponse se réduisant à vue d'œil, les instituts de sondages se devaient de mobiliser plus de moyens humains et financiers pour atteindre leurs objectifs. Ce qui avait pour effet de gonfler les coûts de production de manière considérable (Groves, 2011).

Les tendances de ces dix dernières années, comme indiqué dans le tableau de la page suivante, donnent un aperçu de l'évolution des modes de sondages choisis pour récolter de l'information. On observe tout naturellement une forte croissance du taux de sondages réalisés par Internet qui avoisinait 27% en 2016, soit un sondage sur quatre. Aux dépens des modes de sondages plus traditionnels comme les enquêtes par téléphone et les enquêtes en face à face qui représentent chacun moins de 10% de part de marché (Esomar, 2017, pp.154-155).

Global Market Research Methodologies	2006	2013	2014	2016
Total quantitative	**83%**	**74%**	**73%**	**70%**
– Online	16%	24%	23%	27%
– Automated digital/electronic	–	19%	21%	12%
– Phone	19%	12%	9%	8%
– Face-to-face	12%	9%	8%	7%
– Mobile/smartphone	–	-	3%	5%
– Postal	4%	3%	2%	1%
– Online traffic/audience	–	2%	2%	6%
– Other quantitative	32%	5%	5%	4%
Total qualitative	**14%**	**16%**	**16%**	**16%**
Other	**3%**	**10%**	**11%**	**14%**

Tableau 1 : Évolution des pratiques en matière de sondages

Sources : les données chiffrées reprises sur ce tableau recomposé proviennent de trois sources différentes. Pour les données de 2006 et 2014, elles proviennent du rapport Esomar mais ont été prises sur le site « www.marketstratégies.com » (Hon ho, 2016) ; pour les données de 2013, elles proviennent directement du rapport Esomar 2014 (Esomar, Rapport annuel, 2014, pp.122-125) ; et pour les données de 2016, elles proviennent directement du rapport Esomar 2017 (Esomar, Rapport annuel, 2017, pp.154-155). Le lecteur tiendra également compte du fait que ces moyennes proviennent de données recueillies à l'international, mais certains pays ou entreprises n'ont pas fourni les informations nécessaires à Esomar (Esomar, Rapport annuel, 2017). Ces données permettent néanmoins de nous donner un ordre d'idée sur les tendances du secteur.

Ce tableau indique par ailleurs que les enquêtes par téléphone et les enquêtes en face à face deviennent des outils relativement obsolètes. En l'espace de dix ans, on a observé une diminution de 11% pour les enquêtes par téléphone dont le ralentissement se faisait déjà ressentir aux États-Unis courant 2006 (Hon ho, 2016). Et une réduction de 5% pour les enquêtes en face à face qui, selon Hon ho (2016), seraient plutôt utilisées lors de mini-enquêtes auprès des passants. Toutefois, selon ce même expert, ces deux modes de sondages restent encore la norme dans les pays en voie de développement. Une affirmation qui s'explique assez logiquement. À titre d'exemple, l'accès à Internet dans certains pays africains oscillerait entre 20% et 30% (Frintz, 2015), ce qui, inévitablement, freine l'usage d'Internet pour les sondages.

Les dernières années ont également vu apparaître un nouveau mode de sondages, les enquêtes via mobile (Hon ho, 2016). Il ne s'agit pas ici d'enquêtes par téléphone, mais plutôt d'enquêtes en ligne qui se réalisent via les applications mobiles. L'attrait pour ce mode d'enquêtes ne cesse de s'accentuer, car il permet d'atteindre des répondants en déplacement.

Autre élément pertinent qui ressort de ce tableau, le taux d'enquêtes quantitatives qui, combiné à celui des « autres » types d'enquêtes (établies également sur de grandes populations), représente 84% de l'ensemble des études réalisées dans le cadre de sondages

en tous genres. Ce qui démontre une plus grande attractivité de la part des clients d'instituts de sondages pour des informations qui peuvent être extrapolées sur l'ensemble d'une population. Pour rappel, les enquêtes qualitatives permettent de récolter des informations bien plus détaillées que les enquêtes quantitatives, mais uniquement auprès de petits échantillons. Cette petite taille d'échantillon n'autorise pas l'extrapolation des résultats sur l'ensemble d'une quelconque population étudiée.

b.2. Apparition des nouvelles technologies

Grâce aux récents progrès technologiques, de nouveaux outils ont vu le jour et sont venus renforcer les instituts de sondages dans leurs tâches quotidiennes. Ces instruments de mesure (qui peuvent toutefois comporter quelques imperfections) ouvrent de nouvelles perspectives dans l'univers des sondages. Ils permettent entre autres d'analyser des individus sans même les interroger, en utilisant par exemple le « eye-tracking », un outil qui permet de comprendre le comportement d'achat d'une personne en se contentant d'observer le mouvement de ses yeux (Hon ho, 2016). De telles méthodes d'observation permettent aux chercheurs d'éviter d'influencer d'une quelconque manière les réponses de personnes faisant l'objet d'études.

L'organisme Esomar nous fait part également de deux nouvelles tendances observées auprès des chercheurs. La première consiste à surveiller les réseaux sociaux, la seconde à utiliser des outils d'analyses du web. Ces tendances indiquent, toujours selon Esomar, une volonté à peine voilée de la part de l'industrie du sondage d'automatiser une partie de son activité, avec à la clé un gain de temps et de précision. Néanmoins, rappelle Esomar, même si l'on reconnait l'efficience des nouveaux outils technologiques que ce soit pour la transmission de questionnaire ou pour éviter d'importuner les répondants, l'humain aurait toujours sa place dans l'univers des sondages. La récolte de données à elle seule ne suffit pas, encore faut-il comprendre pour quelles raisons une personne agit de telle ou telle façon. C'est cette expertise en particulier que l'intelligence artificielle ne peut soustraire à l'intelligence humaine (Esomar, Rapport annuel, 2017).

Après lecture du paragraphe ci-dessus, il devient très compliqué de ne pas faire de lien avec la technologie du Big Data (dont nous ne définirons pas les contours à ce stade-ci, étant donné la complexité du concept). En effet, mesurer l'activité du web dans sa globalité, que ce soit via les réseaux sociaux ou autres sites internet, consiste à récolter des masses de données. Ces données doivent, ensuite, être rassemblées et analysées dans le même temps afin de pouvoir en extraire des informations consistantes, qui à leur tour devront permettre aux mandateurs de prendre des décisions. Les chapitres suivants permettront de nous en dire un peu plus sur cette technologie et son impact sur l'industrie. Mais avant, il nous faut nous attarder quelque peu sur l'utilité de notre rétrospection.

1.6. Rétrospection, un passage obligatoire

Comme nous l'avons mentionné au tout début de cet historique, il nous a paru nécessaire de prendre connaissance de l'histoire des sondages afin de mieux comprendre l'état actuel de cette industrie, et de mieux entrevoir son avenir. Sans ce passage historique, d'aucuns pourraient penser que l'avenir du sondage traditionnel est tout tracé. Celui-ci serait bien

évidemment remplacé par les nouvelles technologies (le Big Data pour ne pas le nommer) qui n'auraient plus besoin de faire appel aux anciennes pratiques, jugées révolues, d'un autre temps. Une erreur de jugement patente que nous estimons non-raisonnable. D'autres pourraient considérer que les sondages n'intéressent plus personne. Alors qu'encore une fois, il s'agit bien là d'une erreur d'appréciation.

Dans un premier temps, l'histoire des sondages nous a démontré à maintes reprises qu'elle a su se renouveler et s'adapter à son temps. De l'enquête en face-à-face, aux enquêtes par téléphone, en terminant par les enquêtes en ligne et toute une pléthore de nouveaux outils, nous ne pouvons que constater que cette industrie s'est approprié au fil du temps les nouvelles technologies et techniques de sondages qui font d'elle un outil indispensable dans les sciences politiques et sociales, mais également dans le monde de l'entreprise où les sollicitations ne cessent d'être émises par les départements marketing en particulier.

Dans un second temps, le passé nous apprend que la quantité n'est pas toujours synonyme de qualité. Pour rappel, Gallup nous en avait fait la démonstration en 1936 avec une étude auprès d'un petit échantillon de lecteurs d'une revue. Ses pronostics concernant l'élection du futur Président des États-Unis se sont révélés corrects, à la différence de ceux émis par le détenteur de la revue en question qui avait interrogé des millions de ses lecteurs (Meynaud et Duclos, 2007). Le Big Data, qui pour certains représente l'avenir du sondage, avait échoué dans l'estimation des intentions de vote pour les élections présidentielles des États-Unis, en tout cas dans sa version originelle. Si le Big Data a déjà failli par le passé, il n'est pas impensable qu'il faillisse à nouveau, quand bien même les technologies se sont considérablement améliorées.

Enfin, notre rétrospection a également mis en exergue l'appétit croissant de la société pour la connaissance. Les Autorités publiques françaises se sont intéressées depuis la période prérévolutionnaire à sa population, en essayant par exemple d'évaluer son taux d'insatisfaction (Meynaud et Duclos, 2007). De plus, ajoute Jacques Antoine (2005), les sondages permettent aux politiques et gouvernements de mieux mesurer le pouls de l'opinion publique, afin de savoir s'ils auront leur soutien lors des prochaines élections ou lors de leurs prochaines démarches gouvernementales. Le secteur privé est lui aussi demandeur de sondages. D'ailleurs, les études commandées par le privé représentent l'essentiel du chiffre d'affaires des instituts de sondage (Antoine, 2005).

Les sondages semblent avoir de l'avenir devant eux, reste à savoir sous quelle forme ?

2. Sondages traditionnels, mode d'emploi

De manière générale, nous pouvons considérer que l'industrie du sondage réalise, d'une part, des sondages d'opinion, et d'autre part, des études de marchés. Ces deux terminologies couramment usitées dans les médias peuvent paraître similaires aux non-initiés, mais elles possèdent toutefois des significations bien distinctes. Les sondages d'opinion ont pour objectif de mettre en exergue les opinions de la population étudiée dans un contexte particulier afin de répondre à une problématique définie en amont (ce qui la distingue du Big Data comme nous le verrons plus loin (cf. infra p.49)). En guise

d'exemple le plus courant, les sondages politiques commandés régulièrement par les médias ou les partis politiques qui souhaitent obtenir un état des lieux de la situation politique du pays, d'une région ou d'une commune en particulier. Les études de marchés, quant à elles, s'insèrent dans l'action marketing. Elles sont d'ailleurs l'une des premières démarches marketing, si ce n'est la première. L'étude de marché a pour objectif d'identifier les besoins du marché, sous-entendu ceux des consommateurs, en termes de produits et services, mais également d'établir un profilage des consommateurs (leurs motivations d'achat, les comportements adoptés, etc.) pour permettre au marketing de mieux cerner le marché dans lequel il opère et de proposer un produit ou un service approprié. Les exemples sont ici très nombreux : étude sur l'appréciation d'un nouveau soda, étude sur le comportement menant à l'achat d'un nouveau véhicule, étude de satisfaction d'un service rendu par une société, etc. (Vandercammen et Gauthy-Sinéchal, 2014).

2.1. Mode de fonctionnement

Quelle que soit la demande de leurs partenaires, l'industrie du sondage procèdera selon les huit étapes (Vandercammen et Gauthy-Sinéchal, 2014) que nous décrivons brièvement ci-dessous :

1. identification du problème : où l'objectif est de cerner la problématique à analyser. À ce stade, les chargés d'études se doivent de bien spécifier avec leurs clients l'objet d'étude. Cette spécification devra pour ce faire respecter trois principes : la précision, la non-restriction et l'exclusion de tout a priori ;
2. formulation des hypothèses : cette étape a pour but d'émettre quelques possibilités de réponse à la problématique qui aura été identifiée lors de l'étape précédente et qui seront bien entendu confirmées, infirmées ou nuancées en fin du processus de recherche ;
3. élaboration du plan de recherche : à ce stade, il sera question d'identifier :
    - les informations nécessaires à récolter,
    - les variables qui devront être mesurées,
    - les sources d'informations utiles, c'est-à-dire les informations primaires (résultant d'une étude de marché) et les informations secondaires (qui existent déjà comme des statistiques sur la répartition de la population),
    - la méthode de collecte des données, c'est-à-dire des études de type :
        – quantitatif : réalisées auprès d'un grand échantillon de la population ; selon les besoins identifiés, il s'agirait par exemple d'interroger 500 – 1000 – 2000 individus de la population de référence. Les enquêtes sont généralement très courtes, une dizaine, voire une quinzaine de minutes, et comportent en majorité des questions fermées. L'objectif est ici de

projeter les observations (par exemple, en termes de comportement des individus) sur l'ensemble de la population étudiée ;

– qualitatif : réalisées auprès d'un petit échantillon (par exemple, une enquête auprès de 50 médecins) ; des interviews généralement de longue durée, avec une majorité de questions ouvertes. L'objectif sera de mieux comprendre les facteurs qui amènent un individu à agir de telle ou telle façon, sans qu'aucune projection ne soit réalisée sur l'ensemble de la population ;

- les méthodes d'échantillonnage : comme l'échantillonnage par quotas, l'un des plus usités selon notre expérience, qui a pour objectif de constituer un échantillon restreint, mais représentatif de la population (par exemple, lorsque la population totale étudiée possède 52% de femmes, la taille de l'échantillon qui sera constituée et analysée, admettons dans le cas où n = 1000, comportera 520 femmes sur 1000 individus interviewés. D'autres types d'échantillonnage existent et sont utilisés dans le cadre de sondages tels que l'échantillonnage aléatoire, stratifié ou non, etc. ;

- les méthodes d'enregistrement des informations ;

- et enfin, les méthodes d'analyse et de dépouillement des informations ;

4. élaboration de l'instrument de collecte des données : dans la pratique, il s'agirait par exemple d'utiliser un questionnaire papier dans le cadre d'études quantitatives ou d'un questionnaire en ligne. De manière générale, l'on distingue quatre grands groupes, à savoir : les enquêtes par Internet, les enquêtes en face à face, les enquêtes par téléphone, les enquêtes par correspondance (qui ne sont, comme nous avons pu l'observer (cf. tableau supra p.12), pratiquement plus utilisées) ;

5. enregistrement des données recueillies : généralement, les données sont récoltées à l'aide de logiciels (plusieurs existent sur le marché) et préservées dans une base de données ;

6. traitement et analyse des résultats : cette étape consiste à nettoyer la base de données, croiser les données recueillies, dégager l'essentiel, etc. ;

7. interprétation des résultats : cette étape s'inscrit dans la continuité de la précédente et a pour but de mettre à jour les éléments répondant à la problématique initiale ;

8. et enfin, la remise et la présentation des résultats auprès du client.

De cette courte description du mode de fonctionnement de l'industrie du sondage, le lecteur retiendra essentiellement trois caractéristiques qui la distinguent du Big Data. Dans un premier temps, il s'agit du questionnement initial qui précède tout étape de recherche (soit le point 1). Autrement dit, les chercheurs considèrent en amont la direction à prendre pour parvenir à trouver une solution à leur problématique, c'est-à-dire en posant

le problème et en définissant ses contours. Nous verrons qu'il n'en est pas de même avec la technologie du Big Data qui se veut athéorique (cf. infra p.49).

Dans un second temps, la construction de l'échantillonnage représente une autre ligne de démarcation avec le Big Data. Le chercheur, via la méthode traditionnelle, n'a d'autre choix que de sélectionner une infime partie de la population qu'il souhaite étudier. Il établira dès lors les caractéristiques de cette population de référence en termes de répartition au sein des catégories d'âge et de genre, des catégories socioprofessionnelles, etc., pour ensuite construire un échantillon représentatif de cette population de référence. Le Big Data, quant à lui, prétend analyser les données directement auprès de l'ensemble de la population de référence (cf. infra p.49).

Enfin, la méthode de récolte de données est également différente. Dans le cas du sondage traditionnel (ou de l'étude de marché), le modus operandi consiste à réaliser des questionnaires et à les transmettre aux répondants (que ce soit par téléphone, par Internet, etc.) qui sont conscients d'être interrogés. Alors que le Big Data permettrait de soustraire l'information à l'insu de la population étudiée (cf. infra p.49).

2.2. Les qualités et faiblesses du sondage traditionnel

Outre sa capacité à innover et à se renouveler (cf. supra p.13), l'industrie du sondage possède plusieurs cordes à son arc qui lui permettent, selon toute vraisemblance, d'être encore à ce jour un concurrent de poids face au Big Data (nous verrons par la suite si c'est effectivement le cas). Les points qui suivent abordent brièvement quelques-unes des qualités de cette industrie, ainsi que les faiblesses de celle-ci qui pourraient faire pencher la balance en sa défaveur.

a. Les qualités du sondage traditionnel

Pour parvenir à leurs fins, les chercheurs font usage de petits échantillons construits de manière raisonnée. Cette procédure permet à ces derniers de restreindre les coûts de récolte et de traitement des données. Dès lors, même si un échantillon plus large, voire totalement exhaustif, serait bien plus précis, la construction d'un échantillon restreint autorise les chercheurs à proposer des tarifs raisonnables à leurs clients souhaitant des études en tous genres (Mooi, Sarstedt et Mooi-Reci, 2018).

De plus, Edwards Deming (cité par Strong, 2015) affirme que l'échantillonnage permet d'interroger plus en profondeur les individus, en établissant les incohérences de ceux-ci dans leurs réponses, la véracité de la réalité dont ils font part, etc. Ce processus améliorait significativement la qualité de la recherche, en tout cas mieux qu'un recensement auprès de l'intégralité de la population. Strong (2015) ajoute également que la majorité des erreurs occasionnées lors d'enquêtes proviennent d'études qui sont réalisées sur l'ensemble de la population (soit sans aucune marge d'erreur). Une réalité quelque peu difficile à concevoir, mais qui découlerait d'une sélection intelligente de l'échantillon (Mooi, Sarstedt et Mooi-Reci, 2018).

Autre évidence en faveur de la construction de petits échantillons, la rapidité avec laquelle les chercheurs obtiennent, traitent et présentent à leurs clients les résultats obtenus. En effet, quelle que soit la technologie utilisée, les chercheurs récolteront et traiteront bien plus rapidement les données provenant de petits échantillons puisqu'en toute logique celles-ci sont moins nombreuses (Strong, 2015).

Aussi, comme l'indique le tableau ci-dessous, l'exhaustivité d'un échantillon n'apporte pas toujours un avantage substantiel aux chercheurs. En utilisant un échantillon de 200 personnes les statistiques nous démontrent que la marge d'erreur pour une proportion observée de 50% est de 6.9%. Elle est de 3.1% pour un échantillon de 1.000 et de 1.4% pour un échantillon de 5000 personnes (Strong, 2015, p.21), soit un gain de précision de 55% pour un échantillon multiplié par cinq.

Taille échantillon « n »	Proportion « p » de la population mère				
	p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5
100	0.059	0.078	0.090	0.096	0.098
200	0.042	0.055	0.064	0.068	0.069
300	0.034	0.045	0.052	0.055	0.057
400	0.029	0.039	0.045	0.048	0.049
500	0.026	0.035	0.040	0.043	0.044
600	0.024	0.032	0.037	0.039	0.040
700	0.022	0.030	0.034	0.036	0.037
800	0.021	0.028	0.032	0.034	0.035
900	0.020	0.026	0.030	0.032	0.033
1'000	0.019	0.025	0.028	0.030	0.031
1'200	0.017	0.023	0.026	0.028	0.028
1'600	0.015	0.020	0.022	0.024	0.025
2'000	0.013	0.018	0.020	0.021	0.022
3'000	0.011	0.014	0.016	0.018	0.018
4'000	0.009	0.012	0.014	0.015	0.015
5'000	0.008	0.011	0.013	0.014	0.014
7'500	0.007	0.009	0.010	0.011	0.011
10'000	0.006	0.008	0.009	0.010	0.010
12'000	0.005	0.007	0.008	0.009	0.009
14'000	0.005	0.007	0.008	0.008	0.008
16'000	0.005	0.006	0.007	0.008	0.008

Figure 1 : Les marges d'erreur

Source : François Daniel Giezendanner. (2012). *Taille d'un échantillon aléatoire et Marge d'erreur.* Récupéré le 30 mai 2018 de http://icp.ge.ch/sem/cms-spip/IMG/pdf/taille-d_un-echantillon-aleatoire-et-marge-d_erreur-cms-spip.pdf

Enfin, vouloir interroger la population dans son ensemble pourrait s'avérer ingérable dans la pratique. D'ailleurs, à supposer que cela soit possible, le temps d'interroger l'ensemble de la population sur une émotion d'un moment et d'analyser ces données, il se pourrait que cette même population ait déjà changé de point de vue au moment où le chercheur transmettrait les résultats (Accountlearning.com, 2018).

b. Les faiblesses du sondage traditionnel

S'il est vrai que le sondage traditionnel possède de nombreuses qualités, l'outil n'en reste pas moins dénué de tout défaut. En effet, les sondages traditionnels se sont à maintes reprises montrés inefficaces et ont perdu en crédibilité vis-à-vis tant du public (qui croit moins aux sondages préélectoraux) que des professionnels. Tronchet (2017) nous rappelle par ailleurs certains des échecs récents des sondeurs comme le Brexit, l'élection de Donald Trump, les primaires françaises, etc.

Plusieurs facteurs peuvent expliquer les contre-performances des sondeurs en matière de prédiction. La première n'est autre que les biais liés à la sélection de l'échantillon. Strong (2015) présente quelques-unes des sources de biais :

- le biais d'autosélection : qui survient lorsque les individus proposent eux-mêmes aux instituts de faire partie de leur panel afin de pouvoir participer régulièrement à des sondages d'opinion ou autres études de marchés. Les instituts de sondages préfèrent bien évidemment constituer ce type de panel et entretenir celui-ci plutôt que procéder par recrutement aléatoire. La constitution de panels de répondants est donc financièrement plus rentable et pratiquement plus simple à mettre en place, mais statistiquement moins fiable. Tronchet (2017) ajoute que ces individus qui acceptent de répondre régulièrement aux sondages contre rémunération mènent à une forme de professionnalisation. « Être sondé » représenterait un moyen comme un autre d'obtenir quelques avantages financiers ou en nature. Une opinion que nous partageons, et qui nous oblige dans la pratique de notre métier à réaliser de plus en plus de contrôles afin de repérer ces professionnels (qui ont tendance à répondre plus rapidement que la normale ou à répondre positivement à toutes les questions posées, etc.) qui par moments n'hésitent pas à mentir par appât du gain (Tronchet, 2017) bien que ce montant reste assez dérisoire selon nous ;
- le biais de sous-couverture : qui résulte de la difficulté qu'ont les instituts de sondages à construire des échantillons représentatifs de la population et de ses sous-constituants. Par exemple, lorsque par souci de facilité un sondage politique est réalisé par Internet au sein d'une des communes belges, il arrive qu'une proportion significative d'une sous-population ne soit pas interrogée et ne figure dès lors pas dans l'échantillon (ex. : les individus votant pour de petits partis, issus d'un milieu plus défavorisé, etc.), ce qui logiquement fausse les résultats. Ce biais arrive tant par Internet que via les autres modes de sondages. En guise de second exemple, la méthode d'enquêtes par téléphone. Si un institut de sondages décide d'enquêter uniquement via téléphone fixe, l'échantillon interrogé sera plus que probablement constitué d'une population âgée, voire très âgée, et dès lors non-représentatif. La solution serait bien entendu de mixer les méthodes de sondages (téléphone fixe et mobile, Internet, face-à-face) pour obtenir une meilleure qualité d'échantillonnage, mais d'un point de vue purement financier cela représente un coût supplémentaire non-négligeable ;

- enfin, d'autres biais existent, comme le biais de réticence qui est provoqué par le type de question posé. Certaines questions portent effectivement trop sur l'intimité des répondants et poussent ces derniers à mentir, ou en tout cas à éviter d'apporter toutes les précisions souhaitées par le sondeur. Il existe également les biais liés à l'ordre d'apparition des questions ou des modalités de réponses, des biais liés à la manière de formuler les questions ainsi que les modalités de réponse qui peuvent être interprétées différemment selon les répondants, les biais liés au mode de sondage (présence d'un enquêteur qui pose des questions sur les convictions politiques du répondant), etc.

Outre les biais cités précédemment, les sondages traditionnels sont également soumis à une marge d'erreur. Autrement dit, la différence estimée entre les résultats obtenus lors d'un sondage sur un échantillon de la population étudiée et les résultats qui auraient été obtenus si l'ensemble de la population avait été questionné. La formule permettant de calculer cette marge d'erreur (Giezendanner, 2012), qui est principalement utilisée par les sondeurs, est la suivante :

$$e = \frac{t \times \sqrt{p \times (1-p)}}{\sqrt{n}}$$

e = marge d'erreur

t = coefficient de marge déduit du taux de confiance (1.96 est le coefficient le plus utilisé en sondage, ce qui indique un taux de confiance de 95% ; ces éléments nous renvoient aux statistiques, et plus particulièrement à la loi normale avec : Prob [-1,96 ≤ X ≤ 1,96] = 95%, où X représente une variable aléatoire normale (Gilbert et Malcorps, 2012))

p = proportion observée (50% est la proportion la plus utilisée en sondage, ce qui a pour effet de booster la marge d'erreur)

n = taille de l'échantillon étudié

Dans la pratique, si à la veille des élections, un sondage indique que 48% des électeurs se disent prêts à voter pour le candidat A (et 52% pour le candidat B) et que la marge d'erreur calculée est de ±4%, le candidat A serait susceptible d'être élu lors des votes réels du lendemain. En effet, le sondeur, en indiquant la marge d'erreur, précise que les résultats du candidat A se situeraient plus que probablement (il en est sûr à 95%) entre 44% et 52% et que ceux du candidat B se situeraient entre 48% et 56%. Malheureusement, le temps médiatique ne permet pas d'apporter ces quelques précisions lorsque les informations sont présentées au grand public, ce qui a pour effet, lorsque les résultats obtenus diffèrent de ceux estimés, de susciter de nombreux débats.

Les débats concernant la marge d'erreur sont également sujets à discussion entre les statisticiens et les sondeurs. Pour les statisticiens la marge d'erreur évoquée lors des sondages n'existe pas, car elle ne peut être calculée via la méthode des quotas qui est une méthode non-probabiliste. En d'autres termes, le calcul que nous avons développé ci-

dessus ne peut être utilisé que pour un échantillon probabiliste (tous les individus faisant partie de l'univers de référence ont une chance d'être sélectionnés, ce qui n'est pas le cas pour les échantillons utilisés par les sondeurs) via par exemple la méthode aléatoire (Dehon, Droesbeke, Vermandele, 2015). D'ailleurs, milieu des années 70, Leslie Kish (cité par Blondiaux, 1997) dénonçait déjà sur une touche humoristique la méthode des quotas qui, selon lui, n'avait rien de scientifique. Portelli et Sueur (2010) évoquent quant à eux dans un rapport soumis au Sénat français que les marges d'erreur des méthodes par quotas et échantillon aléatoire sont relativement similaires, et qu'il n'y a donc aucun souci à publier les marges d'erreur via la méthode des quotas. Au vu de notre courte expérience, nous estimons toutefois que, au-delà du débat sur la marge d'erreur, les données issues des sondages (s'ils ont été correctement réalisés) présentent à la société des informations à prendre en considération.

III. Big Data

1. Origine du Big Data

De même que la terminologie « Big Data », comme nous le verrons, est loin de faire consensus auprès des experts (Ollion et Boelaert, 2015), l'origine de ce récent phénomène numérique diffère d'un expert à l'autre. Pour certains, il faudrait remonter à l'Antiquité, passer par le Moyen Âge et l'époque Moderne pour enfin atterrir à notre époque. Mais à l'instar de Rob Kitchin, une sommité de l'univers du Big Data, nous considérons qu'il faille poser des limites généalogiques à cette histoire. Le concept du Big Data doit impérativement comporter un certain nombre de caractéristiques. Sans elles, il n'y a donc pas lieu de parler de Big Data (cité par Schafer, 2017, p.23). Les articles qui mentionnent une origine pré-contemporaine ne tiennent pas compte des composants intrinsèques de cette nouvelle technologie dont l'histoire est finalement bien plus courte. À titre de comparaison, lorsque nous recherchons l'origine de l'automobile, nous ne souhaitons pas remonter à l'époque des chars romains. Nous nous arrêtons plutôt à l'époque où les chevaux étaient dans le capot, et non devant celui-ci.

Si l'histoire du Big Data, où en tout cas de sa terminologie, remonte au milieu des années 1990 (Kitchin et McArdle, 2016), nous noterons qu'avant cette période, depuis 1944, beaucoup d'articles ont fait mention de l'accroissement du nombre d'informations générées par la société. Cet intérêt pour cette masse d'informations a entraîné la communauté scientifique et celles d'autres spécialistes à imaginer quelle serait la quantité de données produite et récoltée dans le futur. Une récolte massive de données qui suscitait la réflexion en matière de moyens pour stocker ces masses hétérogènes de données (Press, 2013).

En 1990, Peter J. Denning dans son article intitulé « Saving All the Bits » tracera les contours de ce qui plus tard se nommera le Big Data. Dans cette courte publication, il mentionnera qu'imposer à la communauté scientifique la sauvegarde de l'ensemble des données produites entraîne celle-ci dans une impasse. Le déluge d'informations étant tel qu'il inonderait l'entièreté du système d'information. À cette impossibilité technique, il

rassurait la communauté scientifique en affirmant qu'il n'était pas impensable d'élaborer des outils informatiques pouvant suivre l'afflux de données en temps réel. Il envisageait également que ces outils pourraient identifier certains linéaments propres à ces masses de données (Denning, 1990).

La première publication faisant apparaître le terme « Big Data » date de 1997. Toutefois, cette dernière n'est pas celle qui a été retenue comme officielle. En effet, John Mashey aurait lors de diverses interventions publiques dans le courant des années 1990 utilisé cette terminologie (Dontha, 2017). Preuve en est, la mise en ligne en 1998 (que nous retiendrons comme étant la date officiellement reconnue de cette terminologie) d'une série de diapositives avec pour titre « Big Data and the Next Wave of Infrastress » (Lohr, 2013).

En 2000, Diebold, qui pensait être à l'origine du terme Big Data, se lancera dans une définition relativement proche (tout en restant assez éloigné dans un certain sens) de l'image que nous nous en faisons aujourd'hui (Dontha, 2017). Il présentera ce concept comme une immense masse de données à notre disposition qui, le cas échéant, pouvait être de très bonne facture, et dont la récolte était rendue possible par les nombreuses avancées technologiques (Diebold, 2000). Un an plus tard, en 2001, Laney décrira les trois composants majeurs du Big Data (sans pour autant le nommer) : les « 3-V », à savoir le volume, la vélocité et la variété des données (Laney, 2001). Ces composants sont encore acceptés et employés aujourd'hui par les spécialistes pour faire référence aux éléments intrinsèques du Big Data (Press, 2013).

Enfin, quelques années plus tard, en 2005 plus exactement, Tim O'Reilly publiera un article qui sonnera le départ de la course aux données (cité par Dontha, 2017). Suivront ensuite une pléthore d'articles et d'études tournant autour de la thématique du Big Data. Ceux-ci s'attarderont sur le phénomène de manière générale, en posant des questions sur les possibilités actuelles et futures de cette technologie. D'autres aborderont la question plus délicate de l'éthique des données ou bien encore essayeront d'identifier les éléments qui composent, et dès lors autorisent l'appellation Big Data (Press, 2013).

2. Typologie des données

Avant de poursuivre cette immersion dans l'univers du Big Data, il nous semble judicieux de marquer l'arrêt sur un élément essentiel qui compose cette terminologie, à savoir : « Data ». En français, ce terme qui signifie « données » (au pluriel !) ne doit pas être confondu avec « l'information ». Cette dernière, comme nous l'expliquent Laudon, Laudon, Fimbel, Costa et Canevet-Lehoux (2013), étant composée d'un ensemble de données « présentées sous une forme utile et utilisable par les personnes » (Laudon et al., 2013, p.22). Les données, quant à elles, représentent « des valeurs à l'état brut correspondant à des évènements qui ont lieu dans ou en dehors de l'organisation » (Laudon et al., 2013, p.22). À la différence de l'information, les données ne peuvent pas toujours être comprises par les individus, et ne présentent pas systématiquement d'utilité pour ces derniers (Laudon et al., 2013).

Dans son sens traditionnel, le terme « Data » (ou données), nous rappelle Kitchin (2014), est souvent compris comme un ensemble d'éléments qui sont issus de l'observation, d'études, d'enregistrements, etc. Des données que nous serions venus soustraire de leur environnement afin qu'en découle de l'information. Toutefois, si nous revenions quelque peu sur nos pas, nous constaterions stupéfaits que l'emploi du mot « Data » n'était en aucun cas le plus approprié. « Data » est issu du mot latin « dare » qui signifie « donner » (Rosenberg, 2013). « Donner » étant l'opposé de « prendre », ce qui est le cas pour le Big Data, il aurait fallu utiliser le terme « Capta », insiste Kitchin (2014), qui en latin signifie « prendre ». Dès lors, cette récente technologie qui fait tant parler d'elle aurait dû prendre la dénomination de « Big Capta » (même si cela sous-entend que les données sont toujours prises de façon active, alors que dans la réalité les activités humaines laissent des traces numériques qui permettent à cette technologie de traiter celles-ci). Une nouvelle terminologie donc qui, malgré ses défauts, aurait pu être plus bénéfique pour cette technologie que ne l'est celle actuellement utilisée. Le Big Data, en raison de sa formulation, reste pour le moins un terme tronqué et trompeur. Il ne serait pas improbable qu'en demandant aux premiers venus quels mots leur viennent à l'esprit lorsque le mot « Big Data » est prononcé, que ces derniers répondent « beaucoup de données », et s'en contenteraient. Alors que comme nous l'établirons (cf. infra p.26), le Big Data est loin de posséder une définition homologuée et reconnue par l'ensemble du monde scientifique, ni celui des experts (Ollion et Boelaert, 2015).

Bien évidemment, l'objet de notre étude étant tout autre, il ne sera pas question d'entrer ici dans une quelconque polémique étymologique quant à la désignation de ce concept. Nous nous contenterons, à l'instar de Rob Kitchin (2014), d'utiliser le terme « Data » pour parler des données. Des données qu'il nous reste encore à définir, car celles-ci ne sont pas toutes identiques. Elles se distinguent notamment par leur forme, leur structure, leur source, leur producteur et leur type (Kitchin, 2014). Les points suivants expliquent brièvement ces différentes notions afin de nous permettre de mieux entamer la partie concernant le Big Data.

2.1. Formes de données

Du point de vue de leur forme, les données s'insèrent dans deux catégories bien distinctes. Ces données peuvent être soit de type quantitatif, soit de type qualitatif. Les données quantitatives sont des enregistrements numériques. Ceux-ci comportent des données faisant référence à des propriétés physiques (taille, poids…) ou des propriétés non-physiques (le niveau d'éducation, le type d'emploi…). Aux côtés de ces données dites quantitatives se trouvent les données de type qualitatif. Ces dernières représentent toutes les données non-numériques comme des images, l'art ou des vidéos (Kitchin, 2014).

À la différence des données quantitatives, les données qualitatives comportent bien plus de richesse en matière d'analyse. Elles peuvent toutefois être transformées en données quantitatives aux dépens de ce qui fait leur valeur (Kitchin, 2014). Kitchin (2014) rappelle d'ailleurs que, dans la pratique, l'analyse de données qualitatives se fait surtout sur le matériel originel. Cette distinction entre quantitatif et qualitatif nous renvoie également

aux sondages traditionnels qui observent les mêmes distinctions. Les sondages qualitatifs permettant d'explorer bien plus en profondeur la pensée des personnes interrogées, mais sur un nombre limité d'individus. Alors que les sondages quantitatifs récoltent auprès de nombreux individus (1000, 2000, 3000…) des réponses qui, malgré leur pertinence, sont moins riches dans leurs apports.

2.2. Structures de données

Dans son ouvrage, Kitchin (2014) répertorie trois types de structures de données. Celles qui octroient le plus de facilité en matière de stockage de données, de traitement, d'analyse (et autres) seront considérées comme des données structurées. Cette première structure dont le modèle est défini à l'avance comporte des données assez basiques comme du texte ou des chiffres qui peuvent très facilement être mis en relation (par exemple : nombre d'hommes ayant tel âge et habitant telle région).

À mi-chemin, nous retrouvons les données semi-structurées qui, à la différence des données structurées, ne possèdent pas de modèle prédéfini. Cette configuration empêche l'intégration de ces données dans une base de données relationnelles, en tout cas pas aussi directement que la première structure. Néanmoins, ces données disposent de certaines caractéristiques communes qui, malgré les irrégularités de leur structure, vont permettre d'être traitées plus facilement que des données non-structurées. Il s'agit par exemple des données personnelles (âge, métier…) disponibles sur le web, mais dont la structure est différente d'un site web à l'autre. Dès lors, cette structure n'autorise la récolte et la mise en commun de ces données qu'à l'aide de logiciels informatiques spécialisés (Kitchin, 2014).

En dernier lieu, nous retrouvons les données non-structurées qui restent les plus compliquées à gérer en termes de récolte, d'analyse, etc. Celles-ci ne possèdent pas d'identificateurs clés leur permettant d'être regroupées avec d'autres données, et encore moins de modèle prédéfini. Ces données sont de manière générale de type qualitatif et, comme c'est quelque fois le cas dans les sondages qualitatifs, sont catégorisées et classifiées aux dépens d'une perte d'information (Kitchin, 2014).

2.3. Sources de données

« Captured, exhaust, transient and derived data » (Kitchin, 2014, p.6) représentent les quatre différentes sources de données. Dans l'ordre de citation, nous avons les captured data, soit des données qui sont directement extraites de l'environnement (par exemple : la récolte de données provenant d'un questionnaire auto-administré). Ensuite, les exhaust data qui sont des sous-produits des captured data. Pour illustrer pratiquement la différence entre ces deux sources, nous pouvons prendre le cas d'achat d'articles sur le site XYZ. La fonction première du panier d'achat virtuel est d'additionner l'ensemble de vos articles et de vous inviter à payer le montant total de vos achats (= captured data ; soit une information primaire). Entre-temps le système mis en place par XYZ fera en sorte d'extraire d'autres analyses de votre expérience d'achat, en regardant si, par exemple, vous avez retiré des produits de votre panier, ou si les stocks pour les produits que vous avez commandés sont épuisés, auquel cas il faudrait renouveler ces stocks (= exhaust

data ; soit une information secondaire, car elle découle des achats réalisés) (Kitchin, 2014).

En troisième lieu, nous retrouvons les transient data qui représentent l'ensemble des données en transition. En raison d'un certain nombre de facteurs (volume de données, structure des données, coûts de traitement…), ces données s'évaporent aussi vite qu'elles n'apparaissent. Leur intérêt n'est palpable qu'au moment de leur affichage (par exemple : une montre sportive qui calcule seconde après seconde votre rythme cardiaque. Nul n'est besoin de garder l'ensemble des données générées, car seule la moyenne sur une période précise intéresse l'utilisateur). Pour terminer, certaines données doivent avant d'être produites faire l'objet d'analyses et de traitements supplémentaires, qui sont effectués sur des données brutes, soit sur des captured data. Il s'agit donc là de données générées à partir d'autres données, soit des derived data (Kitchin, 2014).

2.4. Producteurs de données

Les chercheurs ont à leur disposition trois types de données : des données primaires, secondaires et tertiaires. Les données primaires sont quelque peu similaires aux données brutes. Les chercheurs à ce stade utilisent des outils (par exemple : un formulaire) pour générer des données que l'on qualifiera de primaires. Une fois ces données récoltées, ce premier groupe de chercheurs peut transmettre ces données à d'autres chercheurs. Pour ce second groupe de chercheurs, les données transmises seront considérées comme secondaires, car celles-ci n'auront pas été directement générées par ces derniers. Enfin, les données qualifiées de tertiaires sont celles produites, entre autres, par des bureaux statistiques. Ce type d'agences s'adonne à la récolte de données primaires et secondaires qu'il communique de façon partielle en n'omettant pas de préserver la confidentialité des données. En pratique, il s'agit par exemple de données secondaires confidentielles (issues d'un recensement) pour lesquelles la diffusion sous leur forme complète est interdite. Intervient alors une agence statistique qui récupère ces données, les compile et ensuite les diffuse sous un format confidentiel, c'est-à-dire en intégrant les données dans des catégories distinctes (exemple : nul ne connait que tel individu habite à tel endroit, mais il est possible d'avoir des estimations du nombre d'individus de sexe masculin habitant dans une commune) (Kitchin, 2014).

2.5. Types de données

Dernière distinction que nous pouvons effectuer sur les données : la classification selon le type de données. Kitchin (2014) indique que l'on répertorie les données selon trois types. Le premier type concerne les données indexées, c'est-à-dire celles qui possèdent un identifiant unique qui est lié à l'un ou l'autre individu (par exemple : le numéro de registre national qui est unique pour chaque citoyen belge). Le second type concerne les données dites attribuées. Celles-ci décrivent certaines caractéristiques (âge, sexe…) qui seront associées aux données indexées (par exemple : lieu d'habitation indiqué sur un passeport). Enfin, les métadonnées représentent le dernier type de données. Il s'agit de données qui sont générées afin d'informer sur d'autres données (par exemple : un selfie pris au pied de la Tour Eiffel représenterait une donnée, à cette donnée l'appareil utilisé

pour effectuer la photo ajoutera la date à laquelle cette photo a été prise, soit une donnée qui renvoie vers une donnée) (Kitchin, 2014).

3. Définition du Big Data

Alors qu'il y a quelques années le Big Data ne représentait rien dans l'esprit du citoyen lambda, aujourd'hui, il n'est pas rare de retrouver cette terminologie dans les journaux, à la radio ou bien encore à la télévision. D'ailleurs, des Salons lui sont entièrement dédiés, ce qui démontre en quelque sorte un certain engouement pour ce phénomène, ou du moins un questionnement pour les plus dubitatifs (Ollion et Boelaert, 2015).

La question qui convient d'être posée à ce stade de notre étude est bien entendu celle qui interroge sur la signification même de cette nouvelle technologie. Soit une définition de ce nouveau concept qui empêcherait tout amalgame, toute confusion dans nos esprits que nous soyons directement ou indirectement concernés par le Big Data. Malheureusement, c'est ici que les choses se corsent, car à ce jour il n'y a toujours pas de définition concise et acceptée de tous (Kitchin et McArdle, 2016).

À l'origine, comme nous l'indique Kitchin et McArdle (2016), trois caractéristiques venaient décrire le Big Data, à savoir : le volume de données, la vélocité (soit la capacité à gérer des données en temps réel) et la variété des données (qui pouvaient être de type structuré, semi-structuré ou non structuré). Mais depuis lors, d'autres spécialistes sont venus proposer des caractéristiques supplémentaires à cette technologie, comme par exemple l'exhaustivité, la granularité, l'extensivité et autres. Ces derniers traits sont bien moins connus que les trois composants originels (volume, vélocité, variété), mais semblent relativement importants au regard de Kitchin et McArdle (2016). Car pour ces experts, le volume des données et leur variété n'est pas forcément synonyme de Big Data, alors qu'au contraire la vélocité et l'exhaustivité seraient les caractéristiques les plus importantes de cette technologie. Par ailleurs, ces spécialistes nous apprennent également que le Big Data est malencontreusement utilisé comme un mot « fourre-tout » (Kitchin et McArdle, 2016, p.9). D'aucuns considérant que tout ce que l'on désigne par Big Data comporterait des caractéristiques parfaitement identiques. Autrement dit, en émettant l'hypothèse que, pour ces personnes, le Big Data ne serait représenté que par les 3-V originels (cf. supra p.21), tous les Big Data possèderaient forcément ces trois caractéristiques principales. Kitchin et McArdle (2016) ne sont pas du même avis. Après avoir effectué une analyse sur plus d'une vingtaine de données de types différents, ils en arrivent à la conclusion qu'aucun des types de données analysés ne possédait les mêmes traits caractéristiques qu'un autre. Il y a donc une pléthore de Big Data dont chacun serait caractérisé par un certain nombre de traits qui lui sont propres, et non pas une seule et unique forme de Big Data qui, dans chaque domaine d'utilisation, contiendrait les mêmes caractéristiques (Kitchin et McArdle, 2016).

Dans le cadre de cette tentative de définition du Big Data, nous avons opté pour celle proposée par Rob Kitchin (2014) dans son livre « The Data Revolution ». Celle-ci, tout en partant des trois caractéristiques originelles, ajoute 4 composants supplémentaires pour délimiter les contours du Big Data. Cette définition plus exhaustive que celle largement

reconnue nous paraît bien plus intéressante, car elle permet de montrer un outil qui serait bien plus impactant sur l'industrie du sondage. En effet, en ne considérant que les trois premiers composants, nous pourrions très facilement critiquer les apports du Big Data au sondage traditionnel. Par exemple, si nous ne tenons compte que du volume de données et non pas, comme le préfèrent Kitchin et McArdle (2016), de l'exhaustivité, nous pourrions établir assez rapidement un parallèle avec la première ère des sondages (cf. supra p.7) où Gallup avec un échantillon bien plus restreint que celui de la revue Literary Digest avait réussi à identifier qui serait le futur président des États-Unis. Dès lors, le volume de données, quand bien même serait-il très impressionnant (ex. : des millions de données), ne nous permettrait pas de nous s'assurer qu'il reflète bien la réalité de l'univers de référence étudié. L'exhaustivité, comme l'affirment Kitchin et McArdle (2016), nous semble bien plus adéquate pour rivaliser avec les sondages traditionnels.

Les points qui suivent abordent chacun des 7 composants du Big Data tels que les présente Rob Kitchin.

3.1. Le volume des données

La première idée qui vient à l'esprit lorsque l'on parle du Big Data, c'est le volume des données. En d'autres termes, le Big Data devrait être en mesure d'exploiter des masses considérables de données. Rob Kitchin (2014) évoque qu'il s'agirait au minimum de térabyte, soit approximativement 1.000 gigabytes de données (pour donner un ordre d'idée de ce que représente concrètement un térabyte et les autres unités de mesure, nous renvoyons le lecteur à la page suivante).

La création massive de données, nous expliquent Bidoit-Tollu et Doucet (2017), a été rendue possible par le développement d'Internet, les smartphones, les objets connectés, etc. Toutes ces technologies stockent continuellement, à chaque instant, des quantités astronomiques de données à travers le monde, qui proviennent de sources diverses et variées (réseaux sociaux, signaux GPS, photos, vidéos…) (Bidoit-Tollu et Doucet, 2017). Alors que le nombre de données générées de toutes parts a connu une croissance sans pareil, il y a lieu, selon Zikopoulos, Eaton, de Roos, Deutsch et Lapis (cités par Kitchin, 2014, p.70), d'espérer que le volume total de données générées atteigne les 35 zettabytes en 2020, soit près de 30 fois le volume total de données présent en 2010 (cf. figure 2 infra, p.27).

Data inflation 2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Figure 2 : Les unités de mesure

Source : The Economist. (2010). All too much. Monstrous amounts of data. *The Economist.* Récupéré de https://www.economist.com/node/15557421

3.2. L'exhaustivité

Quand bien même le volume de données contenu dans le Big Data peut être extrêmement élevé, il n'est pas pour autant exhaustif. Cette exhaustivité, seconde caractéristique du Big Data, représenterait la possibilité pour cette nouvelle technologie de ne plus se contenter d'un échantillon de données (quelle que soit sa taille), mais de l'entièreté de la population étudiée, soit en langage statistique « N=n » (Kitchin, 2014). Concrètement, si l'univers de référence étudié (par exemple, dans un sondage politique) présente une population totale de 5 millions d'individus (=N), le Big Data ne se limiterait pas, comme le font les instituts de sondages actuellement, d'un petit échantillon représentatif de 2500 personnes (=n) ou d'un échantillon de 2 millions d'individus (=volume de données), mais bien de tout l'univers de référence (= 5 millions).

Une telle exhaustivité semble pour le moins illusoire, mais Mayer-Schonberger et Cukier (cités par Kitchin, 2014, p 72) nous font comprendre que par exhaustivité, il faut entendre des tailles d'échantillon bien plus grandes que celles fournies par l'intermédiaire du Small Data. Cette dernière appellation fait référence à des volumes de données dont la taille peut être relativement grande, mais dont l'exhaustivité reste au stade embryonnaire (cf. tableau infra p.29) (Kitchin, 2014).

Cette course à l'exhaustivité part du principe qu'au plus il est possible d'emmagasiner des données, au mieux cela permet aux chercheurs d'obtenir une vision plus fidèle de la réalité. Toutefois, vouloir à tout prix détenir l'intégralité des données de l'univers de référence étudié pose question. Notamment, outre celle des coûts des infrastructures, celle

concernant l'éthique dans la préservation et l'utilisation des données. Cette question sera abordée plus largement dans un prochain point (cf. infra p.50).

	Small data	Big Data
Volume	Limited to large	Very large
Velocity	Slow, freeze-framed/ bundled	Fast, continuous
Variety	Limited to wide	Wide
Exhaustivity	Samples	Entire populations
Resolution and indexicality	Course and weak to tight and strong	Tight and strong
Relationality	Weak to strong	Strong
Extensionality and scalability	Low to middling	High

Figure 3 : Comparaison entre Small Data et Big Data

Source : Kitchin, R. et McArdle, G. (2016). *What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets.* Récupéré de http://journals.sagepub.com/doi/pdf/10.1177/2053951716631130

3.3. La granularité

À côté de cet idéal d'exhaustivité des données se trouve celui de leur granularité, soit un niveau de détail à l'unité près. Plus spécifiquement, le Big Data serait en mesure de traiter chaque donnée de manière distincte, en rendant chaque donnée totalement unique et identifiable. Pour illustrer ces propos, nous pouvons prendre l'exemple d'une gamme de barres chocolatées qui, actuellement, n'est identifiée que par un code barre unique, mais identique pour tous les chocolats de cette gamme. Avec la technologie Big Data, il serait très aisé d'octroyer à chacune de ces barres chocolatées de la même gamme un identifiant unique, sans aucune contrainte de traitement (Kitchin, 2014).

Cette identification à l'unité près, nous explique Kitchin (2014), pourrait permettre à l'entreprise de suivre à la trace l'intégralité du chemin parcouru par un produit d'une gamme. Depuis sa conception et sa mise en emballage jusqu'aux mains de la clientèle, la chaîne de distribution ne devrait plus avoir de secret pour ses gestionnaires. Une technologie qui devrait s'avérer utile pour la détection des problèmes issus de leur chaîne logistique. Outre ces possibilités au sein du monde commercial, Kitchin (2014) évoque le fait que la granularité pourrait également permettre aux gouvernements d'étudier des zones bien plus spécifiques lors de recensements. Ils pourraient ainsi réaliser des analyses sur ces recensements, non plus à l'échelle de villes ou de villages, mais de rues. Ce qui devrait sans doute faciliter la tâche des gouvernements quant aux problèmes d'ordre social, économique (ou autres) rencontrés par les citoyens vivant dans une zone bien spécifique (Kitchin, 2014). Encore une fois, ce type de pratique pose question et suscite de réelles craintes.

3.4. Le degré de relation entre les données

Cette quatrième composante du Big Data mesure le degré de relation entre les différentes données. Il s'agit donc d'évaluer à quel point des données d'origine différente peuvent être mises en relation pour répondre à des questions nouvelles. Des questions qui sans ce type d'approche n'auraient que (très) difficilement pu trouver une suite favorable ou n'auraient même pas traversé l'esprit des chercheurs. Cette caractéristique du Big Data représente l'un des éléments essentiels de cette technologie, car elle permet d'extraire des informations pertinentes sur des sujets d'étude. Le Big Data permet en effet de croiser des données personnelles (caractéristiques propres à l'individu : âge, sexe, métier…) avec d'autres concernant par exemple les achats de l'individu, ses interactions sociales, ses déplacements dans le temps et l'espace... En résumé, une sorte de pistage complet du sujet d'étude qui nous informerait de manière plus explicite sur les raisons qui poussent tel ou tel individu à adopter un type de comportement en particulier (Kitchin, 2014).

Cette capacité de mise en relation que possède le Big Data va bien au-delà des possibilités procurées jusqu'alors par des outils traditionnels. Kitchin (2014) nous explique d'ailleurs que cette technologie ouvre de nombreuses perspectives et rendrait tout à fait concevable l'interrelation entre différents types de données, et ce, quelle que soit leur structure. Pour démontrer quelque peu le potentiel du Big Data en matière de mise en relation, nous pouvons porter un regard sur les deux campagnes successives menées par Obama lors des élections présidentielles de 2008 et 2012. Durant ces campagnes, il était question de relever l'ensemble des informations issues du terrain (intentions de vote des individus, données gouvernementales…), des données provenant de data brokers (dont l'objet social est la commercialisation de nos données (Roux, 2017)), des activités menées par les membres de l'équipe de campagne du futur président (campagnes d'affichages menées dans telle ou telle zone, rassemblements, conférences…), ainsi que pléthore d'autres données venant de tous horizons. Au total, par individu, près de 80 variables étaient répertoriées et analysées par le Big Data. Ce qui a permis aux équipes de campagne d'établir avec plus de précision quelles étaient les personnes susceptibles de voter pour Obama, celles qui hésitaient encore et celles qui ne souhaitaient pour rien au monde voter pour ce candidat. Une fois en possession de ces informations cruciales, le discours pour ces différents profils était adapté à ces derniers pour, entre autres, répondre aux différents questionnements que ceux-ci pourraient avoir sur la politique du candidat à la Maison Blanche. Une démarche qui coup sur coup s'est avérée être un véritable succès pour ce candidat qui remporta par deux fois l'élection présidentielle. Sans le Big Data, la mise en relation de toutes ces données n'aurait pu être envisagée, car il s'agissait ici de masses astronomiques de données dont les origines étaient multiples (Kitchin, 2014).

3.5. La vélocité

Si le Big Data attise tant de convoitises, il est fort probable que la vélocité y soit pour quelque chose. Cette caractéristique peut être définie comme étant la capacité pour le Big Data d'enregistrer des données en temps (quasi) réel. C'est ce qui la distingue notamment du Small Data (cf. tableau supra p.29), qui n'est qu'en mesure de récolter des données de

manière ponctuelle, soit à un temps précis et dans un espace défini (exemple : un recensement réalisé par un gouvernement toutes les x années) (Kitchin, 2014).

La vélocité peut facilement être observée dans notre quotidien. On retrouve de nombreux systèmes qui enregistrent continuellement, dans les sites web commerciaux, les passages sur le site web, le nombre de clics, les produits mis sur un panier d'achats, etc. Dans d'autres domaines, les équipements médicaux qui suivent en direct le rythme cardiaque, ou bien encore, les magasins physiques qui pour améliorer leur logistique utilisent des outils leur permettant de suivre à la trace l'évolution des ventes, ce qui leur permet d'estimer avec plus de précision le meilleur moment pour réapprovisionner leur stock. La promesse du Big Data suit le même fil conducteur que les exemples précités, à la seule différence près qu'il s'agit ici de reproduire cet enregistrement en temps réel pour des masses de données provenant de sources différentes. Techniquement le défi est de taille, car il faut sans cesse emmagasiner de l'information (qui peut causer des problèmes de surchauffe) dans des proportions surdimensionnées (ce qui peut créer des saturations au niveau du système d'enregistrement) (Kitchin,2014).

3.6. La variété

Autre richesse du Big Data, la variété des données qui sont récoltées. Cette variété est marquée tant par le format de ces données (images, textes, vidéos…) que la source de celles-ci (réseaux sociaux, recensements…). En termes de diversité, le Small Data reste pour le moins similaire au Big Data (cf. tableau supra p. 29). Les deux technologies pouvant collecter de grandes quantités de données. La différence réside surtout dans la capacité du Big Data à pouvoir mettre en relation cette diversité de données que celles-ci soient ou non structurées (Kitchin, 2014). Là où le Small Data éprouverait bien plus de difficultés, voire se trouverait dans l'impasse la plus totale.

3.7. La flexibilité

Dernier critère que nous évoquerons dans le cadre de cette définition du Big Data, et non des moindres : la flexibilité qu'offre cette technologie. Selon Kitchin (2014), le Big Data peut assez facilement s'accommoder des changements liés à son environnement, en ce sens qu'une modification ou un ajout d'élément nouveau à une base de données ne constituent pas d'obstacle à proprement parler, contrairement aux outils traditionnels qui font davantage montre de rigidité. Un exemple patent : celui du recensement. Ce dernier se compose d'un certain nombre de questions qui seront posées à une population définie. Une fois les questionnaires distribués, il ne sera plus possible d'ajouter une quelconque question, ce qui en fait un modèle très rigide (Kitchin, 2014).

En outre, la flexibilité du Big Data se caractérise également par sa capacité à évoluer en s'adaptant aux nouvelles générations de données, ainsi qu'aux hausses de la demande de données. À la différence des outils traditionnels dont la capacité d'enregistrement limitée empêche toute perspective d'évolution significative (Kitchin, 2014).

4. Big Data, mode d'emploi

Outre l'attrait pour les promesses théoriques du Big Data, de nombreux facteurs sont à l'origine du développement de cette technologie. Les infrastructures technologiques par le biais des avancées en informatique (matériels et logiciels) représentent l'un des facteurs clés. Ces outils ne sont cependant pas les seuls. La possibilité de stocker, d'analyser et de traiter des quantités astronomiques de données ne saurait être possible si aucune source (où des données n'attendraient qu'à être extraites) n'était disponible pour ces technologies. Afin de mieux appréhender l'univers du Big Data, nous proposons de parcourir son mode d'emploi dans les points qui suivent ; cette démarche ne consistera pas à rentrer dans les détails, mais à donner au lecteur un aperçu des infrastructures et sources qui ont rendu possible le développement du Big Data.

4.1. Les infrastructures technologiques du Big Data

Selon Laudon et al. (2013), les infrastructures technologiques peuvent être conçues au sens premier, soit un ensemble d'équipements qui doit permettre l'utilisation « des technologies de l'information et de l'informatique » (Gouvernement du Québec, 2018), soit dans un sens plus large, à savoir « un ensemble de services offerts à l'entreprise, budgétisés par les managers et incluant les ressources humaines et techniques » (Laudon et al., 2013, p.166). Les auteurs préfèrent le second sens au premier car celui-ci englobe toute une série de services autorisant une meilleure grille de lecture des apports de la technologie envers l'entreprise. Ces services comportent, entre autres, les plateformes technologiques (matériels informatiques), les services de télécommunication, de gestion de données, des logiciels, etc. (Laudon et al., 2013). Dans le cadre de notre étude, c'est ce second sens qui nous semble le plus à même d'identifier les outils technologiques susceptibles de faciliter la tâche du Big Data. En effet, nous ne pouvons nous cantonner à une définition qui ne concevrait le Big Data qu'en termes de matériels informatiques permettant de traiter des données. Une des étapes essentielles du Big Data est la récolte de données provenant de multiples sources, dont le réseau informatique.

a. Les ordinateurs

Depuis la fin de la Seconde Guerre Mondiale, le monde de l'informatique n'a cessé d'évoluer. Ce qui nous permet aujourd'hui de disposer d'ordinateurs fixes, portables, de smartphones dont la puissance dépasse de loin leurs prédécesseurs. Outre cette avancée de l'outil en lui-même, le coût de ces nouvelles technologies a décru dans des proportions considérables. Cette diminution de coût représente dès lors un avantage conséquent, car elle facilite et encourage la recherche et l'investissement dans des technologies devant permettre l'utilisation du Big Data (Kitchin, 2014). Cependant, malgré les avancées technologiques et ce qu'elles permettent en matière de captation et de traitement de l'information, Fanet et Duranton (2017) estiment que la technologie dont nous disposons aujourd'hui serait loin de couvrir l'intégralité des exigences du Big Data.

b. Réseau interconnecté

Autre outil indispensable du Big Data est bien entendu l'Internet. Sans ce dernier, il n'aurait pas été possible de transmettre aisément des quantités d'information à très haut débit et à des prix raisonnables. Ce mode de communication conçu à l'origine pour servir des intérêts militaires relie aujourd'hui plusieurs millions de réseaux à travers le monde, grâce notamment à sa démocratisation via la création du web (Laudon et al., 2013). Outre cette expansion de la connectivité, la vitesse de transfert de données, soit de la bande passante, devrait doubler tous les six mois d'après George Gilder (cité par Kitchin, 2014, p.82). Cette amélioration a, en toute logique, eu des répercussions positives au sein de la technologie Big Data qui, comme nous l'avons évoqué plus tôt (cf. supra p.30), se doit pour être considérée comme telle de disposer d'une très grande vélocité.

c. Internet of things

Par l'intermédiaire de capteurs intégrés dans les appareils digitaux que nous utilisons quotidiennement (smartphones, montres, réveille-matin…), d'énormes quantités de données sont créées, transférées, stockées pour ensuite être traitées. Ensemble, ces données contribuent à la concrétisation du Big Data. Elles permettent entre autres dans un environnement logistique de suivre en temps réel l'évolution et la position de l'intégralité du stock dont dispose une société. L'Internet of things se retrouve également dans la domotique, comme par exemple dans des régulateurs de température, des compteurs, des détecteurs, etc. (Futura, 2018). Pour qualifier cette nouvelle ère où les objets connectés sont reliés à Internet, Kitchin (2014, pp.83-84) parle d'« omniprésence et d'ubiquité » [traduction libre] des appareils digitaux. L'omniprésence désignant le fait que les objets possèdent des capteurs qui les connectent à un réseau de manière directe ou indirecte, alors que le don d'ubiquité indique que ces objets autorisent la traçabilité de l'appareil digital en question en tout lieu et à tout moment (Kitchin, 2014).

d. Indexation des appareils

Outre l'invasion des objets connectés (dont il devient parfois très difficile de se défaire), l'infrastructure technologique permettrait aujourd'hui d'identifier chacun de ces objets avec une très grande précision (Kitchin, 2014). Selon Gershenfeld, Krikorian et Cohen (2004), l'Internet of things utilise des identifiants uniques qui sont attribués à un seul et unique appareil digital. En pratique, le smartphone modèle 10S de la marque XYZ ne possède pas le même identifiant que celui de la marque ABC, et ne possède également pas le même identifiant que celui de sa propre marque et modèle (modèle 10S – marque XYZ dans cet exemple). Cette précision dans l'identification de l'appareil utilisé permet d'obtenir une meilleure granularité dans la récolte de données et permet également de croiser les données récoltées avec d'autres pour obtenir des informations plus détaillées de l'univers étudié. De plus, cette récolte massive de données se fait automatiquement. Il suffit que l'objet en question soit directement ou indirectement relié à Internet pour que la transmission de données se fasse (Kitchin, 2014). Pour exemple, le signal GPS d'un véhicule donne à quelques mètres près la position du véhicule (transmission directe de données), alors qu'une montre connectée transmettra la fréquence cardiaque, la vitesse de

course, etc. au moment où celle-ci sera connectée à un ordinateur (connecté à Internet) en vue de mettre à jour les statistiques personnelles (transmission indirecte de données).

e. Stockage de données

La capacité de stockage de données a connu également une forte croissance au cours de ces dernières années. Du système basique qui consistait à encoder les données dans des cartes perforées aux disques durs, le mode de stockage n'a rien à envier aux évolutions qui se sont opérées au sein des autres composants des ordinateurs (cartes mères, cartes graphiques, processeurs…). Ici, également, la diminution du coût de stockage permet à toute entreprise de préserver des masses de données (quelle que soit leur structure) pour des montants relativement faibles (Kitchin, 2014).

Les besoins en collecte de données se sont tellement accrus que bon nombre de centres de données (également connus sous leur appellation anglophone de datacenters) pullulent à travers le monde (Caulier, 2015). Les datacenters présentent aux côtés de leurs baies de stockages d'autres équipements tels que des serveurs et des composants réseaux. Selon certains spécialistes, l'usage de datacenters serait plus bénéfique pour les sociétés qui en feraient appel car ceux-ci donnent aux sociétés l'occasion de diminuer significativement leurs coûts, et ce, tout en améliorant leurs performances (Daydé et Veynante, 2017).

Outre les datacenters, qui sont des infrastructures physiques (Angeles, 2013), le cloud computing permet également le stockage et le traitement d'un nombre très volumineux de données. Cette technologie consiste en une virtualisation de serveurs se démultipliant pour donner l'impression aux utilisateurs qu'ils possèdent tous un ordinateur différent. De plus, cette récente infrastructure est en mesure d'utiliser les « capacités de traitement ou de stockage provenant de plusieurs machines physiques différentes » (Laudon et al., 2013, p.189). Cette capacité d'utiliser simultanément plusieurs ressources dans le cadre d'une quelconque opération permet un véritable gain d'efficacité dans la gestion de données selon Farber, Cameron, Ellis et Sullivan (cités par Kitchin, 2014, p.86). En pratique, ces serveurs virtuels permettent l'exploitation de trois types de services en particulier (Laudon et al., 2013) :

1. Infrastructure as a Service (IaaS) : soit l'infrastructure virtuelle qui donne accès à un service de stockage, de serveurs et de réseaux (Kitchin, 2014),
2. Platform as a Service (Paas) : soit une plateforme qui permet aux informaticiens de se doter d'outils pour développer des applications (Laudon et al., 2013),
3. Software as a Service (SaaS) : soit des logiciels destinés aux entreprises mais également aux utilisateurs lambda (Laudon et al., 2013).

Les données enregistrées dans ces univers virtuels sont réparties à différents endroits du monde virtuel dans lequel ils se situent, ce qui rend impossible l'identification précise de la position de ces données dans cet espace virtuel. Cette répartition granulaire des données optimise, par voie de conséquence, la confidentialité des données et de leurs traitements (Laudon et al., 2013), ce qui présente un avantage pour le moins intéressant pour

l'entreprise et les internautes. Toutefois, pour l'entreprise, il s'agit de peser le pour et le contre d'un tel service cloud, car en fournissant l'intégralité de ses données à une compagnie qui offre ce type de service, elle en devient totalement dépendante (Laudon et al., 2013). Qu'adviendra-t-il dans le cas où le prestataire de service cloud venait à rencontrer des problèmes (faillite, pannes…) ? C'est en tout cas les défis auxquels devra répondre le cloud computing de demain. En attendant, il reste un des outils indispensables pour le Big Data, et ce, parallèlement au développement des bases de données NoSQL qui permettent le traitement de données structurées et non-structurées (Kitchin, 2014).

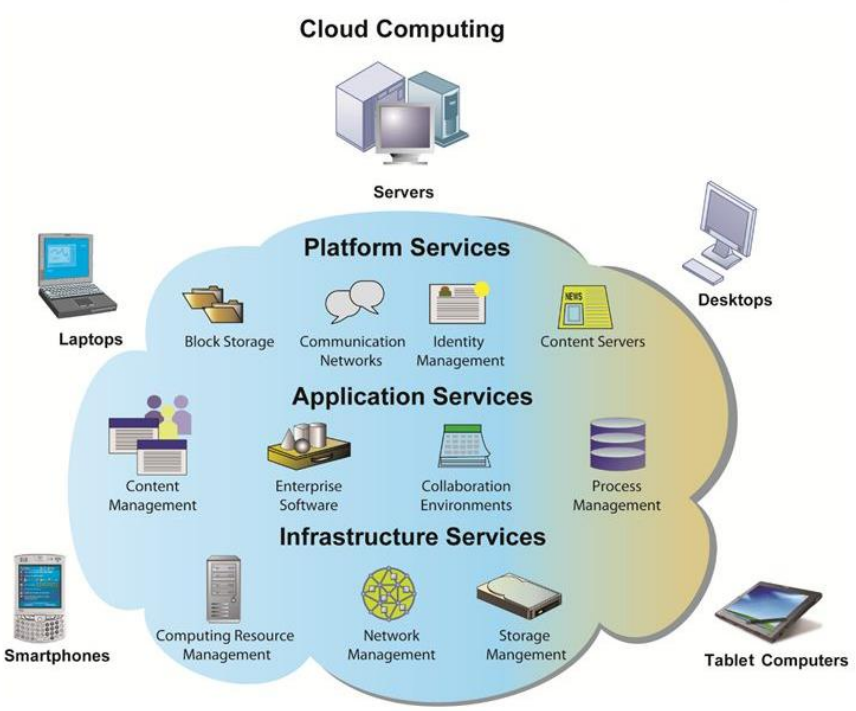


Figure 4 : Les services offerts par le cloud computing

Source : Laudon, K., Laudon, J., Fimbel, É., Costa, S. et Canevet-Lehoux, S. (2013). *Management des systèmes d'information* (13e éd.). Montreuil : Pearson France.

4.2. Les sources du Big Data

Après avoir dressé un tour d'horizon des outils nécessaires à l'utilisation du Big Data, il convient d'établir quels sont les différents types de sources auprès desquelles le Big Data puise ses données. Kitchin (2014) indique que, de manière générale, il est possible de classifier ces sources en trois catégories distinctes : la première étant les Directed Data, la seconde les Automated Data, et enfin, les Volunteerd Data. Les points suivants parcourent chacune de ces catégories, en spécifiant les particularités de celles-ci et les sous-catégories qui les composent.

a. Directed Data

Les Directed Data sont des données très structurées dont l'objectif reste assez restreint. Dans la pratique, nous retrouvons ce type de données lors d'observations réalisées par des individus lambda comme des agents de police, des enseignants, des médecins qui dans le

cadre de leur activité collectent toute une série d'informations bien spécifiques. Le médecin constatera l'état de santé général du patient, le professeur attribuera une note à chacun de ses élèves, alors que l'agent de police dressera un procès-verbal. Les données peuvent également être récoltées par l'intermédiaire d'un outil technologique comme des recensements ou la complétion d'une fiche d'impôt via le site internet du Gouvernement (Kitchin, 2014).

b. Automated Data

À la différence des Directed Data, les Automated Data n'ont plus réellement besoin de la main de l'Homme pour être générées. Comme l'appellation anglophone l'indique, il s'agit de données produites automatiquement à l'aide d'outils technologiques. Par ailleurs, ces données sont non seulement produites automatiquement, mais également traitées et analysées, pour ensuite permettre à leurs détenteurs d'effectuer certaines actions en fonction des résultats obtenus (Kitchin, 2014). Kitchin (2014) sous-catégorise ces Automated Data en cinq parties que nous proposons de développer brièvement ci-dessous.

b.1. La surveillance automatisée

Cette première sous-catégorie se focalise sur la mise en place de systèmes de surveillance autorisant le suivi de certaines activités humaines. Il s'agit par exemple de gestionnaires de trafic, soit des caméras qui permettent à un feu de signalisation de rester un peu plus longtemps au vert d'un côté du carrefour si celui-ci présente un nombre plus important de véhicules. Cette régulation est totalement automatisée et est rendue possible par la gestion instantanée de données venant de multiples sources (ici, les différentes caméras positionnées de chaque côté d'un carrefour). Autre exemple, les radars dont l'objectif est de détecter des infractions au code de la route. Si un véhicule circule à une vitesse trop élevée, le radar le détecte, identifie la plaque d'immatriculation du véhicule, enregistre l'information auprès d'une base de données et génère une contravention qui est automatiquement envoyée à son destinataire, sans la moindre intervention humaine (ou en tout cas très minime). Enfin, l'on retrouve également ce système de surveillance automatisée dans les titres de transports. Les données générées permettent d'identifier l'afflux de personnes auprès de l'une ou l'autre station, ce qui permet par la suite au gestionnaire du réseau de transports d'augmenter la fréquence de passage sur la ligne identifiée (Kitchin, 2014).

b.2. Les appareils digitaux

Les appareils digitaux font désormais partie de notre quotidien. Ils nous permettent de prendre des photos, de filmer, de téléphoner, de nous réveiller, de mesurer la température de notre lieu de résidence, etc. Ces appareils digitaux constituent notre seconde sous-catégorie, car ces derniers génèrent automatiquement, et de différentes manières, des données. Des données qui, selon les cas, seront enregistrées, analysées, combinées à d'autres afin par exemple d'analyser le comportement des utilisateurs (Kitchin, 2014). Dans la pratique, un individu qui ferait usage de son smartphone lors de ses vacances en Espagne postera des photos sur sa page Facebook depuis ce lieu. Sa position étant connue

et liée à la photo, il serait dès lors envisageable (à la suite du transfert de données) pour une compagnie commerciale de lui proposer des activités, des produits ou services dans le lieu où il se trouve.

b.3. Les données captées

Troisième sous-catégorie des données automatisées, les données captées sont issues des capteurs et détecteurs et font partie intégrante de l'Internet of Things. Ces technologies permettent entre autres de suivre l'évolution de la consommation d'eau, de gaz, d'électricité, du niveau de pression, etc. Ces capteurs et détecteurs enregistrent de façon continue des données et transmettent ces informations de manière passive (il s'agira de venir scanner l'information) ou active (les données sont transmises à intervalles réguliers). Ces outils technologiques peuvent par exemple être intégrés sous des ponts afin d'en mesurer l'état de corrosion, et dès lors prévoir sa rénovation. On les retrouve également au sein de véhicules comme les camions pour estimer le nombre d'heures durant lequel un chauffeur n'a pas arrêté de conduire, plus communément appelé boîte noire. Autre exemple, les puces RFID (radio frequency identification) qui permettent de suivre à la trace les produits sur lesquels elles ont été apposées. Le suivi continuel de l'acheminement des produits dans un système logistique va permettre de détecter les fraudes éventuelles, mais également la fiabilité et l'efficacité du système mis en place (Kitchin, 2014).

b.4. Les données scannées

À l'instar des données captées, d'autres outils permettent le suivi de produits, il s'agit de notre quatrième sous-catégorie, à savoir les données scannées. Dans la pratique, un code-barre sera présent sur chacun des produits vendus par une société (Kitchin, 2014). Une fois le produit scanné à la caisse, les données sont enregistrées et transmises. Le gestionnaire du magasin est directement informé de l'évolution de son stock (qui pourrait éventuellement être relié à un partenaire pour automatiser l'approvisionnement du produit en question).

Cette source de données citée par Kitchin (2014) ne permet pas la granularité des données, car le code-barre, comme nous l'avons vu (cf. supra p.33), n'autorise que l'identification du type de produit, sans aucune unicité par rapport aux autres produits du même type (un chocolat Mars, reste un chocolat Mars). Toutefois, le nombre de données générées reste très élevé et est produit en continu, ce qui répond à quelques-uns des éléments qui composent le Big Data (nous rappelons encore une fois qu'en ce jour, ils existent plusieurs formes de Big Data qui répondent à certaines des caractéristiques (cf. supra p.26) de cette technologie).

b.5. Les données d'interaction

Cette cinquième et dernière catégorie comporte les données issues de l'interaction entre l'utilisateur et la technologie. On l'observe lorsqu'un utilisateur se rend sur Internet pour acheter ou se renseigner sur l'un ou l'autre produit. Des cookies (qui peuvent être refusés par l'utilisateur, ce qui ampute alors l'expérience de celui-ci) vont permettre de suivre à

la trace l'ensemble des actions réalisées par l'utilisateur, par exemple : les produits insérés dans son panier d'achat, ceux qui ont été introduits puis retirés, etc. L'ensemble de ces informations est collecté et traité par les outils technologiques dont dispose le gestionnaire du site web. Ils permettront par la suite à celui-ci d'effectuer un meilleur ciblage des consommateurs (réunis selon leurs spécificités) ou de réaliser des études sur le comportement d'achat des internautes. Les données d'interaction se retrouvent aussi lors de transactions financières (retraits d'argent, virements, dépôts, etc.) ou lors d'appels téléphoniques où toutes les informations concernant ces appels sont enregistrées (comme le numéro appelé et la durée de la conversation téléphonique). Les données d'interaction sont donc des données générées automatiquement, et ce, sans que l'utilisateur en prenne conscience (Kitchin, 2014).

c. Volunteerd Data

Alors que dans les deux précédentes catégories, l'individu faisant l'objet d'études n'était pas acteur de la production de données, il en devient ici le producteur principal. Le Volunteerd Data désigne, en effet, l'ensemble des données qui sont volontairement transmises à une institution ou une organisation par un individu qui, d'une certaine manière, consent à révéler des informations d'ordre privé. De nos jours, il existe plusieurs occasions qui incitent les individus à faire part de leurs données (Kitchin, 2014). Les points suivants décrivent brièvement ces différentes occasions.

c.1. Les transactions

Lors de transactions commerciales, il est de manière générale demandé aux acheteurs de fournir un certain nombre d'informations les concernant, dont certaines sont obligatoires. Il peut s'agir des nom et prénom, mais également du lieu de résidence, du lieu de livraison, l'âge, le genre, la profession, etc. Toutes ces informations seront liées au compte des acheteurs et permettent au gestionnaire du site web de dresser le profil des internautes (type d'achat effectué, fréquence d'achat, etc.). Par la suite, le gestionnaire du site web peut solliciter les acheteurs afin d'entretenir la relation et établir quelles sont les actions qui pourraient être entreprises pour améliorer leur expérience d'achat. Lorsque les acheteurs se prêtent au jeu, ils fournissent des données qui constituent une valeur ajoutée pour le gestionnaire du site web qui, dès lors, bénéficie d'une meilleure vision de son environnement. D'autres situations peuvent également inciter les internautes à faire part d'informations privées, comme les sites promettant des coupons, des bons de réduction, et autres (Kitchin, 2014).

c.2. Les réseaux sociaux

À l'heure de la digitalisation de nos rapports sociaux, les réseaux sociaux sont devenus un passage obligatoire pour l'industrie du marketing (Balagué, 2017), et de ce fait de l'industrie du sondage. Les réseaux tels que Facebook, Twitter, YouTube, etc. fourmillent d'internautes qui sont tantôt acteurs de leur environnement digital, tantôt producteurs de celui-ci. Cette modification du comportement des internautes qui, à l'origine, étaient des acteurs passifs (bénéficiant des apports du web) est due à l'avènement du Web 2.0. Ce renouveau de la toile est venu littéralement révolutionner leur regard sur ce monde virtuel

(Kitchin, 2014). Par ailleurs, cette révolution est telle que les métiers afférents au marketing se voient dans l'obligation d'adapter leurs modes de fonctionnement afin de bénéficier des retombées cognitives de ces réseaux sociaux. En effet, ces plateformes sociales sont des lieux d'échanges où les individus peuvent exprimer librement leur satisfaction envers l'une ou l'autre expérience vécue, partager leur avis sur un produit, un service, un événement, etc. Dès lors, il devient essentiel pour l'entreprise de disposer d'outils technologiques, mais également de formations appropriées, pour récolter, traiter et analyser une multitude d'informations venant de plateformes différentes, et parmi ces outils, le Big Data (Balagué, 2017).

c.3. La sousveillance

La sousveillance s'oppose à la surveillance dans le sens où elle viendrait contrebalancer le pouvoir des autorités qui scrutent en continu la population (par exemple : les caméras de surveillance disposées à maints endroits dans les villes). Ce concept de sousveillance désigne dans la pratique la volonté de l'individu d'utiliser les technologies mises à sa disposition (caméras, appareils photos, smartphones…) pour collecter les données issues de son environnement. Cette collecte de données pouvant dès lors servir de preuve à charge ou à décharge dans le cas, par exemple, d'une altercation avec les forces de l'ordre (Mann, Nolan et Wellman, 2003).

Kitchin (2014) élargit quelque peu ce concept de sousveillance qui peut aussi désigner l'utilisation d'outils technologiques pour s'autosurveiller. Dans la pratique, il s'agit entre autres de montres connectées qui nous informent de notre fréquence cardiaque, du nombre de calories perdues, d'outils mesurant la pression artérielle ou le niveau de glycémie, etc. Toutes ces données personnelles, une fois récoltées, sont retransmises aux fournisseurs d'applications, de logiciels, etc. qui utilisent celles-ci pour avoir un meilleur aperçu des comportements humains, et dès lors, envisager des actions commerciales (ou autres) auprès de ces individus. Toutefois, ces outils de sousveillance que nous transportons au quotidien ne sont qu'à leurs débuts, et ne permettent pas d'exploiter au maximum ces données (Kitchin, 2014).

c.4. Le crowdsourcing

Si le crowdfunding se caractérise par une contribution financière de la part du grand public, le crowdsourcing se caractérise par une contribution intellectuelle. L'objectif du crowdsourcing est donc de récolter de l'information, des idées, des données auprès des citoyens (qui, le cas échéant, peuvent être rémunérés) afin de permettre la résolution d'une problématique (Kitchin, 2014). En faisant ainsi appel à la communauté, l'entreprise bénéficie d'idées innovantes, mais qui plus est peu onéreuses, voire gratuites. La vague d'engouement à l'égard de ces plateformes collaboratives n'a cessé d'augmenter au cours des dernières années. Pour preuve, entre 2013 et 2014, la demande des entreprises envers le crowdsourcing s'est accrue de 48% (Signoret, 2015).

Le recours au crowdsourcing se fait souvent sur base de challenge où les participants sont invités à laisser s'exprimer leurs talents, que ce soit pour la création de contenus vidéos,

d'applications ou autres (Signoret, 2015). Kitchin (2014) répertorie trois principaux types de crowdsourcing :

- la première est celle qui, par la collaboration, est génératrice d'idées (exemple : Wikipédia),
- la seconde fait appel à la communauté pour donner leur avis sur l'un ou l'autre produit ou service (exemple :TripAdvisor reviews),
- la troisième cherche quant à elle une ou plusieurs solutions pour répondre à une problématique (exemple : InnoCentive).

c.5. Science citoyenne

À l'instar du crowdsourcing, la science citoyenne fait appel à la communauté, mais cette fois-ci au bénéfice de la science. Dans le cadre de recherches scientifiques de grandes envergures, où les outils, le personnel et/ou les moyens financiers sont insuffisants, le grand public est invité à contribuer activement à la recherche (Kitchin, 2014). Les demandes sont diverses et variées comme, par exemple, celle de la Nasa qui invitait les internautes à trouver la 9ème planète du système solaire (Lucchese, 2017). On retrouve aussi les demandes des instituts météorologiques qui appellent les citoyens à faire des relevés météorologiques. D'autres demandes émanent d'organisations qui souhaitent répertorier les animaux, insectes, plantes vivant dans un territoire donné (Kitchin, 2014).

Cette science citoyenne, également connue sous l'appellation de science participative, est donc le fait de non-professionnels volontaires apportant une aide non-négligeable à la communauté scientifique. Par l'intermédiaire de cette participation massive d'individus dans la récolte de données, le Big Data est, largement et de manière continue, alimenté en données hétérogènes. Ces dernières peuvent ensuite être traitées, analysées et répertoriées selon leurs particularités. La science citoyenne (participative) est par ailleurs essentielle pour combler les lacunes technologiques auxquelles nous sommes confrontés en ce jour, car il est de toute évidence bien plus aisé à un individu d'apprécier les non-dits figurant dans un texte (les ressentis positifs ou négatifs) que de laisser la technologie effectuer cette tâche (Amer-Yahia, Ganascia, Ogier, 2017).

4.3. Les domaines d'application

Nombreux sont les domaines où le Big Data pourrait apporter un avantage compétitif aux entreprises et individus qui en feraient bon usage. Du secteur privé au secteur public, les demandes et les utilisations de cette technologie ne cesseraient de s'intensifier auprès des différents acteurs du marché. Nous souhaitons, dès lors, aborder ci-dessous quelques-uns des domaines où le Big Data viendrait directement rivaliser avec l'industrie du sondage traditionnel tels que l'activité politique, le secteur public, le secteur de la santé et le secteur privé. Bien évidemment, nous avons volontairement omis d'évoquer certains domaines d'application du Big Data, comme celui de la recherche scientifique, car ceux-ci outrepassent le cadre de notre objet d'étude, car ne représentant aucunement une menace pour l'industrie du sondage traditionnel.

a. L'activité politique

S'il y a bien un domaine qui, depuis longtemps, s'est intéressé à l'opinion publique, c'est bien celui du monde politique. Jusqu'alors, pour mesurer le pouls de la société, plusieurs choix s'offraient aux politiques, dont celui qui nous concerne directement, les sondages. Cette pratique, très largement utilisée en période électorale, permet aux politiciens de réévaluer leur stratégie et d'adapter celle-ci en fonction de l'électorat. Par ailleurs, les sondages permettent aux politiciens de ne pas avoir de mauvaises surprises lors des résultats définitifs. Toutefois, l'efficacité des sondages est de plus en plus souvent remise en question. Les erreurs d'estimation lors de récents événements tels que le référendum britannique sur le Brexit, l'élection présidentielle américaine, l'élection aux primaires de François Fillon (Les Républicains) ou celle de Benoît Hamon (Parti socialiste), etc. en sont les principaux facteurs (Tronchet, 2017).

Pour autant, l'engouement pour les données électorales ne s'est pas amoindri. En effet, ce qui change pour les politiciens, c'est surtout la méthode de sondage. Car, aujourd'hui, les politiques semblent tout doucement se rapprocher du Big Data qui serait, selon Liegey (cité par Georis, 2017), le meilleur outil mis à la disposition des politiciens. Technologie qui fut d'ailleurs utilisée lors de la campagne d'Obama en 2012, et qui a entre autres permis le succès du candidat. La stratégie du président réélu consistait à récolter une quantité de données sur l'ensemble des électeurs (profils, registres électoraux publics…), pour ensuite mettre ces données en relation (pour bénéficier de croisements pertinents, de comparaisons…), et enfin en faire bon usage (campagnes d'emailings ciblées auprès de ses supporters, porte-à-porte avec le discours approprié…) (Berry, 2012). Depuis lors, d'autres en ont fait l'usage, dont le Président Macron, Hillary Clinton, politiciens du Parti Socialiste, etc. (Georis, 2017).

Pour le monde politique, la récolte de données reste incontournable. Lors des campagnes électorales, il est nécessaire pour le politique d'identifier les individus qui seraient prêts à voter pour le candidat, mais également ceux qui souhaiteraient participer activement à la campagne (par le biais du volontariat) et ceux disposés à effectuer des dons financiers. De plus, lors d'une campagne, il est important de prévoir à quels types de profils seront confrontés les volontaires qui feront du porte-à-porte ainsi que le type de communication à adopter pour transmettre un message qui dans le fond restera similaire, mais dans la forme sera différent (Nickerson et Rogers, 2013).

Le Big Data semble bel et bien ancré au sein du monde politique qui, selon Liegey (cité par Georis, 2017), a compris l'utilité de cette technologie et devrait à l'avenir en faire plus largement appel. Cet avenir du Big Data, au vu des éléments développés plus haut, nous paraît radieux, et représente dès lors une menace significative pour l'industrie du sondage, sans cesse mise à mal par les multiples remises en question. Face à une granularité permettant l'orientation de l'action politique de manière ciblée et précise, quel poids jouerait celui du sondage traditionnel qui ne peut se targuer d'une telle efficacité et se contente d'estimations globales teintées d'approximations ?

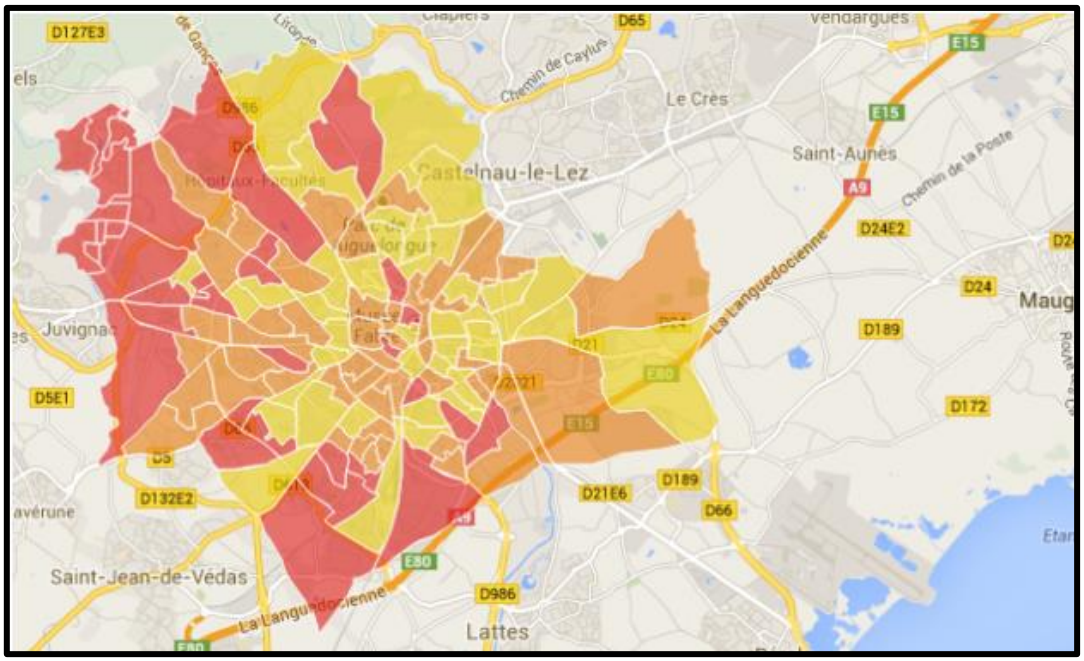


Figure 5 : Logiciel « 50+1 » utilisé lors d'une stratégie électorale en porte-à-porte

Source : Chamay, A. (2014). Des logiciels pour gérer sa campagne comme un produit commercial. *01net.com.* Récupéré de https://www.01net.com/actualites/des-logiciels-pour-gerer-sa-campagne-comme-un-produit-commercial-616428.html

b. Le secteur public

Si les sondages traditionnels et le Big Data intéressent fortement nos politiques dans le cadre des prévisions électorales, ces outils leur plaisent tout autant lorsqu'il s'agit d'établir des politiques sociales, économiques, etc. Pouillon (1951) le faisait déjà remarquer il y a près de 70 ans lorsqu'il évoquait l'objectif le plus manifeste des sondages, à savoir la faculté d'établir ce qui compose la pensée des individus au regard de questions sociétales ou d'ordre politique.

Les acteurs publics ont toujours été très demandeurs en matière de données. Qu'il s'agisse de données en vue de gérer et de contrôler leur population ou de données permettant de mieux desservir celle-ci, les acteurs publics n'ont cessé d'amasser les faits économiques, sociaux et autres. L'une des méthodes qui leur a permis jusqu'alors de se représenter une image fidèle de la société était la quantification systématique des réalités sociales ainsi que les avancées technologiques (Kitchin, 2014), et bien entendu toutes les techniques qui sont propres aux sondages (travaux en psychologie, études comportementales…).

Avec le Big Data, soit l'utilisation massive de données plutôt qu'un échantillon restreint, de nouvelles perspectives s'offrent aux entreprises afin d'améliorer la gestion des villes, et leur développement. Un progrès technologique patent, jusqu'alors inimaginable pour les acteurs publics, et qui se traduit par une meilleure compréhension du citoyen dans sa globalité. Par ailleurs, on observe l'émergence de villes devenues des laboratoires très largement digitalisées afin de récolter un maximum de données, et dont l'objet serait de découvrir les nouvelles applications possibles de ces masses de données (Eveno, 2017).

La question qu'il serait légitime de se poser à ce stade serait : « Quel avenir pour les sondages traditionnels dans le secteur public, à l'heure du Big Data ? ». Cet avenir nous semble pour le moins peu reluisant. En effet, le Big Data rivalise avec des arguments de poids. Frega et Tsoukiàs (2017) en proposent trois. Ils indiquent que, d'une part, cette technologie, dans le cadre d'études publiques, permettrait d'analyser de manière bien plus précise les souhaits des citoyens, que ce soit en matière de transport, soins de santé, d'énergie, etc. D'autre part, grâce à l'une des particularités du Big Data, la granularité, il serait envisageable pour les acteurs publics de personnaliser les services à la personne en proposant des services ajustés qui tiennent compte des spécificités de chaque citoyen. Enfin, l'innovation au sein du secteur serait largement stimulée avec à la clé de nouvelles politiques qui, en raison de leur complexité, nécessitent la manipulation de volumes considérables de données (Frega et Tsoukiàs, 2017). Une nouvelle fois, selon nous, l'industrie du sondage traditionnel semble quelque peu mise en danger par ce mastodonte de la manipulation de données, le Big Data.

c. Le secteur de la santé

À l'instar de toute entreprise à caractère commercial, le secteur pharmaceutique est lui aussi friand de sondages en tous genres. Nous avons d'ailleurs eu à maintes reprises l'occasion de le constater au sein de Dedicated, institut de sondages dans lequel nous exerçons notre rôle de chargé d'études. De manière générale, ces sondages commandés par l'industrie pharmaceutique ont pour objectif de s'informer sur les comportements de trois groupes-cibles :

1. les particuliers : c'est-à-dire les individus susceptibles de contracter une maladie en particulier ou qui sont (ont été) atteints par la maladie en question. À ce stade, les sondages proposent d'identifier les comportements de ces patients en matière de compliance face aux traitements, d'habitudes d'achat, les relations que ceux-ci entretiennent avec leur médecin et pharmacien, etc.

2. les médecins : soit tous les médecins, qu'ils soient ou non spécialistes. L'objectif étant ici d'établir quelles sont les habitudes de prescription des médecins. En d'autres termes, pour quelles raisons prescrivent-ils tel médicament en de telles quantités plutôt qu'un autre médicament (d'une autre marque) dans une quantité différente, etc.

3. les pharmaciens : tout comme les médecins, l'intérêt est de recueillir des données nécessaires au mode de fonctionnement des pharmaciens : établir leur impact lors de recommandations de médicaments avec ou sans prescription, les conseils qu'ils procurent aux malades, les médicaments qu'ils préfèrent proposer et les raisons de ces choix, etc.

Selon Corvol (2017), le Big Data est le principal fournisseur de données du secteur médical. Cette technologie permettrait notamment à la médecine de répondre aux besoins contemporains du monde médical, à savoir le développement d'une médecine qui se dit « prédictive, de précision, participative et préventive » (Corvol, 2017, p.247). En effet, si

tels sont les souhaits de cette nouvelle médecine, le Big Data est en mesure de collecter les masses de données hétérogènes provenant de millions de malades, et de traiter et d'analyser ces données pour en dégager des informations consistantes pour le monde médical.

En matière d'études prédictives, le Big Data autorise le suivi et le croisement d'une multitude de données relatives aux patients comme, par exemple, leurs antécédents, leurs réactions à tels types de médicaments, mais également l'évolution de leur état de santé et les conditions environnementales dans lesquelles ces patients se trouvent. L'agrégation de toutes ces données devrait permettre au monde médical (et par extrapolation au monde pharmaceutique) de mieux prévenir des épidémies dans telle ou telle région. Outre ce bénéfice, le Big data permet un gain de précision grâce à cette spécificité qu'est la granularité. Dans la pratique, cette précision se manifeste par une réelle personnalisation du traitement médical qui tient compte tant des facteurs globaux que des particularités de l'individu. Enfin, l'apparition du phénomène d'auto-contrôle en matière de santé, rendu possible par les objets connectés, donne aux mondes médical et pharmaceutique la possibilité de suivre au quotidien l'état de santé de la patientèle. Ce qui devrait permettre d'identifier les comportements à risques et les habitudes des patients afin d'établir l'une ou l'autre corrélation pertinente (Corvol, 2017).

Au vu des avantages fournis par le Big Data (et les promesses de celui-ci) pour le secteur pharmaceutique et le secteur médical, il devient légitime de questionner l'utilité du sondage traditionnel. En d'autres mots, quel pourrait être l'apport supplémentaire d'une enquête quantitative réalisée auprès d'un millier d'individus répondant de manière approximative à des questions d'ordre médical ?

d. Le secteur privé

À l'ère où l'information est primordiale dans cet environnement très concurrentiel, il nous semble très difficile d'imaginer que le secteur privé puisse se passer d'études de marchés. Notre expérience nous a effectivement fait l'état d'une très forte demande du privé, et ce, quel que soit le secteur d'activité concerné : secteur du commerce de détail, secteur automobile, secteur du service, etc. Qu'il s'agisse d'un nouveau produit ou service, ou de produits ou services existants, les entreprises estiment à juste titre qu'il est indispensable de recourir aux sondages. Les objectifs de ces derniers sont directement liés aux besoins des entreprises qui souhaitent établir le niveau de satisfaction de leur clientèle, le niveau d'attractivité des produits et/ou services qu'ils proposent, les habitudes et connaissances des consommateurs concernant un marché en particulier, la notoriété de certains produits, etc. L'engouement du secteur privé pour les études de marchés n'étant plus à prouver, le seul point qu'il nous reste à éclaircir étant, selon nous, sous quelle forme, selon quel processus celles-ci continueront à se dérouler.

Une première piste de réponse viendrait tout droit du secteur automobile, en tout cas selon Montcheuil (2014) qui démontrait quelques années auparavant l'utilité du Big Data au sein de l'industrie automobile. Cette technologie permettrait, selon cet expert, de récolter des données tout au long du processus d'achat du véhicule, en commençant par sa

conception, sa mise en production, son acheminement auprès des différents revendeurs et la vente du véhicule en question. Outre cette première étape dans la collecte de données, les capteurs intégrés aux véhicules ainsi que les comportements digitaux des futurs acheteurs et prospects (par exemple : visite d'un site web de la marque, complétion d'un formulaire en ligne…) viendraient compléter une base de données déjà bien fournie en amont. L'agrégation des données récoltées permettrait ensuite une analyse concise du marché dans sa globalité, mais également dans sa granularité. Au cas par cas, un concessionnaire pourrait dès le premier contact avec son client être au courant des désidératas de ce dernier et, dès lors, lui proposer un véhicule et des services dédiés qui conviennent au mieux à l'utilisateur final du véhicule. De plus, l'analyse du comportement des conducteurs comme la conduite excessive (ex. : poids lourds) ou le nombre d'heures de conduite quotidienne aiderait les compagnies d'assurance à établir de meilleures estimations concernant les risques d'accident qu'encoure un individu en particulier. Le prix des assurances serait adapté à cet individu sans pour autant impacté les tarifs octroyés aux autres conducteurs n'adoptant pas le même type de comportement au volant (Montcheuil, 2014).

Une autre piste de réflexion viendrait de l'industrie des biens de grande consommation plus connus par leur acronyme anglais FMCG (Fast-Moving Consumer Goods). Selon Floridi (2018), le Big Data serait en mesure d'améliorer le processus d'apprentissage du marché en exploitant les données issues des différentes transactions commerciales ainsi que le comportement des consommateurs (en magasin ou via l'e-commerce). Les données récoltées autoriseraient déjà les Brand Managers et Product Managers à mieux concevoir leur clientèle et aux entreprises de rationaliser leurs structures de coût. Les masses de données permettraient également aux entreprises d'innover et de mieux cibler leurs consommateurs (Floridi, 2018).

Tout comme nous l'avions évoqué pour le secteur public, le secteur de la santé et celui du monde politique, l'avenir du Big Data au sein du secteur privé nous semble, aux dépens de l'industrie du sondage traditionnel, très prometteur. En effet, si les promesses du Big Data sont exactes, et en supposant que les coûts d'extraction et d'analyse des données soient relativement abordables (actuellement ou dans un futur proche), nous ne pouvons envisager un avenir prospère à cette industrie traditionnelle qui aura fait son temps !

4.4. Les qualités et faiblesses du Big Data

Après avoir présenté le Big Data sous toutes ses coutures, il nous reste à établir quelles sont ses qualités et promesses d'une part ainsi que ses faiblesses et désillusions d'autre part. Les éléments que nous développons ci-dessous permettront au lecteur d'avoir une meilleure vision des avantages et inconvénients qu'offre le Big Data sans pour autant tomber dans le piège tendu par les partisans et les opposants de cette technologie qui ne cessent de s'affronter arguments à l'appui (Ollion et Boelaert, 2015).

a. Qualités et promesses du Big Data

Force est de reconnaître que l'une des qualités majeures du Big Data est la récolte de données auprès d'individus qui ne sont pour la plupart pas, ou en tout cas pas totalement, conscients que des informations les concernant sont entreposées dans des serveurs en vue d'être décortiquées et analysées à des fins diverses et variées. Pinker (2017) indique que les individus, du fait qu'ils ne se sentent pas surveillées, ont tendance à laisser s'exprimer leurs opinions, leurs avis sur des sujets sans prendre le risque d'être jugés par d'autres. Cette qualité propre à ce qu'autorise cette technologie nous est pratiquement impossible à gommer lorsque des sondages d'opinion sont effectués via la méthode de sondages traditionnelle. D'aucuns pouvant considérer à tort qu'il existe une réponse, parmi tant d'autres, qui serait plus juste, plus acceptable au vu de l'interrogateur que celle qui se trouve dans leur for intérieur. Une supposition qui, malheureusement, biaise les résultats !

Pour étayer quelque peu cette qualité, il nous suffit par exemple de remonter aux élections présidentielles américaines, et plus particulièrement celles de Barack Obama (en 2008 et 2012). Lors de ces élections, qui verront par deux fois un président noir être élu à la tête de l'une des plus grandes puissances mondiales, l'un des questionnements qui traversait les esprits à l'époque était celui s'interrogeant sur l'origine ethnique du candidat. Pour la majorité des Américains, selon des sondages traditionnels menés avant et après les élections présidentielles, l'origine ethnique des candidats n'avait eu aucune influence sur leur choix d'élire ou non ce candidat. Toutefois, cette réalité idyllique qui nous a été proposée par les sondages traditionnels s'est malheureusement avérée erronée. En effet, à la suite de recherches menées (à l'aide de Google Analytics) auprès de différents États américains sur les propos racistes qui pullulent sur la toile, on découvre une corrélation positive entre le taux de personnes se qualifiant de racistes ou ayant utilisé des termes racistes lors de leurs recherches sur Internet et le taux de personnes qui, dans un État en particulier, vote pour un candidat en raison de l'appartenance ethnique de ce dernier. Lorsque dans un État, le taux de racisme à l'égard des individus originaires d'Afrique noir était plus élevé, le candidat Obama obtenait un moins bon score que le candidat blanc auquel il était confronté. Cette corrélation était la seule à pouvoir expliquer la sélection d'un candidat au détriment de son opposant politique lorsque les individus analysés affichaient un même profil sociodémographique (Stephens-Davidowitz, 2017). Si les futurs candidats ne se fiaient qu'aux sondages traditionnels, ce type de données passerait totalement inaperçu, alors qu'il s'agit bien là d'informations d'ordre capital. Quelle que soit l'origine du candidat, celui-ci a le devoir de s'informer sur son électorat (potentiel) afin d'adapter sa communication auprès de celui-ci.

D'un point de vue social ou commercial, on remarque également que le Big Data peut venir confirmer ou infirmer certaines des affirmations émises par l'individu lambda. Lorsque dans un sondage de type traditionnel (mené aux États-Unis), l'on demande à des individus le nombre de contraceptifs masculins utilisés lors de rapports intimes au sein de leur couple, les chiffres varient grandement. Ainsi les résultats du sondage ont indiqué que selon le témoignage des femmes 1,1 milliard de préservatifs seraient utilisés par an, alors que les hommes sont beaucoup plus optimistes et prétendent en utiliser 1,6 milliard

par an (Stephens-Davidowitz, 2017, p.5). La loi du juste milieu pourrait nous proposer de retenir une consommation moyenne se situant entre les deux précitées. Cependant, la vérité est tout autre ! Il semblerait que la consommation américaine annuelle de contraceptifs masculins se situerait aux alentours de 600 millions d'unités selon les chiffres publiés par l'institut Nielsen (cité par Stephens-Davidowitz, 2017, p.5). Cette différence pose question quant à la connaissance qu'ont les individus sur leur propre comportement, en supposant bien entendu que ceux-ci ont répondu de manière honnête, ce que ne pense pas Stephens-Davidowitz (2017) qui prétend que les individus ont cette fâcheuse inclination vers le mensonge volontaire. D'ailleurs, toujours selon ce chercheur, les individus mentent continuellement lorsqu'ils sont questionnés sur des sujets sensibles ou non tels que le niveau d'activité sportive, le nombre de livres lus, le prix de leurs vêtements, etc. Pour appuyer ses propos, Stephens-Davidowitz (2017) n'hésite pas à faire mention d'une étude comparative réalisée auprès des habitants de Denver (aux États-Unis) en 1950 qui démontre des différences significatives entre les résultats obtenus à la suite de sondages et les résultats obtenus à la suite de décomptes officiels :

	Reported on survey	**Official count**
• Registered to vote	83%	69%
• Voted in last presidential election	73%	61%
• Voted in last mayoral election	63%	36%
• Have a library card	20%	13%
• Gave to a recent Community Chest charitable drive	67%	33%

Tableau 2 : Différence entre les résultats de sondages et décomptes officiels

Source : Stephens-Davidowitz, S. (2017). *Everybody lies* (p.106). Londres : Bloomsbury Publishing..

Comme nous pouvons l'observer dans le tableau précédent, les chiffres varient fortement entre les sondages traditionnels et les décomptes officiels. Des variations d'une dizaine de pourcents, mais qui peuvent très vite passer du simple au double, comme c'est le cas dans le cas des dotations aux œuvres de charité. De cette perspective, il nous paraît évident que les analyses extraites par l'intermédiaire du Big Data présentent une plus grande valeur ajoutée d'un point de vue purement qualitatif (les données sont extraites à l'insu des répondants, alors qu'ils ne se sentent pas observés), mais également quantitatif (les masses de données permettent une plus grande précision à l'aide de l'exhaustivité).

Nous aimerions, toutefois, avant de poursuivre sur les qualités du Big Data, nuancer les propos de Stephens-Davidowitz (2017). Ce spécialiste en data science, en voulant prêcher pour sa paroisse, estime que les différences détectées entre sondages traditionnels et décomptes officiels sont uniquement dues aux mensonges volontaires des individus, d'où le titre de son ouvrage « Everybody lies ». Les mensonges biaisent effectivement les résultats que les instituts de sondages obtiennent. Cependant, il nous faut relativiser ces dires. Les individus ont selon notre expérience plutôt tendance à mentir sur des sujets

intimes (en regard de leur sexualité) ou personnels (convictions politiques), à exagérer des faits dans certains cas, mais ont également leur propre perception de la réalité (qui peut être assez éloignée de la vérité), sont sujets aux erreurs de jugement, d'interprétation, etc. Ce sont tous ces éléments qui portent préjudices aux sondages traditionnels, et non la simple volonté de mentir. À titre d'exemple, lorsque notre institut de sondages interroge les individus sur leur dernier vote aux élections communales, une proportion significative de répondants fait mention d'un parti politique qui n'était pas présent aux élections communales (mais régionales). Un oubli involontaire du participant qui est persuadé d'avoir voté pour ce parti aux élections communales. Ce qui, pour rappel, n'est pas toujours le cas en Belgique où les partis peuvent exister au niveau régional, mais pas au niveau communal, et inversement. La confusion existe bel et bien dans les esprits, et il faut en tenir compte quel que soit le procédé utilisé pour sonder la population.

Autre qualité reconnue à la technologie Big Data, le pouvoir de prédiction qui, selon Strong (2015), offrirait de nouvelles perspectives aux entreprises. Car, effectivement, même si la connaissance du marché actuel est primordiale pour les sociétés, l'évolution de celui-ci l'est tout autant. Par l'intermédiaire du Big Data, les entreprises disposeraient de plus d'éléments factuels leur permettant de mieux appréhender les consommateurs de demain auprès desquels elles proposeront leurs produits et services. Cette tendance qui consiste à utiliser le Big Data se fait de plus en plus sentir tant au niveau commercial qu'au niveau scientifique (Strong, 2015).

En guise d'exemple, le cas de la société American Express qui souhaitait mettre en exergue les similitudes que pouvaient avoir les Américains qui ne s'affranchissaient pas de leur dette, à la suite de l'utilisation de leur carte de crédit. Après avoir analysé des masses de données, une caractéristique apparaît et expliquerait le lien de causalité entre détenteur de carte de crédit et mauvais payeur. La société découvrira que les utilisateurs de l'American Express ayant une adresse de facturation en Floride étaient les plus susceptibles d'être en défaut de paiement (profitant d'une loi favorable de cet État à l'égard de ces derniers). American Express n'est pas la seule société à avoir bénéficié du Big Data, Amazon société dont la réputation n'est plus à démontrer recourt également à cette technologie. Celle-ci leur permet d'anticiper les demandes des consommateurs pour l'un ou l'autre produit qui, concrètement, aboutit à une réduction du temps de transport et de coûts de chacune des étapes de leur chaîne logistique (Strong, 2015).

Toutes les données récoltées par les entreprises, qui font appel au Big Data, leur permettent de mieux anticiper les mouvements opérant au sein du marché et de créer par la même occasion un avantage compétitif vis-à-vis de leurs concurrents (Strong, 2015). Le Big Data propose à ce stade un potentiel énorme en matière de récolte de données pertinentes qui nous semble difficilement envisageable à l'aide du sondage traditionnel. Par exemple, en supposant que nous atteignons les mauvais payeurs dans le cas de l'American Express (à l'aide de la méthode traditionnelle), rien ne nous dit que ces personnes accepteraient de participer à un sondage sur le crédit, quand bien même ce sondage serait rémunéré.

À l'origine, le Big Data présentait également un certain nombre de promesses pouvant être regroupées selon trois types dont nous font part Ollion et Boelaert (2015) qui, comme nous le verrons dans le point suivant (cf. infra p.52), viendront nuancer quelque peu ces promesses émises par certains experts :

- des promesses de type empirique : la technologie permet d'accumuler d'immenses quantités de données. Une fois ces données accumulées, celles-ci représenteraient pour leurs utilisateurs une nouvelle source d'informations qui devraient leur permettre de mieux identifier les comportements des individus et de déceler des particularités propres à ces derniers. Ce qui n'aurait pu être envisageable via la méthode de sondage traditionnelle. C'est le cas notamment avec les capteurs intégrés dans les véhicules, les cartes de transport, etc. qui renvoient automatiquement, instantanément et de façon granulaire (un capteur traçant le comportement d'un individu) des données relatives aux personnes et leurs caractéristiques comportementales (Ollion et Boelaert, 2015). Un sondage traditionnel est loin de pouvoir rivaliser avec une précision telle que proposée par le Big Data,
- des promesses de type méthodologique : alors qu'il y a peu les entreprises, les analystes et autres chercheurs devaient se limiter à une quantité d'informations assez restreinte, le Big Data permettrait d'analyser l'intégralité de ces données en temps réel. Ce qui autoriserait les utilisateurs de cette technologie à ignorer les dilemmes d'antan où, en raison de nombreux facteurs, seule une infime partie des données, parmi celles disponibles, étaient capturées pour dans un second temps être analysées. Le Big Data, comme nous l'avons évoqué plus tôt (cf. supra p.28), propose d'offrir aux chercheurs et entreprises le traitement de l'échantillon total (N = tout) au lieu d'un échantillon restreint (N = n) (Ollion et Boelaert, 2015),
- des promesses de type théorique : le Big Data aurait un impact positif sur la connaissance. Par l'intermédiaire de cette masse de données collectées, il deviendrait envisageable de faire apparaître de nouvelles vérités, de nouvelles corrélations qui n'auraient pu être identifiées à l'aide d'anciennes pratiques (Ollion et Boelaert, 2015). Chris Anderson (cité par Strong, 2015) partage également cet avis, mais va encore plus loin. Il affirme en effet que la théorie précédant un acte de recherche ne présente plus aucune utilité, il suffirait d'assembler les données émanant de la population étudiée pour en tirer les tendances et caractéristiques propres à cette population. Enfin, Cukier et Mayer-Schönberger (cités par Strong, 2015) estiment aussi qu'il ne faille plus s'attarder sur la raison d'un événement ou d'un trait comportemental, il suffit de les détecter et de les accepter tels quels. Dans la pratique, si l'on observe que les conducteurs ayant consommé un verre d'une boisson alcoolisée ont plus de chances d'occasionner un accident que les conducteurs totalement sobres, la démarche scientifique s'arrête ici même. Dès lors, aucune utilité d'étudier les effets de l'alcool sur le système nerveux, afin de compléter cette science fraîchement acquise.

b. Faiblesses et désillusions du Big Data

Si l'un des principaux atouts du Big Data est, comme nous l'avons évoqué au point précédent (cf. supra p.46), la récolte de données à l'insu des personnes, il se pourrait bien que cette immixtion dans la vie privée des citoyens soit (ou devienne) sa plus grande faiblesse. Comme l'indique Buléon (2017), cette technologie n'est pas toujours sans risque pour la population qui, par moments, la considère comme le Big Brother de notre génération. Ce sentiment d'insécurité a tout naturellement fait naître de nombreux questionnements au sein de la population qui s'interroge de plus en plus sur les bienfaits de cette technologie et les risques encourus en termes de confidentialité. Selon Buléon (2017), deux logiques transparaissent des débats publics, la première d'ordre sécuritaire (mesures de protection des citoyens contre les activités terroristes qui gangrènent la toile), la seconde étant l'exploitation commerciale des données. Si la première logique ne nous concerne pas directement, la seconde nous intéresse particulièrement dans le cadre de notre étude.

En tant que concurrent du sondage traditionnel, le Big Data ou en tout cas ses plus fervents défenseurs se targuent de pouvoir récolter une multitude de données à travers la toile pour en extraire des informations pertinentes, et ce, à l'insu des citoyens. Cela a effectivement été le cas récemment lorsque l'agence Belga (2017) nous apprenait que des données de patients étaient revendues à une multinationale alors que celles-ci étaient confidentielles. Des données qui apportaient de nombreuses indications relatives aux traitements suivis par les patients, qui par la suite devaient être revendues à des sociétés pharmaceutiques. Cette information a fait réagir les instances de l'État, qui ont appelé à une plus grande protection des données, ainsi que la population qui s'inquiète tout naturellement sur l'utilisation de ses données personnelles. Une autre affaire assez récente également étant celle qui s'est déroulée aux États-Unis où la société Cambridge Analytica est accusée d'avoir collecté des données de millions de comptes Facebook. L'objectif de cette récolte massive était de réaliser un profilage de cette population pour ensuite tenter d'influencer leur choix politique (L'Express.fr, 2018). Par ailleurs, une autre épine est venue se planter au pied de cette technologie, à savoir le Règlement Général pour la Protection des Données, plus connu sous l'acronyme RGPD ou GDPR (en anglais, General Data Protection Regulation).

Le RGPD, qui a vu son implémentation à l'échelle européenne le 25 mai 2018, a pour principaux objectifs la protection des informations personnelles des individus ainsi que le respect de leur vie privée. Dans la pratique, ce texte autorise les citoyens à faire usage de ce règlement pour se protéger des immixtions non-désirées des entreprises. En outre, les citoyens pourront également demander aux entreprises possédant leurs données de les effacer (Dèbes, 2018). D'autres éléments viennent également perturber le fonctionnement des entreprises tels que la durée de conservation des données (qui est désormais limitée), les finalités auxquelles sont utilisées ces données (sauf exception, les données ne pourront être utilisées qu'à celles initialement prévues), la sécurité des données, etc. (Commission européenne, 2018). En cas de non-respect du règlement, les entreprises s'exposeraient à de lourdes sanctions. Dans la pratique, une entreprise pourrait se voir infliger une amende

pouvant atteindre 20 millions d'euros ou 4% du chiffre d'affaires de l'année antérieure, la priorité sera donnée au montant le plus élevé (GDPR Associates, 2018, para.5). Autant l'avouer, en cas d'application correcte, le RGPD représenterait bel et bien une menace pour l'avenir du Big Data et nous amène à relativiser le potentiel de cette technologie qui devra trouver le moyen de contourner ce nouvel obstacle.

Certes, le Big Data possède, comme nous l'avons vu précédemment, un réel potentiel de prédiction. Il n'empêche que ces prédictions peuvent s'avérer fausses, en décalage avec la réalité. Pour preuve, l'outil Google Flu Trends dont l'objectif était de prédire les futures épidémies de grippe à travers le monde. Pour ce faire, les développeurs du programme ont estimé que les individus infectés par le virus de la grippe étaient plus susceptibles d'effectuer des recherches sur la toile. En indiquant leurs symptômes (maux de tête, etc.) sur la barre de recherche Google, ils émettaient un signalement immédiatement reconnu par l'outil développé par Google. Par l'intermédiaire d'algorithmes, Google parvenait à réaliser de bonnes estimations sur les futures épidémies de grippe (Stephens-Davidowitz, 2017). Malheureusement, quelques années après sa mise à disposition sur la toile (tout citoyen était en mesure de suivre l'évolution de la grippe en direct), l'outil démontrait qu'il ne suffisait pas de récolter des données aléatoirement à travers Internet pour pouvoir établir de telles estimations. En effet, en 2013, Google Flu Trends surestima d'environ 50% l'épidémie de grippe aux États-Unis par rapport aux chiffres officiels (Beauté, 2015, para.5). Cette erreur d'estimation était, en partie, due aux médias qui, voulant prévenir leurs concitoyens, répandaient malencontreusement une épidémie de recherches sur Internet de la grippe et ses symptômes (Beauté, 2015). Google Flu Trends n'ayant pas pris en compte cette variable dans ses calculs finira donc par démontrer que cette technologie est loin d'être infaillible en termes de prédiction. Malgré l'affirmation de Stephens-Davidowitz (2017) selon laquelle des chercheurs auraient, depuis ce fâcheux événement (parmi d'autres), réajusté l'outil de prédiction, la courte histoire du Google Flu Trends s'arrêtera quelques temps plus tard (Beauté, 2015). Sans doute les 100 erreurs sur 108 prédictions révélées par la revue Science en 2014 (citée par Beauté, 2015, para.5) ont joué un rôle prépondérant dans l'arrêt prématuré de l'outil de prédiction.

D'aucuns pourraient prétexter que l'échec cuisant du Big Data en matière de prédictions de grippe serait sans doute lié à la singularité ou à la complexité de ce domaine. Toutefois, d'autres exemples, ou plutôt d'autres échecs, démontrent le contraire. Parmi ceux-ci, les élections présidentielles américaines de 2016 où le Big Data fût utilisé afin d'anticiper le vote des électeurs. Les outils utilisés annonçaient Hillary Clinton aux commandes de l'une des plus grandes puissances mondiales, mais encore une fois la prédiction se soldera par un échec. Steve Hilton (2016) nous expose dans une interview accordée à Bloomberg certaines pistes de réflexion. La première étant la non-prise en considération des électeurs honteux du futur président D. Trump. Ceux-ci, en raison de plusieurs facteurs (notamment la pression de certains médias, mais également des autres citoyens), ont préféré cacher leur vote, et ont attendu les urnes pour le dévoiler dans la plus grande discrétion. Cette catégorie n'apparaissait dès lors pas dans les radars du Big Data. Tout comme la seconde piste de réflexion qui nous apprend que beaucoup d'électeurs étaient invisibles au Big

Data, car ces derniers ne croyant pas en la politique, n'avaient aucune raison d'établir une quelconque démarche telle que demander une carte d'électeur, s'inscrire dans un forum, partager un avis sur la politique, etc. D. Trump aurait, selon ce même spécialiste, réussi à raviver la flamme de ces apolitiques. Huffman (2016), durant cette même interview, pointe du doigt des problèmes méthodologiques (sans les nommer) ainsi qu'un facteur qui, semble-t-il, serait passé inaperçu, à savoir l'intensité de l'engagement des partisans pro-Trump sur les forums du site Reddit (dont Huffman est le Chief Executive Officer), plateforme de partage très prisée aux États-Unis (De Fournas, 2018). En résumé, ce second échec en termes de prédiction, nous apprend que le Big Data n'est pas dénué de tout défaut. Cette technologie possède en l'occurrence des imperfections qu'il nous faut saisir, en délimiter l'ampleur tout en évitant les erreurs d'interprétation et d'analyse (Lohr et Singer, 2016) afin de ne pas tomber dans le piège évident du sectarisme technologique.

Enfin, il nous faut encore nuancer les trois promesses du Big Data qui en partie se sont avérées être de véritables désillusions (Ollion et Boelaert, 2015) :

1. au niveau des promesses empiriques : les données recueillies semblent pour le moins superficielles ou en tout cas pas aussi riches que prévues. Ollion et Boelaert (2015) nous expliquent, à titre d'exemple, que les données récoltées concernant les modes de transport utilisés par les citoyens ne renseignent que très peu sur le comportement de ces derniers en matière de déplacements lorsque ces données ne sont pas recoupées par la suite avec d'autres éléments propres aux individus analysés. Selon ces mêmes chercheurs, un sondage traditionnel serait plus bénéfique. D'autant plus que l'utilisation de ces masses de données encombre le travail du chercheur par des procédures de nettoyage, de recodage, etc. assez contraignantes,

2. au niveau des promesses méthodologiques : l'exhaustivité n'étant pas toujours au rendez-vous, c'est une nouvelle désillusion qui s'offre aux chercheurs. L'abondance de données, nous précisent Ollion et Boelaert (2015), n'est pas toujours synonyme de plus de précision, car elle peut entraîner des erreurs de jugement comme ce fût le cas en 1936 lors des élections présidentielles aux États-Unis (cf. supra p.7),

3. au niveau des promesses théoriques : Ollion et Boelaert (2015) indiquent que les données en soi ne suffisent pas à l'apport de nouvelles sciences, il faudrait au préalable établir un cadre théorique avant de se lancer dans la recherche. Autrement dit, définir ce que l'on souhaite découvrir avant de récolter des données, et ne pas se contenter d'une recherche aveugle. Un avis partagé par Strong (2015) qui rajoute par ailleurs que nous évoluons dans un monde complexe, ce qui nous oblige dès lors à constituer un modèle théorique pour ne pas aboutir à de mauvaises interprétations dont l'apparence première ne reflèterait pas la réalité.

IV. Conclusion

À ce stade de notre étude, le lecteur aura eu l'occasion de rafraîchir ses connaissances sur l'industrie du sondage traditionnel (ou de le découvrir, selon son expérience antérieure de cet univers). Il aura par ailleurs obtenu une meilleure vision de ce que représente le Big Data. Aussi, les forces et faiblesses des deux outils ont pu indiquer au lecteur qu'aucune hégémonie n'était présente à l'horizon de part et d'autre. L'industrie du sondage et le Big Data souffrent en quelque sorte de défauts liés à leur mode de fonctionnement, mais cela ne leur empêche pas de disposer de véritables qualités dont la société peut tirer parti.

Par ailleurs, nous avons pu parcourir de manière assez théorique les différents domaines où la technologie Big Data a été utilisée. Ainsi, il est possible d'utiliser cet outil dans le secteur privé, le secteur public, le secteur médical et au sein du monde politique. Comme nous l'évoquions à la fin de chacun de ces domaines d'application, le Big Data représente à ce stade de notre réflexion une menace à prendre en considération par l'industrie du sondage, en gardant bien évidemment à l'esprit que cette technologie nous a démontré qu'elle était loin d'être infaillible.

À présent, il nous faut poursuivre notre démarche et déterminer comment cette technologie Big Data arrive dans la pratique à prendre le relais de l'industrie du sondage traditionnel. Ce qui nous permettra de mieux délimiter les menaces que cette nouvelle technologie représente pour l'industrie dans laquelle nous travaillons depuis quelques années.

Seconde partie : approche pratique

I. Introduction

Cette seconde section va nous permettre d'apporter au lecteur un regard pratique sur la thématique de notre étude. Concrètement, nous commencerons celle-ci en précisant notre question de recherche qui nous servira d'ailleurs de fil rouge tout au long de cette partie afin de ne pas dévier de notre trajectoire principale. Une fois notre question de recherche posée, nous formulerons des hypothèses (qui rejoindront peut-être celles du lecteur) qui seront par la suite infirmées ou confirmées.

Mais pour répondre à cette question de manière méthodique, nous avons décidé de nous inspirer de la méthodologie de recherche propre aux sondages traditionnels. Il s'agira dans un premier temps d'effectuer une recherche d'informations secondaires (desk research) et ensuite une recherche d'informations primaires au moyen d'une étude qualitative. La description complète de notre processus de recherche vous sera bien entendu présentée avant d'évoquer les résultats de ces deux types de recherche d'informations.

À la fin de cette section, le lecteur devrait être en mesure, selon nous, de disposer de tous les éléments de réponse à notre question de recherche sur base de témoignages d'experts, de professionnels, etc. Néanmoins, pour faciliter sa tâche, nous proposerons au lecteur, dans la troisième section de ce document, une synthèse critique et récapitulative basée sur les éléments théoriques et pratiques que nous avons récoltés au cours de notre étude.

1. Question de recherche

Le sondage traditionnel n'a eu de cesse de se renouveler au fil du temps, comme nous avons pu le constater dans la partie lui étant dédiée (cf. supra p.13). Parmi les nombreuses interrogations et les débats qui tournent autour de l'industrie du sondage, il nous semble légitime que certains spécialistes du domaine, comme Wyner (2017), s'interrogent sur la nécessité d'un renouveau de cette industrie. D'autant plus qu'à ce stade de notre travail, nous avons pu démontrer au lecteur qu'un rival de taille, le Big Data, venait empiéter sur les domaines de prédilection de l'industrie du sondage traditionnel.

Notre objectif sera dès lors de tenter de répondre à la question de recherche suivante : « Quel est l'impact du Big Data au sein des domaines de prédilection de l'industrie du sondage traditionnel, à savoir : le monde politique, le secteur privé, le secteur de la santé et le secteur public, et, dès lors, quelles sont les conséquences des apports du Big Data sur les activités de l'industrie du sondage traditionnel ? ». La réponse à cette question de recherche passera bien entendu par la mise en avant de cas concrets (des business cases) qui devront nous démontrer l'efficacité du Big Data dans ces situations particulières. Aussi, nous considérons qu'il faille impérativement compléter notre réponse par le biais d'une comparaison entre ces deux rivaux évoluant dans un terrain identique.

2. Hypothèses

Notre question de recherche étant posée, il nous est permis d'envisager à l'avance des éléments de réponse à cette question. Au vu des informations recueillies dans la première partie de notre travail, nous estimons dans un premier temps qu'il ne serait pas concevable d'imaginer que le Big Data soit l'instrument révolutionnaire venu détrôner à l'aide d'une technologie de pointe supérieure son prédécesseur, le sondage traditionnel. En effet, nous avons pu exposer un certain nombre de lacunes (cf. supra p.50) venant nuancer quelque peu les performances de cette technologie. Par ailleurs, nous sommes toujours chargé d'études au sein d'un institut de sondage, fonctionnant de manière traditionnelle, ce qui pour le moment confirmerait donc notre première hypothèse.

Dans un second temps, nous estimons que le Big Data, malgré ses faiblesses, impactent de manière significative l'industrie du sondage traditionnel. Nous pensons toutefois qu'il faille relativiser cet impact. Le Big Data, selon nous, apporterait des données inédites aux chercheurs, des données qui n'auraient pu être récoltées via la méthode traditionnelle. En d'autres termes, ces données présenteraient des réalités inobservées jusqu'alors. Certes, celles-ci seraient très intéressantes pour les demandeurs (secteur politique, privé, etc.), mais resteraient différentes des apports du sondage traditionnel. Dès lors, nous ferions face à deux outils fonctionnant de manière distincte, et qui tentent à leur manière de répondre le plus précisément possible aux questionnements de leurs demandeurs.

Enfin, de la même manière que l'éviction du sondage traditionnel de ses domaines de prédilection nous parait inconcevable, nous estimons qu'il en est de même de la non-intégration du Big Data dans ces mêmes domaines. Nous devrions de ce fait observer des formes de collaboration entre ces deux instruments (soit par l'exemple, soit de manière théorique). Le Big Data fournirait un élément de réponse à une question globale posée par les demandeurs, et l'industrie du sondage complèterait celui-ci ou viendrait infirmer ou confirmer l'élément de réponse fourni par le Big Data. Cette possibilité doit bien entendu être lue dans les deux sens, le Big Data pourrait lui aussi être chargé de confirmer ou d'infirmer une proposition de réponse émise par l'industrie du sondage traditionnel.

3. Méthodologie de recherche

Pour répondre à notre question de recherche, nous avons décidé de procéder en deux étapes. La première étape consistera à réaliser une « desk research », soit la recherche d'information secondaire. Comme nous l'avons brièvement évoquée plus tôt dans ce document (cf. supra p.15), l'information secondaire existe déjà. Il s'agit entre autres de documents, de statistiques sur la population étudiée, d'articles, etc. qui sont disponibles moyennant quelques recherches. Parallèlement à notre desk research, nous avons souhaité réaliser une enquête qualitative (qui fera office d'information primaire) auprès de la communauté des experts, des spécialistes du domaine qui nous concerne (Vandercammen et Gauthy-Sinéchal, 2014). Les points qui suivent résument les informations récoltées lors de notre recherche et détaillent de manière plus spécifique la méthode que nous avons employée pour atteindre notre objectif.

3.1. Desk research

Notre recherche d'informations secondaires sera composée de deux parties. La première intitulée « Études de cas » aura pour objectif de proposer au lecteur certains cas pratiques où la technologie Big Data a été employée et s'est révélée être utile pour le domaine d'application concerné ; ces différents cas entreront bien entendu en rivalité avec le sondage traditionnel ; les domaines d'application sont identiques à ceux de notre question de recherche, pour rappel : le secteur privé, le secteur public, le monde politique et le secteur de la santé. La seconde partie intitulée « Sondages traditionnels vs Big Data » se concentre, quant à elle, sur les conséquences qu'implique l'apport du Big Data sur cette industrie du sondage traditionnel ; nous évaluerons au travers du regard d'experts si le sondage traditionnel est voué à disparaître pour être remplacé par le Big Data ou si une collaboration reste possible entre les deux outils (et sous quelle forme ?).

a. Études de cas

a.1. Le cas Walmart

- ***Contexte***

Présente dans 28 pays, la multinationale, qui s'est spécialisée dans la grande distribution, possède à ce jour près de 12.000 magasins, évoluant sous 65 enseignes différentes, et emploie plus de 2 millions de collaborateurs (Walmart, 2018, para.1). L'objectif affiché de Walmart est clair. Il s'agit d'offrir à ses clients la meilleure expérience d'achat que ces derniers choisissent d'acheter en magasin, par Internet ou via mobile (Walmart, 2018). Une telle infrastructure nécessite forcément une gestion considérable de données issues de son environnement. Dès lors, il n'est pas étonnant, nous indique Marr (2016), que la société a depuis plusieurs années des vues sur le Big Data. Cette technologie, qui en était encore à ses prémices en 2004, leur a permis d'établir qu'à la suite d'un ouragan, au-delà de la vente naturelle de produits de première nécessité, le taux de vente d'une friandise à la fraise (proposée par Walmart) avait connu une augmentation de sept fois son taux de vente normal, nous précise Dillman, Chief Information Officer de Walmart en 2004 (citée par Hays, 2004). Depuis cet événement riche d'enseignement, Walmart a très largement investi dans les nouvelles technologies afférentes au Big Data (Marr, 2016). Par ailleurs, dès 2015, la multinationale envisageait la création d'un gigantesque cloud privé pouvant traiter pas moins de 2.5 pétabytes de données par heure (Marr, 2016, p.26).

- ***Problématique Walmart***

Le secteur de la grande distribution évolue dans un monde extrêmement compétitif. Chaque jour, ce sont des millions de consommateurs qui se rendent auprès de magasins pour acquérir leurs produits et services. L'enjeu pour la grande distribution, et donc Walmart, est de mettre à disposition le bon produit, au bon moment, au bon prix, et le cas échéant, accompagné du bon service à ses clients (Marr, 2016). Il est évident que pour une entreprise telle que Walmart, le défi logistique est de taille tant la société dispose de magasins disséminés à travers tous les États-Unis principalement, mais également auprès de 27 autres pays (Walmart, 2018, para.1). L'objectif est donc d'éviter coûte que coûte

qu'un client, ne trouvant pas chaussure à son pied, se tourne vers la concurrence (Marr, 2016).

– ***La solution Big Data***

Conscient de la valeur des données, Walmart décida en 2011 de créer un laboratoire de recherche dont la mission était de mieux discerner le comportement des consommateurs, afin de répondre plus efficacement à leurs besoins (Marr, 2016, p.27). Son nom ? Le Data Café. Cette plateforme permettait déjà de gérer, en temps réel, 40 pétabytes de données relatives aux transactions commerciales réalisées au cours des semaines précédentes, provenant de 200 flux de données internes et externes à la société (Marr, 2016, p.27). Pour les analystes Walmart, le temps c'est de l'argent, d'où l'affirmation de Peddamail, Senior Analyst chez Walmart (cité par Marr, 2016, p.27) : *« If you can't get insights until you've analysed your sales for a week or a month, then you've lost sales within that time. »*. Nous comprenons par l'intermédiaire de cette courte affirmation que le Big Data représente l'outil par excellence, car il permet de gérer en temps réel l'immense quantité de données provenant de différents points de vente, contrairement au sondage traditionnel qui nécessite quelques semaines pour la mise en place, la récolte et le traitement des données, sans compter que seule une infime partie des données serait traitée, avec une moindre précision.

Peddamail (cité par Marr, 2016) évoque quelques-unes des réponses apportées par le Big Data aux différents partenaires de Walmart. Un des cas rapportés est celui d'une équipe de vente qui ne parvenait pas à comprendre les raisons de la chute subite des ventes d'un produit spécifique. À la suite des recherches menées par le Data Café, il s'avéra qu'il s'agissait d'une erreur de prix. Après avoir adapté le prix du produit, les ventes seraient revenues à la normale dans un délai très court. Autre exemple d'application du Big Data, le cas de nouveau cookies vendus durant la période d'Halloween. Certains points de vente ne parvenaient pas à vendre ne serait-ce qu'un cookie, ce qui paraissait bien étrange. Les détecteurs ont permis en temps réel d'indiquer une anomalie auprès de la centrale. Une fois le signalement fait, les analystes décidèrent de contacter les points de vente éprouvant des difficultés à vendre ces cookies afin de les questionner sur leur inefficacité à vendre ceux-ci. L'enquête fut de courte durée ! En effet, ils découvrirent que ces nouveaux cookies ne figuraient tout simplement pas dans les rayons du magasin, mais restaient dans les stocks à la suite d'une erreur logistique.

Le Big Data permet également via le Social Genome Project de Walmart de prédire les achats des futurs consommateurs en analysant leurs conversations sur les réseaux sociaux. Autre outil d'analyse, le Shopycat Service de Walmart qui permet d'identifier le niveau d'influence qu'ont certains individus auprès de leurs amis, leurs connaissances en matière de comportement d'achat. En d'autres termes, à quel point une personne est-elle en mesure de pousser une connaissance à acheter tel ou tel produit (Marr, 2016). D'autres applications du Big Data sont mentionnées dans un article publié par l'équipe Walmart (2017). Au niveau de la demande de produits pharmaceutiques, les analyses de données permettent à la compagnie de déceler à quels moments de la journée il y a une plus grande

affluence devant ce département. Ce qui autorise la société à prévoir des renforts, à mieux gérer les plannings, etc. Des analyses similaires sont également effectuées pour gérer les autres départements et caisses des différents points de vente. Grâce à la granularité qu'autorise le Big Data, Walmart arrive à personnaliser l'expérience d'achat de ses clients, notamment en utilisant les données récoltées via les applications mobiles. Enfin, le Big Data permet aussi à la société d'accélérer le processus d'achat de sa clientèle en évaluant les meilleures alternatives en matière d'exposition de la marchandise sur ses étagères.

L'écosystème Big Data dont bénéficie Walmart est à l'image de la société, colossal ! Quotidiennement, comme nous le rappelle Van Rijmenam (2018), des millions de transactions commerciales sont réalisées. Ainsi, des térabytes de nouvelles données sont générés et ajoutés à un historique comportant des pétabytes de données. Toujours selon Van Rijmenam (2018, para.8), plus de 100 millions de mots-clés sont traités tous les jours par Walmart afin d'améliorer l'expérience client. La complexité de cet écosystème, reliant une multitude de données provenant de toutes parts, est représentée dans le schéma suivant :

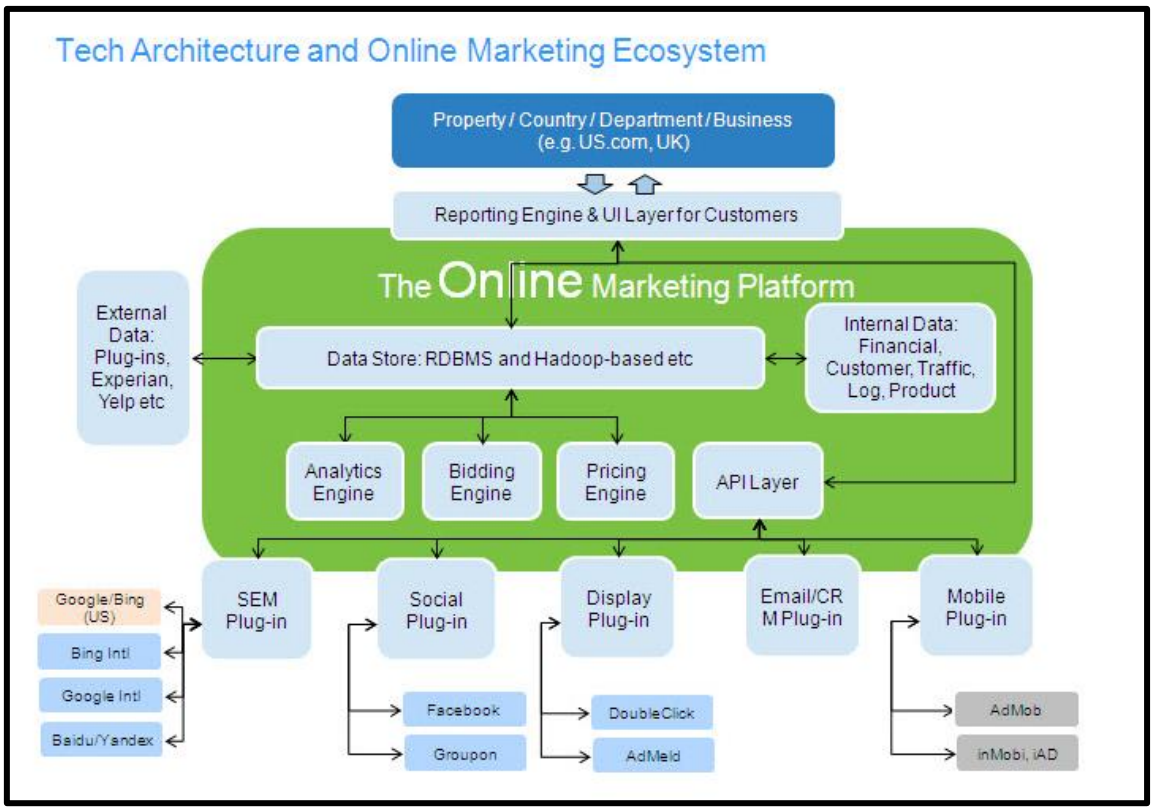


Figure 6 : L'écosystème Big Data de Walmart

Source : Van Rijmenam, M. (2018). *Walmart Is Making Big Data Part Of Its DNA*. Récupéré de https://datafloq.com/read/walmart-making-big-data-part-dna/509

Des données internes provenant par exemple des divers points de ventes, en passant par les données externes telles que celles récupérées à travers la toile, les réseaux sociaux, les applications mobiles, etc., nous pouvons imaginer les bénéfices de cette architecture technologique pour Walmart. Elle lui permet sans nul doute d'identifier des multitudes de corrélations, de collecter plus d'informations sur sa clientèle, et dès lors, améliorer l'expérience client.

- ***Conclusion***

Selon la multinationale Walmart (citée par Marr, 2016), l'implémentation du Big Data au sein de leur structure aurait permis à la société de réaliser d'énormes gains de temps. Concrètement, alors que par le passé environ deux à trois semaines étaient nécessaires pour identifier un problème et proposer une solution, seule une vingtaine de minutes serait aujourd'hui demandée par les nouveaux outils technologiques.

À la lecture de cet exemple, nous ne pouvons que constater l'apport manifeste du Big Data pour ce secteur d'activité. Dans un monde en perpétuel changement caractérisé par des millions de transactions commerciales, réalisées extrêmement rapidement, il devient difficile de penser que la grande distribution se passerait d'une source d'informations telle que le Big Data. Comme nous l'indique Bernard Marr (2016), cette technologie répond parfaitement à certains besoins de la grande distribution. Reste à savoir si l'industrie du sondage traditionnel a encore sa place.

a.2. Le cas King Faisal Specialist Hospital and Research Center

- ***Contexte***

Outre la volonté d'apporter à ses patients les meilleurs soins et les meilleurs services, le centre hospitalier de l'ancien roi saoudien est également connu pour être un centre de recherches bénéficiant des dernières technologies de pointe (King Faisal Specialist Hospital and Research Center, 2018). Si la réputation d'un hôpital se fait entre autres à travers la qualité de son personnel (médecins, infirmiers, etc.), le nombre et la qualité du matériel médical (lits, scanners, outils de diagnostic, etc.) et la qualité de ses rapports scientifiques, il convient néanmoins de disposer d'une bonne organisation au sein même de cet hôpital. Car en cas de dysfonctionnement organisationnel, la patientèle risquerait d'une manière ou d'une autre d'en subir les conséquences plus ou moins graves.

Pour pallier certaines problématiques issues du monde médical, nous avons pu observer au travers de différents articles que les hôpitaux ont, pour diverses raisons, de plus en plus recours à la technologie Big Data. Toutefois, Bresnick (2017) constate, à juste titre selon nous, qu'il est fort difficile de trouver une réelle politique Big Data au sein des hôpitaux ! Les raisons majeures seraient entre autres le manque de ressources ainsi que le temps d'implémentation que nécessite une solution Big Data, en tout cas selon une étude réalisée en 2017 par Dimensional Insight auprès de 104 professionnels travaillant dans des centres hospitaliers (Dimensional Insight, 2017, para.10). Ce ne sera bien entendu pas le cas du centre hospitalier de l'ex-roi saoudien qui, ne manquant pas de ressources, profite déjà des apports de cette nouvelle technologie.

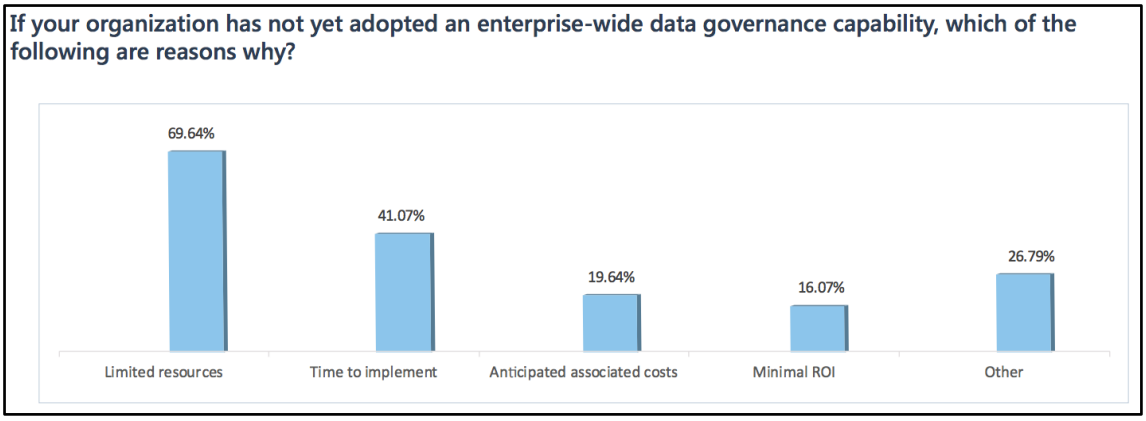


Figure 7 : Principales raisons de la non-intégration du Big Data en centre hospitalier

Source : Bresnick, J. (2017). *56% of Hospitals Lack Big Data Governance, Analytics Plans*. Récupéré de https://healthitanalytics.com/news/56-of-hospitals-lack-big-data-governance-analytics-plans

– ***Problématique du centre hospitalier***

Si dans le secteur privé, on accepte volontiers que « le temps, c'est de l'argent ! », on admettra tout autant, voire bien plus, que dans un hôpital « le temps, c'est la vie ! ». La vie étant par essence précieuse, la médecine a de tout temps évolué pour tenter de la préserver. Mais malgré toute sa bonne volonté, l'inévitable ne peut pas toujours être évité, et parfois pour des raisons autres que les compétences du médecin. C'est le cas au sein des urgences notamment où l'afflux massif de patients empêche significativement le bon fonctionnement de ce service qui, à l'origine, n'est destiné qu'à l'une ou l'autre catégorie spécifique de malades. Ce dysfonctionnement, très connu des centres hospitaliers, peut par moments empêcher ou ralentir le personnel soignant dans ses interventions auprès des patients nécessitant des soins immédiats, au profit (malheureusement) de ceux présentant bien moins d'urgence (ex. : cheville foulée, petits maux de ventre sans gravité, etc.). L'encombrement des urgences représente l'un des défis contemporains majeurs que se doit de relever le monde médical, comme nous le rappellent Khalifa et Zabani (2016).

Le centre hospitalier saoudien n'échappant pas à la règle, il a régulièrement été confronté à cette problématique des urgences. Pour résoudre celle-ci, il a été décidé dans un premier temps de scinder le problème en trois parties servant de base à la construction de trois indicateurs clés. Le premier indicateur précisait les temps d'attente aux urgences avant la prise en charge par le médecin. Le second indicateur portait sur le temps d'auscultation initial réalisé par le médecin avant que celui-ci ne statue sur le cas du patient. En d'autres mots, établir si oui ou non le patient doit être hospitalisé ou peut être renvoyé à son domicile. Enfin, le dernier indicateur mettait en exergue le temps de transport nécessaire entre le moment de l'admission du patient au sein de l'hôpital par le médecin, jusqu'au moment où le patient se trouvait effectivement sur son lit d'hôpital. Les trois indicateurs étant définis, la grille d'analyse servant à évaluer les performances du service d'urgence pouvait être mise en place. Cette dernière consistait par l'intermédiaire du Big Data à mettre en évidence les faiblesses au sein du processus de prise en charge du patient, à

proposer des améliorations possibles, à mettre en place une solution et enfin à contrôler les performances de cette solution (Khalifa et Zabani, 2016).

– *La solution Big Data*

Pour Khalifa et Zabani (2016), la prise de conscience des réalités auxquelles sont confrontées les sociétés (qu'elles soient privées ou publiques) et les raisons de l'existence de ces réalités ne sont plus suffisantes pour les organisations. Les objectifs sont à présent d'établir ce qui se déroule dans l'instant présent, de prévoir ce qu'il arrivera demain et de déterminer les séries d'actions à entreprendre afin d'être le plus efficacement possible préparé à la société de demain.

En vue d'améliorer sa gestion organisationnelle, le centre hospitalier saoudien a souhaité procéder en deux étapes. La première a logiquement consisté à collecter l'ensemble des données relatives aux urgences déjà existantes. Celles-ci étant stockées sous différents formats et à divers endroits devaient avant tout être rassemblées en un lieu commun avec une typologie commune. Une fois l'agrégation complétée, il a fallu procéder au nettoyage des données, à la validation et au traitement de celles-ci pour enfin entamer le processus d'analyse. L'objectif de cette analyse était d'identifier les variables qui présentaient une réelle et consistante significativité afin d'aider les gestionnaires du projet à proposer des solutions au centre hospitalier (Khalifa et Zabani, 2016) ; solutions qui figurent dans la seconde étape de cette étude saoudienne et que nous aborderons juste après avoir dressé le bilan de la première phase de cette étude.

À la suite du processus de nettoyage, 26.948 cas de patients s'étant présentés aux urgences de l'hôpital étaient disponibles pour les analystes (Khalifa et Zabani, 2016, p.760). Selon les auteurs, huit variables étaient communes sur l'ensemble des cas et étaient susceptibles de présenter des caractéristiques contribuant à l'encombrement des urgences. Il s'agissait entre autres du genre du patient, son âge, sa nationalité, le mode d'arrivée aux urgences, etc. Parmi ces huit variables, seules trois, statistiquement significatives, furent retenues en vue de mettre à jour, de manière chiffrée, les principales causes de contre-performance du service d'urgence de l'hôpital saoudien (Khalifa et Zabani, 2016), à savoir :

1. le niveau d'urgence du patient : cinq groupes distincts furent créés, allant du cas le plus grave (réanimation) au cas ne présentant aucune dangerosité (non-urgent),
2. le mode d'arrivée du patient : arrivée par ambulance, via la police, par un membre de la famille, etc.,
3. et l'âge du patient.

Le tableau présent à la page suivante met en évidence l'ensemble des résultats recueillis par les analystes durant l'année 2014 ; on y retrouve le nombre de patients total s'étant présenté au service d'urgence, le type de gravité présenté par le patient et l'admission ou non du patient dans le centre hospitalier (Khalifa et Zabani, 2016).

Code	Acuity level	Admitted patients	%	All ER patients	%	% of admitted to all
1	Resuscitation	95	2.6%	145	0.5%	65.5%
2	Emergent	913	24.8%	2470	9.2%	37.0%
3	Urgent	2636	71.5%	15,489	57.5%	17.0%
4	Less Urgent	38	1.0%	7575	28.1%	0.5%
5	Non Urgent	5	0.1%	1269	4.7%	0.4%
Total		3687	100%	26,948	100%	13.7%

Figure 8 : Taux d'admission en fonction du niveau d'urgence du patient

Source : Khalifa, M. et Zabani, I. (2016). Utilizing health analytics in improving the performance of healthcare services: A case study on a tertiary care hospital. Journal of Infection and Public Health, 9 (6), 757-765. doi : 10.1016/j.jiph.2016.08.016, p762.

Ce premier tableau nous démontre clairement que le service d'urgence de l'hôpital est assez mal compris de la population. En effet, sur l'ensemble des 26.948 patients s'étant présentés aux urgences, seuls 13,7% (soit 3687 patients) ont finalement été pris en charge, c'est-à-dire qu'ils ont été conduits dans une chambre d'hôpital après avoir reçu les premiers soins ou les premiers examens. Parmi les patients admis, 98,9% figuraient dans les catégories 1 à 3 (voir les pourcentages supra 4ème colonne), c'est-à-dire les cas les plus urgents ; assez logique, puisque les patients des catégories 4 et 5 ne présentent aucun risque d'aggravation de leur état de santé et ne nécessitent aucun soin spécifique pouvant être apporté par un centre hospitalier. Aussi, le tableau indique (voir les pourcentages supra 6ème colonne) que 32,8% de l'ensemble des patients (26.948) appartenaient aux catégories 4 ou 5, soit pratiquement un tiers de la population. Enfin, on observe que proportionnellement à leur catégorie respective les patients étant dans une situation plus critique étaient plus nombreux à avoir été admis ; ainsi, 65.5% des patients (voir les pourcentages supra 7ème colonne ; 65,5% = 95/145) affichant le niveau d'urgence le plus élevé ont été admis au centre hospitalier, contre 17% (=2.636/15.489) des patients de la catégorie 3 (Khalifa et Zabani, 2016, p.762).

Certes, nous constatons qu'une certaine forme d'incompréhension subsiste au sein de la population concernant les conditions pouvant amener à une prise en charge d'un service d'urgence. D'ailleurs, si cette étude avait été réalisée en Belgique, nous pensons que des résultats similaires ou du moins allant dans le même sens seraient ressortis d'une telle étude. Cependant, la venue de patients aux services d'urgence n'est pas dénuée de tout fondement. Les recherches menées par le centre hospitalier saoudien a mis en évidence que les patients qui choisissaient de se rendre au service d'urgence étaient confrontés à de longs délais de rendez-vous ou à des difficultés d'obtenir des soins de base. Dès lors, ils se rendaient aux urgences pour bénéficier d'un traitement immédiat et être rassurés quant à leur état de santé (Khalifa et Zabani, 2016). Cet état d'esprit que nous pouvons aisément comprendre ne change pas pour autant notre regard sur les conséquences directes qu'implique cette mentalité au sein d'un service d'urgence. Pour remédier à ce problème, les gestionnaires du centre hospitalier, et plus particulièrement de ce service, décidèrent d'implémenter certains changements du point de vue organisationnel, pour

ensuite en mesurer les performances (Khalifa et Zabani, 2016), ce qui nous conduit à la seconde phase de cette étude.

Dès 2015, le centre hospitalier décida d'allouer de nouvelles ressources connexes à son service d'urgence en aménageant quelque peu l'organisation de celui-ci. Brièvement, ce sont deux solutions qui ont été implémentées. La première, un outil de détection, devait permettre d'identifier très rapidement les patients faisant état d'un (très) faible niveau d'urgence (les catégories 4 et 5) ; 20% des lits du service d'urgence leur étaient destinés afin qu'ils puissent être dispensés des premiers soins par le service médical (Khalifa et Zabani, 2016, p.761). La seconde solution fut de faire patienter l'ensemble des patients, ne nécessitant pas de lit, dans une nouvelle salle d'attente ; pour la prise en charge de ces cas considérés comme moins urgents par l'hôpital, 2 médecins urgentistes furent présents 24h sur 24. L'objectif affiché de cette stratégie était d'utiliser les mêmes ressources physiques (le nombre de médecins et de lits sont volontairement restés inchangés) pour servir de manière plus efficace les patients du service d'urgence (Khalifa et Zabani, 2016).

Le tableau suivant indique le temps d'attente moyen avant la prise en charge par le service d'urgence (cf. supra premier indicateur p.61), trimestre après trimestre du premier trimestre 2014 au premier trimestre 2016.

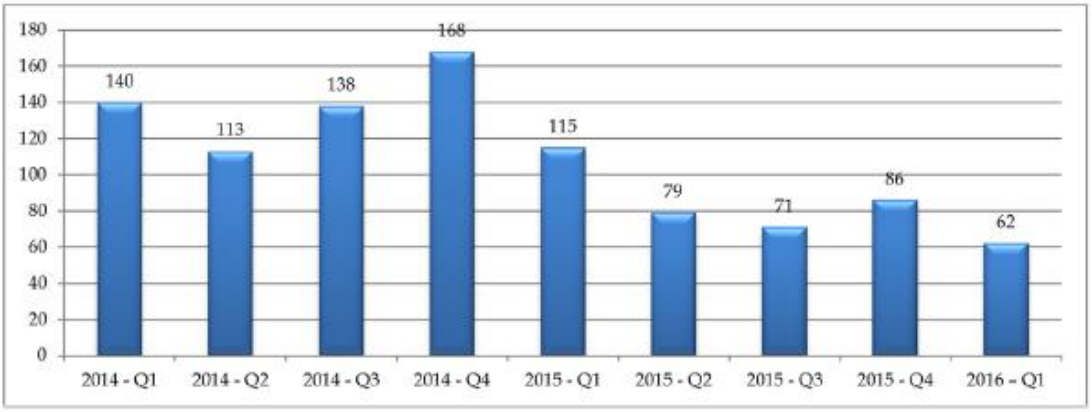


Figure 9 : Temps d'attente moyen en minutes, trimestre après trimestre (2014-2016)

Source : Khalifa, M. et Zabani, I. (2016). Utilizing health analytics in improving the performance of healthcare services: A case study on a tertiary care hospital. Journal of Infection and Public Health, 9 (6), 757-765. doi : 10.1016/j.jiph.2016.08.016, p.763.

Dès 2015, comme le graphique nous le montre, les améliorations se font ressentir. Nous observons en effet une nette diminution du temps d'attente moyen. En comparant les trimestres, non pas l'un après l'autre (car cela nous donne une tendance générale) mais de manière relative par rapport au trimestre de l'année précédente (premier trimestre 2015, par rapport au premier trimestre 2014), nous observons les améliorations suivantes : [voir page suivante]

Année		2014	2015	2016	Évolution
Nombre de minutes	Trimestre (1)	140	115	-	-18%
	Trimestre (2)	113	79	-	-30%
	Trimestre (3)	138	71	-	-49%
	Trimestre (4)	168	86	-	-49%
	Trimestre (1)	-	115	62	-46%

Tableau 3 : Mesure de l'amélioration du temps d'attente moyen en minutes

En ne considérant pas les comparatifs sur les deux premiers trimestres (où nous supposons qu'un temps d'habituation fut nécessaire pour l'implémentation du nouveau mode de fonctionnement, mais qui présentent déjà une forte diminution (respectivement 18% et 30%)), la diminution du temps d'attente fut en moyenne de l'ordre de 50%, et ce, rien que pour un seul indicateur. En agrégeant le temps d'attente moyen pour l'ensemble des trois indicateurs (pour rappel, il s'agissait du temps d'attente, du temps d'auscultation et du temps de transport (cf. supra p.61)), nous obtenons graphiquement les résultats suivants (Khalifa et Zibani, 2016) :

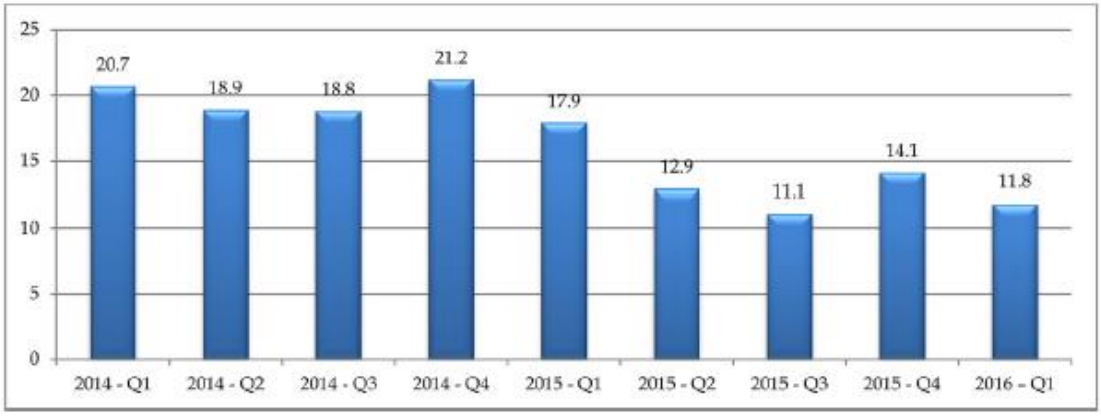


Figure 10 : Temps d'attente moyen en heures sur l'ensemble des trois indicateurs, trimestre après trimestre (2014-2016)

Source : Khalifa, M. et Zabani, I. (2016). Utilizing health analytics in improving the performance of healthcare services: A case study on a tertiary care hospital. Journal of Infection and Public Health, 9 (6), 757-765. doi : 10.1016/j.jiph.2016.08.016, p.763.

Visuellement, la tendance semble pratiquement similaire au graphique que nous avons commenté précédemment. Néanmoins, le niveau de performance est quelque peu plus faible sur les deux autres indicateurs (dont nous ne possédons pas les chiffres) ; le tableau présent à la page suivante démontre tout de même une nette amélioration du service d'urgence à la suite du changement organisationnel (une amélioration globale de plus de 30% en moyenne, hormis le premier trimestre).

Année		2014	2015	2016	Évolution
Nombre de minutes	Trimestre (1)	20,7	17,9	-	-14%
	Trimestre (2)	18,9	12,9	-	-32%
	Trimestre (3)	18,8	11,1	-	-41%
	Trimestre (4)	21,2	14,1	-	-33%
	Trimestre (1)	-	17,9	11,8	-34%

Tableau 4 : Mesure de l'amélioration du temps d'attente moyen en heures sur l'ensemble des trois indicateurs

- ***Conclusion***

Selon Khalifa et Zabani (2016), l'encombrement du service d'urgence représente l'un des facteurs impactant le taux de décès ainsi que le taux de complication chez les patients. La technologie Big Data a permis au centre hospitalier de réaliser un grand bond en avant en mettant en exergue les faiblesses de leur mode de fonctionnement. C'est également le Big Data qui a permis de détecter les variables responsables de cet encombrement et de suivre en continu l'évolution des performances du nouveau mode organisationnel. En d'autres termes, le Big Data a bel et bien sa place au sein du monde médical.

a.3. Le cas Transport for London (TfL)

- ***Contexte***

L'organisme de transport public londonien (Transport for London) a pour mission de mettre en œuvre le plan de mobilité décidé par le Maire de Londres, Sadiq Khan (Transport for London, s.d.). Au sein de ce gigantesque réseau, ce sont plus de 31 millions de voyages qui sont réalisés chaque jour par les résidents et non-résidents de la capitale du Royaume-Uni (Transport for London, s.d., para.3). Pour faire face à une telle demande, le réseau londonien s'est engagé dans plusieurs axes de développement. Ainsi, nous retrouvons des bus, des trams, le Tube (London Underground), des trains, des vélos, etc. Tous ces moyens de transport visent donc à améliorer la mobilité londonienne qui, selon la stratégie mise en place par Sadiq Khan, devra d'ici 2041 passer exclusivement (80%) via les transports publics (Transport for London, s.d., para.5). Pour parvenir à ces fins, Transport for London bénéficie du déluge de données en provenance des titres de transport scannés par les usagers. Il s'agit au vu du nombre de voyageurs de millions de données permettant ainsi au gestionnaire du réseau de retracer le parcours des usagers, d'identifier les zones à forte concentration (où le besoin en transport est forcément plus conséquent), les modes de transport privilégiés, etc. C'est donc en toute logique que le Big Data s'est imposé au sein de la société (Marr, 2016).

- ***Problématique Transport for London***

Gérer le quotidien d'une ville en matière de transport n'est pas chose facile, surtout si la population de cette ville s'accroit très rapidement. C'est le cas de la ville de Londres où

la population devrait selon les prévisions atteindre 10,8 millions d'individus en 2041 (Greater London Authority, Rapport, 2018, p.17). Il paraît évident que pour parvenir à mieux anticiper les mouvements de population et les besoins en matière de transport de celle-ci, la technologie, et plus particulièrement le Big Data, représente un des éléments essentiels pour capturer les données et les transformer en information (Sager Weinstein, 2017). Marr (2016) indique qu'en l'occurrence la collecte et l'analyse de données via la technologie Big Data devront permettre à la société de transport londonien de planifier les différentes gammes de services et d'informer comme il se doit les usagers du service public.

– *La solution Big Data*

Au travers des données récoltées par les titres de transport, l'Oyster Smartcard, la société de transport public londonien va être capable de tracer le quotidien des usagers. Cette carte électronique sans contact récolterait pas moins de 19 millions de données issues des multiples pointages des voyageurs (Marr, 2016, p.181). La granularité qu'autorise ce système permet donc d'identifier le déplacement de chacun des utilisateurs avec une précision inégalée auparavant, tout en préservant l'anonymat des individus. En outre, l'agrégation de ces données améliore la vision globale de la mobilité au sein de la capitale. En effet, si par le passé, l'oblitération de tickets jetables ne permettait pas aux analystes d'identifier les changements de ligne, de mode de transport, etc. réalisés par les usagers, il devient tout à fait possible aujourd'hui, par l'intermédiaire des cartes électroniques personnalisées, de suivre au jour le jour le comportement d'un seul et même individu (Marr, 2016).

La société de transport londonien a également fait usage d'un tout nouveau mode de collecte de données, à savoir les connexions wifi dans les stations du Tube. Durant cette première phase pilote, d'une durée de quatre semaines, 54 stations ont enregistré plus de 500 millions de demandes de connexion aux bornes wifi, provenant de 5,6 millions d'appareils mobiles, pour un total d'environ 42 millions de déplacement au sein du réseau londonien (Sager Weinstein, 2017, para.16). En faisant usage du Big Data, les analystes ont pu entrevoir le potentiel futur des données dont ils disposaient. Celles-ci leur ont déjà permis d'établir avec plus de précision le quotidien des voyageurs. Concrètement, ils ont été en mesure d'identifier les entrées et sorties des individus des différentes stations, le passage d'une station à l'autre, le changement de ligne, les sorties de train, les moments où les trains et les gares sont bondés, le taux d'occupation des différentes lignes ainsi que les contraintes faisant suite aux perturbations apparues au cours d'un voyage (Transport for London, 2017). Le tableau présent à la page suivante démontre le gain de précision apporté par cette nouvelle méthodologie en comparaison avec la méthode de récolte originelle (l'Oyster Smartcard).

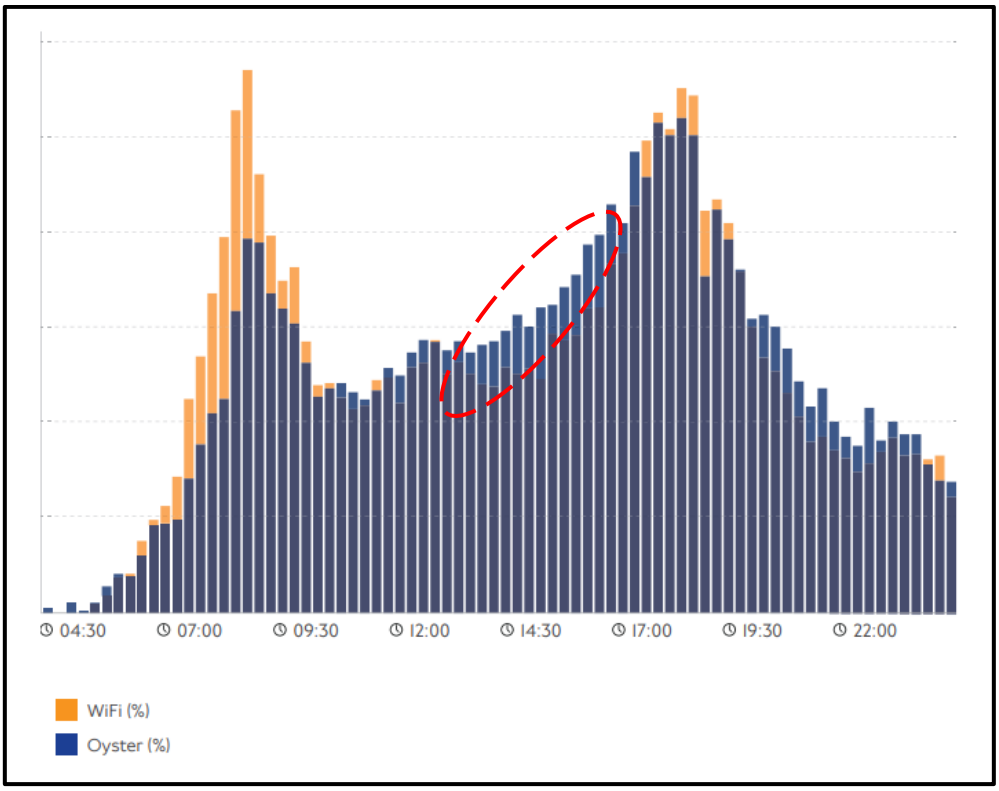


Figure 11 : Taux d'occupation de la station Oxford Circus à Londres

Source : Transport for London. (2017). *Review of the TfL WiFi pilot*. Récupéré le 7 juillet 2018 de http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf, p.25.

Le graphique précédent nous démontre l'efficacité de cette nouvelle méthode de récolte de données réalisée à l'aide des bornes wifi. Nous observons que globalement la majorité des données collectées via cette méthode est pratiquement identique à celles des données issues des titres de transports électroniques. Néanmoins, le wifi viendrait apporter une plus grande précision durant certaines heures, comme c'est le cas entre 07h00 et 9h30. Cela signifierait peut-être que durant ces heures les individus s'adonneraient à d'autres activités, comme par exemple l'achat de produits au sein des différents points de vente de la station. Mais selon le rapport publié par Transport for London (2017), l'Oxford Circus est surtout connu pour être une station où s'entrecroisent plusieurs lignes de Tubes. En tout cas, quelle que soit la raison, l'information reste très intéressante pour le gestionnaire du réseau qui doit veiller à une bonne fluidité au sein de chacune des stations. Enfin, nous constatons également un certain manque de précision de la part de la méthode wifi (cercle en rouge sur le graphique qui signale des barres bleu clair, précisant ainsi que les valeurs obtenues par la méthode wifi sont inférieures à la méthode Oyster) ; aucune information n'émane à ce sujet de la part de la société de transport public (s'agit-il d'une défaillance du système wifi ou bien d'usagers préférant utiliser la 4G/5G… Peut-être devrions-nous avoir recours au sondage traditionnel pour comprendre le phénomène ?).

Selon le gestionnaire de réseau londonien, la collecte de données au travers des bornes wifi afficherait une réelle utilité au niveau de la gestion de la gare et de la sécurité des usagers. Cette méthode étant plus précise (d'après Transport for London) peut montrer en temps réel la congestion d'une station et permettrait dès lors de prendre des mesures de précaution comme cela a été le cas le 30 Novembre 2016, aux alentours de 18h40, à la station Euston (Transport for London, 2017, p.38). Le gestionnaire découvrira qu'il s'agissait très probablement de la fermeture d'une station qui poussa les usagers à se rendre auprès de la station Euston afin d'atteindre leur destination finale (Transport for London, 2017) ; le graphique suivant nous montre le taux d'occupation de la station d'Eupen à la date du 30 novembre 2016 (en bleu) comparativement à la moyenne d'occupation de cette station durant l'année (en rouge).

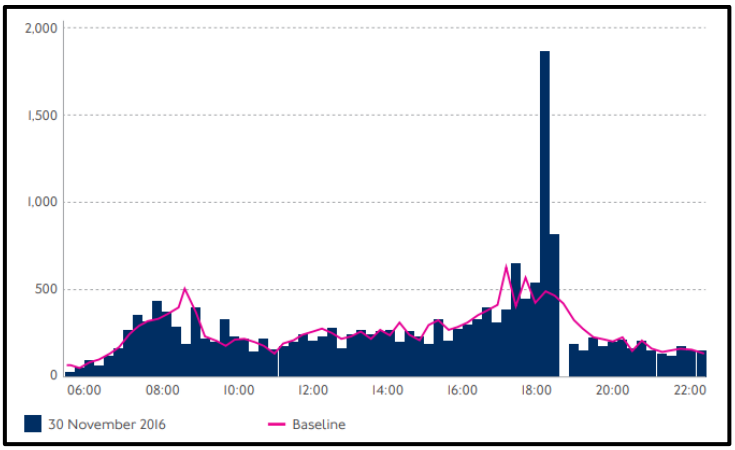


Figure 12 : Taux d'occupation de la station Euston à Londres

Source : Transport for London. (2017). *Review of the TfL WiFi pilot*. Récupéré le 7 juillet 2018 de http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf, p.39.

Lorsque le réseau du service de transport londonien est fortement perturbé, engendrant ainsi de longs retards pour les usagers, Transport for London a pour obligation de mettre en application sa politique de remboursement des voyageurs lésés. Cette mesure est, nous l'imaginons, très complexe à mettre en place, en plus d'être onéreuse. Le Big Data est donc une fois de plus une opportunité pour cette société qui, dans de telles circonstances, se doit d'une part d'identifier les personnes éligibles pour le remboursement (cela nous éloigne quelque peu de notre objet d'étude mais pas totalement, car nous pourrions imaginer que le sondage traditionnel soit utilisé pour mesurer le taux de personnes qui entamerait les démarches visant à être remboursé), et d'autre part, d'estimer précisément le comportement des usagers en cas de perturbation afin d'adapter le service de transport public. Cette dernière possibilité qu'autorise le Big Data permettrait à la compagnie de détecter, voire de prévoir, en temps réel les perturbations, et de proposer dans l'instant qui suit (ou de manière préventive) des alternatives aux usagers (Marr, 2016).

– ***Conclusion***

Par l'intermédiaire de cette nouvelle technologie, la société de transport londonien peut désormais avoir une vision plus globale et plus granulaire de l'environnement dans lequel

elle opère. Elle est à présent en mesure d'identifier avec précision le nombre d'individus utilisant quotidiennement leurs services, d'établir le niveau de congestion des gares, mettre en place des systèmes provisoires efficaces en cas de perturbations, etc. Sager Weinstein (cité par Marr, 2016) indique par ailleurs en guise d'exemple que le Big Data fut utilisé lors de la réparation d'un pont très fréquemment traversé par ses bus. Grâce à la technologie, Transport for London a su proposer des alternatives aux usagers qui pour la moitié se rendaient dans des zones à proximité du pont en question, selon Sager Weinstein (cité par Marr, 2016, p.182). Aussi, des messages personnalisés ont été envoyés aux usagers afin de les prévenir des répercussions causées par ces travaux nécessaires de la voie publique (Sager Weinstein, cité par Marr, 2016). Selon Marr (2016, p.182) 83% des usagers auraient émis un avis favorable ou très favorable quant à l'utilité de ce service d'informations personnalisé employé par la société de transport.

L'étude de ce cas pratique nous a démontré l'efficacité du Big Data au sein du secteur public. Les applications sont d'ores et déjà multiples alors qu'elles n'en sont qu'à leurs prémices. La technologie permet, sans irruption dans la vie privée des usagers (Sager Weinstein, 2017), de retracer le parcours de ces derniers à l'identique. Une information qui ne nous semble pas être possible de récolter à l'aide du sondage traditionnel. Dans cette dernière méthode, nous pourrions au mieux retracer les chemins habituels, mais quid des chemins alternatifs empruntés par les usagers ? Quid du temps resté dans une gare pour telle ou telle raison (un usager pourrait s'être souvenu d'être resté plus longtemps qu'à l'accoutumée au sein d'une gare, mais combien de temps exactement ?) ? Quid du taux de congestion d'une gare ? Etc. Les exemples qui ont été exposés ne représentent bien entendu qu'un sous-domaine d'application de cette technologie dans le secteur public, mais ils démontrent en tout cas un apport indéniable du Big Data pour les sociétés exerçant une activité dans ce domaine.

a.4. Le cas politique

– *Contexte*

Est-il possible pour un candidat à la présidentielle de faire usage du Big Data pour estimer le nombre d'individus qui voteront pour lui ? Pourrait-on utiliser cette technologie pour analyser le profil des votants de tel ou tel parti ? Serait-il envisageable d'influencer d'une manière honnête le vote d'électeurs en établissant quelles sont les demandes spécifiques de ceux-ci en matière de politique sociale, économique, environnementale (et autres) ? Ces questions, ainsi que bien d'autres qui gravitent autour du monde politique, pourraient éventuellement trouver une réponse auprès des algorithmes qu'utilise le Big Data, mais avec quel degré de précision ?

Pour cette étude de cas dans le domaine politique, nous avons volontairement exclu le cas de Cambridge Analytica en raison de l'immoralité de la pratique ; nous reviendrons tout de même brièvement sur cette affaire, qui a fait la une de plusieurs journaux, lors de la conclusion de ce cas pratique pour signaler les dérives possibles du Big Data. Aussi, nous n'avons pas voulu faire usage de l'étude de cas proposée par la société Contemporary Analysis qui indique avoir aidé un candidat (possédant de faibles ressources) lors d'une

campagne électorale en l'informant sur les axes prioritaires de travail (Contemporary Analysis, 2018) ; cette étude de cas ne comporte aucune donnée technique que nous pourrions utiliser pour évaluer le potentiel réel du Big Data en matière politique.

Notre étude de cas présente une application pratique du Big Data qui vise à démontrer les qualités de cette méthode d'étude au niveau politique. Néanmoins, à aucun moment les résultats de cette étude n'ont aidé d'une manière ou d'une autre un candidat lors d'une campagne électorale (ou autres applications politiques). Il s'agit d'une étude réalisée par trois chercheurs de l'Université de Cambridge qui, comme nous allons le voir, ont tenté d'utiliser les données d'un réseau social pour établir un profil sociodémographique, politique et comportemental des internautes.

– ***Problématique***

Comme nous l'avions indiqué précédemment (cf. supra p.46), l'une des qualités du Big Data est de s'immiscer dans l'intimité des internautes, et d'ainsi récolter des données les concernant. Nous savons également, comment nous le rappellent Kosinski, Stillwell et Graepel (2013), que bon nombre d'activités contemporaines font un détour obligatoire par la toile. Selon le comportement digital des individus, nous pouvons trouver ici et là des publications sur les réseaux sociaux (photos, images, etc.), des commentaires, des réactions, des inscriptions à certaines pages ou sites, les mentions du sexe et de l'âge de l'individu, etc. Pour autant, serait-il imaginable d'identifier le profil des individus pour prédire quel serait le vote de ceux-ci lors des prochaines élections ? C'est en tout cas ce qu'ont tenté de réaliser les chercheurs dont l'objectif n'était pas seulement d'identifier la tendance politique de l'internaute, mais de dresser un profil plus large de celui-ci (l'étude pourrait nous indiquer si le procédé utilisé serait utile politiquement ; ce qui serait le cas si les profils sociodémographiques et politiques sont proches de la réalité).

– ***La solution Big Data***

Par le biais de cette étude les chercheurs ont souhaité prouver qu'il était tout à fait possible de définir le profil des utilisateurs du réseau social Facebook en ne faisant usage que des likes (mention « j'aime ») réalisés par ceux-ci. De manière générale, les likes sont utilisés pour de multiples raisons, comme par exemple, le fait d'aimer un contenu audio, une vidéo, un restaurant, un auteur, des livres, un artiste, les photos publiées par un ami, etc. Selon les chercheurs qui ont mené cette étude, les informations issues de ces likes seraient pratiquement identiques à celles récoltées par les algorithmes des moteurs de recherche (tels que ceux de Google) qui peuvent démontrer la tendance, l'attractivité d'un sujet, d'un artiste, d'un chanteur, etc. Les likes présenteraient également des similitudes avec les sites web tels que YouTube (Kosinski, Stillwell et Graepel, 2013) ; la tendance observée en termes de vues sur ce site hébergeant des millions de vidéos serait dès lors à l'image de la tendance observée au niveau des likes de Facebook.

Pour vérifier leurs dires, cette équipe de chercheurs a donc récolté l'ensemble des likes de 58.000 utilisateurs du réseau social Facebook (Kosinski, Stillwell, et Graepel, 2013, p.1) ; ces derniers étaient bien entendu d'accord pour participer à cette recherche qui, pour

rappel, n'avait aucune finalité politique. Par ailleurs, pour établir l'exactitude de leurs observations, il a été demandé aux répondants de compléter des questionnaires en ligne ainsi que des tests de psychomotricité et de fournir des précisions sociodémographiques (âge, sexe, religion…) (Kosinski, Stillwell et Graepel, 2013) . Le tableau ci-dessous présente les premiers résultats de cette étude.

Variables analysées	Niveau de précision
Caucasian vs. African American	95%
Gender	93%
Gay	88%
Democrat vs. Republican	85%
Christianity vs. Islam	82%
Lesbian	75%
Smokes Cigarettes	73%
Drinks Alcohol	70%
Single vs. In Relationship	67%
Use drugs	65%
Parents together at 21	60%

Tableau 5 : Taux de précision des likes Facebook sur le profil des utilisateurs

Source : Kosinski, M., Stillwell, D. et Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110 (15), 5802-5805. doi : 10.1073/pnas.1218772110, p.2.

Comme nous pouvons l'observer dans le graphique précédent, certains des résultats issus des likes du réseau social sont extrêmement proches de la réalité provenant des données exactes fournies par les sujets de cette étude (via les questionnaires, tests de personnalité, données sociodémographiques, etc.). Ainsi, les likes semblent suffire à l'identification du groupe ethnique de l'internaute (95%), de son sexe (93%), de son orientation sexuelle (88%, s'ils sont des hommes) ainsi que sa couleur politique (85%). Même si l'objet de l'étude ne présentait aucune finalité politique, nous pourrions imaginer que, en faisant usage d'un tel outil, les analystes pourraient découvrir des corrélations entre les électeurs d'un parti politique et leur origine ethnique, leur confession religieuse ou d'autres caractéristiques qui leur sont propres. Cependant, les seuls likes ne seront selon nous pas suffisants pour rivaliser avec l'industrie du sondage ; les algorithmes doivent également être en mesure d'identifier les publications et commentaires ainsi que d'autres données digitales, et ce, au sein d'un maximum de réseaux sociaux, étant donné que, par exemple, certains individus sont plus actifs sur Twitter que sur Facebook.

Le graphique suivant expose d'autres résultats issus de cette même étude et qui relativise les performances de ce type d'étude fonctionnant uniquement sur base de likes :

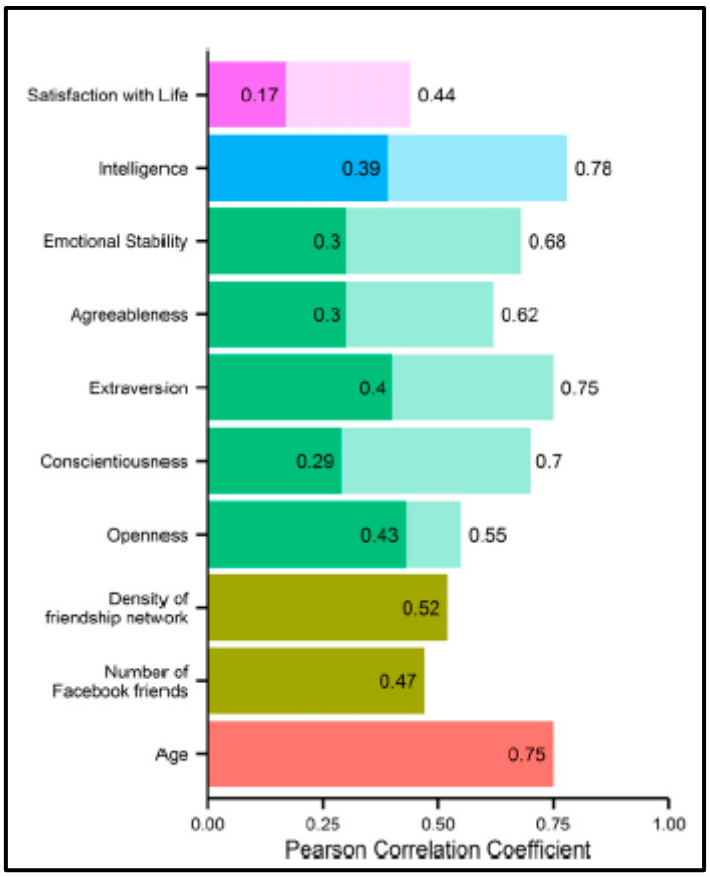


Figure 13 : Taux de précision des likes Facebook sur la personnalité des utilisateurs

Source : Kosinski, M., Stillwell, D. et Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110 (15), 5802-5805. doi : 10.1073/pnas.1218772110, p.2.

Dans le graphique qui précède, nous pouvons observer qu'hormis le critère d'âge (75%), les autres variables sont fort éloignées de la réalité (satisfaction quant à leur qualité de vie, stabilité émotionnelle, etc.) ; les critères d'ordre purement psychologique ne peuvent être calculés précisément à l'aide de tests psychométriques, d'où les barres transparentes qui indiquent le degré de certitude quant à ce critère. Nous constatons néanmoins que seuls 12% séparent le trait de caractère « Openness » (soit le fait d'être ouvert) des estimations des tests psychométriques réalisées parallèlement à cette étude.

– *Conclusion*

La seule utilisation de likes ne s'avère pas être une menace pour l'industrie du sondage. Il est en effet clair qu'une technique aussi simpliste est insuffisante pour présenter un réel intérêt chez les politiciens. Cependant, les chercheurs ont démontré par le biais de cette étude que les traces digitales que nous laissons derrière nous sur la toile sont susceptibles d'apporter des indices quant à notre orientation politique ainsi que de nombreuses autres caractéristiques (Kosinski, Stillwell et Graepel, 2013). Si d'aventure les algorithmes

analysaient plus largement notre activité sur Internet en croisant des données émanant de multiples sources, il deviendrait probable pour une société d'identifier avec une assez grande précision notre couleur politique, notre profil sociodémographique ainsi que d'autres comportements qui nous sont propres, afin par exemple d'aider un politicien à proposer des réformes préélectorales pour convaincre les électeurs, ce que prétend faire Contemporary Analysis (2018).

– ***Complément : l'affaire Cambridge Analytica***

Dans un entretien vidéo accordé au journal britannique The Guardian (2018, 17 mars), Christopher Wylie, lanceur d'alerte, ancien data analyst de la firme Cambridge Analytica, délivre des éléments plus que troublants concernant les pratiques de son ex-employeur. L'objectif des responsables de la société était avant tout d'apporter des changements au niveau politique. Pour ce faire, il fallait (selon la mentalité des gestionnaires) d'abord changer la culture de la population ; changement culturel qui ne pouvait se concrétiser autrement que par la modification de l'état d'esprit des individus. Dès lors, l'idée fut de développer un outil capable de collecter un maximum d'informations sur la population afin de mieux cerner chacune des mentalités de manière individuelle.

L'outil en question était une application Facebook qui, une fois utilisée par l'internaute, autorisait Cambridge Analytica à accéder à l'ensemble des activités Facebook tels que les commentaires, publications, likes, ainsi que certains des messages privés des utilisateurs. Ainsi, en l'espace de quelques mois, ce sont près de 10 millions d'individus qui ont fait usage de cette application et qui ont à leur insu transmis toutes leurs données à Cambridge Analytica ; et comme un malheur n'arrive jamais seul, ces 10 millions d'internautes lésés ont infecté l'ensemble de leurs contacts, ce qui portera le nombre total de profils récoltés à 50 millions d'internautes (Wylie, 2018, 17 mars). Enfin, une fois les données traitées et analysées, la société créa de nouveaux sites web spécifiquement dédiés aux différents profils pour tenter de les influencer (concrètement, à voter pour D. Trump). Toutefois, Wylie (2018, 17 mars) ne peut confirmer ou infirmer la réalité de cette influence.

b. Sondages traditionnels vs Big Data

Après avoir passé en revue différents cas pratiques d'utilisation du Big Data (entrant en concurrence avec l'industrie du sondage traditionnel), il convient à présent de mesurer les répercussions causées par cette nouvelle technologie. Bien évidemment, calculer ce type d'impact nécessite des données chiffrées qui exprimeraient l'évolution du marché du Big data et celui de l'industrie du sondage, et qui de surcroît devraient démontrer que le ralentissement observé au sein de l'univers du sondage traditionnel serait dû à l'arrivée d'un nouveau concurrent, le Big Data. Ce type d'observation ne nous est en l'occurrence pas accessible. Toutefois, nous pourrions de manière indirecte établir si la technologie Big Data a changé l'état d'esprit des experts (d'un bord comme de l'autre) qui, par exemple, pourraient estimer que les apports de cette technologie seraient largement supérieurs à ceux du sondage traditionnel. D'où nous pourrions déduire que l'impact serait déjà assez conséquent et que l'avenir ne serait pas du tout reluisant pour l'industrie

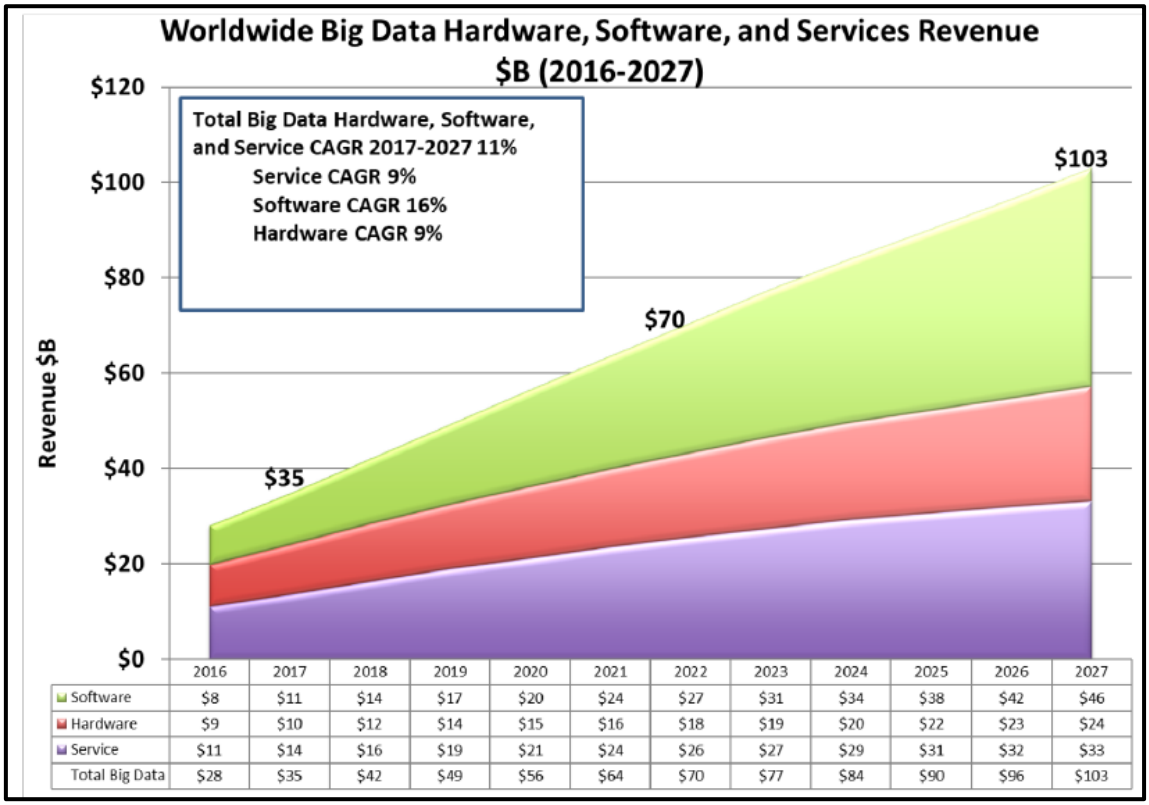
du sondage traditionnel. Pour éclaircir cette zone d'ombre, nous avons décidé de parcourir la toile à la recherche de différents experts ayant déjà publié leur vue sur cette thématique.

b.1. Vers une disparition du sondage traditionnel ?

Pour Devault (citée par Van Rijmenam, 2018), il n'y a pas le moindre doute concernant l'avenir du sondage traditionnel. Pour cette spécialiste, il conviendrait même de mettre toutes les études comportementalistes aux oubliettes. Un point de vue pour le moins radical qui part du principe suivant : les chercheurs n'ont plus aucune réelle utilité à s'attarder sur la compréhension de l'être humain et des raisons qui le poussent à agir de telle ou telle façon, il leur suffit de configurer des algorithmes qui récolteront en continu les activités de l'Homme et restitueront une image fidèle de celui-ci. Van Rijmenam (2018) rajoute pour sa part que la technologie Big Data est venue remettre en question l'existence même du sondage traditionnel. Une industrie qui devrait en tout état de cause disparaître de notre horizon, selon ce spécialiste, à moins qu'elle ne décide de se réformer ; une réforme qui passerait par le recrutement de data analysts afin de proposer les mêmes services que le Big Data (une explication qui ne nous semble pas être une réforme, mais une mutation, un abandon de l'outil traditionnel au profit du Big Data).

Pour sa part, Daboll (2013) énumère cinq raisons qui expliqueraient la fin prochaine de l'univers du sondage traditionnel qui, selon cet expert, se focaliserait trop sur le mode de collecte de données, alors que le Big Data cogiterait déjà sur le mode d'intégration des données (qu'il possède déjà) et l'analyse de celles-ci. La première raison serait la taille de l'échantillon et de la problématique analysée. L'échantillon serait selon lui soit trop faible par rapport à certaines thématiques étudiées, soit suffisant mais qui, dès lors, empêcherait une vision plus globale de l'environnement. La seconde raison serait le manque de pertinence de cet outil désuet. Le temps demandé par le sondage traditionnel n'est pas en mesure de répondre aux besoins des départements marketing qui souhaiteraient obtenir des informations concrètes dans l'immédiat, sans attendre le processus long et fastidieux du sondage traditionnel. Par ailleurs, la longueur de nombreux questionnaires laisse à désirer. Ceux-ci peuvent prendre jusqu'à une heure et ne tiendraient pas compte du public faisant objet de l'enquête qui n'a pas autant de temps à consacrer à de tels entretiens. La troisième raison serait l'impossibilité de gérer la complexité. Les sondeurs, en raison des outils qu'ils ont à leur disposition, s'accommoderaient de réponses sommaires pour tenter de définir un monde complexe. La quatrième raison renvoie aux compétences des employés en matière de traitement des données. Ces derniers ne possèderaient pas toutes les qualités requises pour effectuer des analyses aussi complexes que celles réalisées par les data analysts. Enfin, la dernière raison serait le manque d'engouement de l'industrie pour le changement. En d'autres termes, ils ne s'adapteraient pas à la demande de leur nouvel environnement qui, en matière de technologie, a considérablement évolué. Néanmoins, l'auteur reconnaît qu'actuellement le Big Data éprouve certaines difficultés, que des erreurs d'interprétation ne sont pas impossibles, mais qu'à la différence de l'univers du sondage traditionnel, le Big Data devrait trouver une issue favorable à sa problématique.

Selon Hing Lo (2016), aucun doute n'est à émettre quant à l'avenir du Big Data qui bénéficie d'un réel ancrage au sein de la société. Pour preuve, Hing Lo (2016, para.3) évoque le rapport émis, en 2016, par la société PWC qui indique que les activités d'analyse de données avaient augmenté de 350% entre 2012 et 2015. Cette tendance peut aussi être démontrée à l'aide des prévisions du chiffre d'affaires du Big Data pour les années à venir, comme nous l'indique le graphique suivant :



	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027
Software	$8	$11	$14	$17	$20	$24	$27	$31	$34	$38	$42	$46
Hardware	$9	$10	$12	$14	$15	$16	$18	$19	$20	$22	$23	$24
Service	$11	$14	$16	$19	$21	$24	$26	$27	$29	$31	$32	$33
Total Big Data	$28	$35	$42	$49	$56	$64	$70	$77	$84	$90	$96	$103

Figure 14 : Prévision du chiffre d'affaires du Big Data 2016-2027

Source : Kobielus, J. (2018). *Wikibon's 2018 Big Data Analytics.* Récupéré le 25 juillet 2018 de https://wikibon.com/wikibons-2018-big-data-analytics-trends-forecast, p.3.

Comme l'indique le tableau précédent, le chiffre d'affaires du Big Data afficherait un taux de croissance à deux chiffres (11%) pour les dix prochaines années (entre 2017-2027). En 2018, nous devrions atteindre près de 42 milliards de dollars de chiffre d'affaires tous types de revenus confondus (softwares, hardwares et services), et celui-ci serait de 103 milliards en 2027, soit une augmentation de près de 250%. La production de softwares engrangerait la plus grande part du chiffre d'affaires en 2027 (45% environ, soit près de la moitié des revenus), alors que les services représenteraient près d'un tiers du chiffre d'affaires. L'importante augmentation en softwares (16% sur les dix prochaines années) indiquerait selon nous que de plus en plus d'entreprises feront appel à cette technologie (qui serait utilisée en interne) et que de nouvelle start-up verront le jour proposant aux sociétés l'outil Big Data ; ce qui, par la même occasion, expliquerait l'augmentation des revenus pour les services (9% sur les dix prochaines années).

– ***Critique***

Pour Crawford (2013), il est illusoire d'imaginer que le simple amoncellement de données puisse être la clef de tous les secrets. Cette manière de penser mènerait d'ailleurs vers une forme de fondamentalisme, selon elle, qui revendiquerait que l'analyse de données est en mesure de présenter en toute objectivité la vérité vraie. Pour étayer son point de vue, cette spécialiste nous rappelle le cas Twitter. Lors de l'ouragan Sandy qui frappa de plein fouet une partie des États-Unis le 29 octobre 2012 (Le Monde.fr, 2013), près de 20 millions de tweets avaient été postés sur le réseau social entre le 27 octobre et le 1er novembre (Crawford, 2013, para.3) de cette même année. L'activité identifiée au travers des tweets ne reflétait pas réellement la réalité. Selon cette experte, la proportion de tweets en provenance de Manhattan était bien trop élevée par rapport à d'autres villes plus sévèrement touchées, ce qui par conséquent tronquait la vision globale de l'événement. Par ailleurs, dans de nombreuses zones traversées par l'ouragan, les habitants étaient à court d'électricité, ce qui par conséquent empêchait ces derniers de charger leur smartphone, ordinateur, etc. pour envoyer des tweets. Aussi, la comparaison de l'activité sur Twitter et sur Foursquare, application de géolocalisation (Foursquare, 2018), montrait, comme nous allons le voir, quelques divergences non-négligeables (Crawford, 2013).

Pour Grinberg, Naaman, Shaw et Lotan (2013), lorsqu'une catastrophe de l'ampleur de l'ouragan Sandy touche une zone habitable, des faits observables liés à cet événement se répercutent sur les réseaux sociaux. Pour appuyer leurs dires, ces spécialistes ont analysé les activités avant et pendant l'ouragan sur les applications Twitter et Foursquare. À la suite de leurs observations, ils constatent qu'une divergence significative apparaît entre les deux réseaux sociaux lorsque la population est soumise à des conditions inhabituelles. Le graphique présent à la page suivante expose visuellement ces divergences d'activités. Avant la tempête, l'activité sur les deux réseaux sociaux sont relativement similaires, mais dès l'étape de préparation de gros écarts d'activités se font ressentir. On observe des pics d'activités sur Foursquare entre le 27 et le 30 octobre 2012 (cf. infra graphique p.78 : « grocery shopping » et « shopping »), alors que les tweets ne suivent pas du tout cette tendance. Aussi, au moment où la tempête frappe la ville de New-York, un décalage subsiste durant la nuit entre l'application Foursquare et l'application Twitter (en léger retard), ainsi qu'un pic d'activité au sein des magasins pour Foursquare non-suivi par Twitter. C'est pour ces raisons que ces spécialistes soulignent l'importance de croiser les données afin d'obtenir des informations plus précises sur les comportements humains et tenter d'éviter ainsi toutes formes d'erreurs d'interprétation.

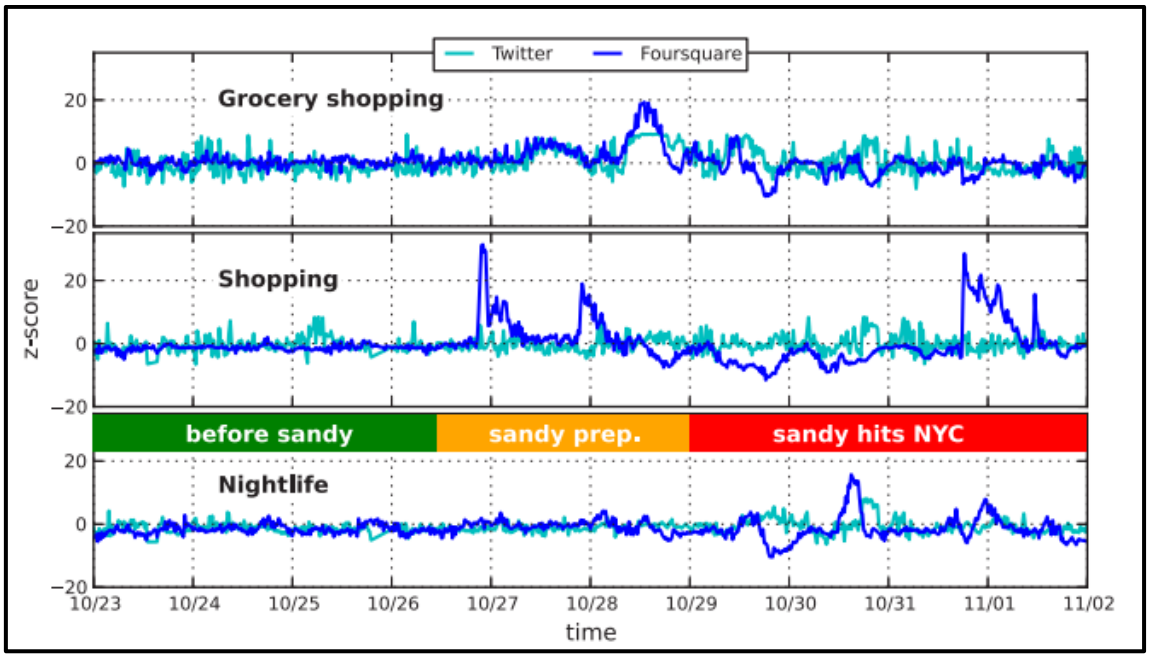


Figure 15 : Taux d'activité sur Twitter et Foursquare avant et pendant Sandy

Source : Grinberg, N., Naaman, M., Shaw, B., et Lotan, G. (2013). *Extracting Diurnal Patterns of Real World Activity from Social Media.* Récupéré le 8 août 2018 de https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6087/6359, p.213.

Enfin, nous évoquerons ici un dernier élément de preuve, selon nous, qui tend à démontrer que certains analystes obnubilés par le Big Data commettent des erreurs d'interprétation patentes. C'est le cas notamment de Stephens-Davidowitz (2017) qui affirme par exemple que, à la suite d'un attentat terroriste, le taux d'anxiété au sein de la population ne varierait pas du tout. Sa méthodologie fut d'utiliser le moteur de recherche Google, société pour laquelle il a travaillé en tant qu'analyste, pour évaluer le nombre de fois que le terme « anxiété » ainsi que tous les synonymes y afférant étaient utilisés par les internautes avant et après un attentat majeur tel que ceux de Paris ou de Bruxelles (que ce soit en Europe ou aux Etats-Unis) ; nous souhaitons également préciser que ce spécialiste part du principe que, à la suite de tels événements, les individus se précipiteraient sur leur clavier ou leur smartphone pour écrire le terme anxiété (ou autres synonymes), un raisonnement qui se répercute bien entendu sur ses interprétations. Les résultats obtenus étaient, selon lui, si univoques qu'ils représentaient à eux seuls une preuve suffisante pour juger de l'état d'esprit de la population ; une population qui ne souffrirait donc d'aucune crainte quant à sa sécurité.

La simple lecture du court passage retrouvé dans le livre « Everybody lies » de Stephens-Davidowitz (2017) nous a pour le moins laissé assez dubitatif, car dans nos souvenirs, lors des attentats de Bruxelles, la détresse était assez palpable au sein de la population. Les reportages télévisés, les journaux, les radios, les témoignages, la présence de militaires et de policiers nous rappelaient sans cesse l'état d'urgence dans lequel nous étions plongés, le drame que nous avions vécu au cœur de notre capitale. Si une telle situation n'avait eu aucun impact sur notre niveau d'anxiété, nous serions en droit de nous

poser des questions sur notre humanité (« Serions-nous devenus des surhommes, voire des machines incapables de ressentir la peur ? »).

Fort heureusement, Dedicated avait réalisé deux études, sous la direction de Dumoulin et Sterckmans, auprès de la population belge un an avant et un an après les attentats de Bruxelles qui avaient en partie pour objectif de mesurer l'état d'anxiété dans lequel se trouvait la population ; pour information, ces 2 études avaient été réalisées sur un échantillon représentatif de près de 3.000 Belges (Baromètre politique trimestriel, 2017, p.22). Les résultats obtenus sont en contradiction avec les affirmations de Stephens-Davidowitz (2017). Il apparait en effet que 77% des Belges estimaient que leur pays était « assez » ou « très » fortement exposé à la menace terroriste en mars 2017 (soit tout juste un an après les attentats), contre 69% en mars 2015 (Dumoulin et Sterckmans, Baromètre politique trimestriel, 2017, p.22). Une augmentation de 8% qui ne prouve pas directement que le taux d'anxiété a augmenté au sein de la population, mais qui peut de manière indirecte nous indiquer une prise de conscience de la part des citoyens qui seraient davantage sur leurs gardes, plus craintifs. Pour estimer plus concrètement le niveau d'anxiété, il suffit d'analyser les chiffres concernant les comportements adoptés par la population à la suite de cet événement anxiogène ayant touché l'aéroport de Zaventem et le métro bruxellois. Nous apprenions ainsi que 23% des Belges indiquaient éviter de se rendre au sein des grandes villes et des lieux publics en 2017, contre 12% en 2015 ; il en va de même pour les transports en commun pour lesquels 11% de la population mentionnaient éviter (ou moins utiliser) les transports en commun en 2017, alors qu'ils n'étaient que 5% par le passé ; les Belges adopteraient ainsi tout un tas d'autres comportements à caractère préventif (Dumoulin et Sterckmans, Baromètre politique trimestriel, 2017, p.25). Les résultats de cette étude nous semblent bien plus réalistes, plus cohérents que ceux évoqués plus tôt via la méthode Big Data. Ce type d'erreur commis par les adeptes des algorithmes est dû selon nous à une méthodologie athéorique (cf. supra p.52) qui part du principe que l'identification d'un événement, d'une corrélation entre des faits, etc. est suffisante pour extrapoler des vérités, sans qu'aucune distance critique ne soit prise.

b.2. Vers une collaboration : sondage traditionnel – Big Data ?

À ce stade de notre réflexion, il nous semble peu probable de voir disparaître le sondage traditionnel, en tous cas pas dans un futur proche. Nous avons pu le constater tout au long de ce document, cette industrie a su s'adapter et relever les défis de son temps. Par ailleurs, Silver (2018) nous fait remarquer que les sondages politiques aux États-Unis, vivement critiqués par la presse, ont certes connu des échecs à de nombreuses reprises, mais ils restent néanmoins aussi proches de la réalité qu'auparavant. Pour confirmer ses propos, Silver (2018, para.7), qui suit de très près l'univers du sondage en matière politique au travers de son site web Fiverthirtyeight.com, a récolté pas moins de 8.500 sondages réalisés les 21 derniers jours qui ont précédé les élections américaines (élections présidentielles, élections du gouverneur, des membres du sénat, etc.) entre 1998 et 2018. Après avoir nettoyé la base de données des sondages qui lui paraissaient peu crédibles, il constate que l'erreur moyenne (par rapport aux résultats définitifs) n'a été que de 6% tous

types d'élections confondus (mêmes les primaires où les erreurs ont tendance à être bien plus significatives en raison des difficultés éprouvées à atteindre ce type de population) entre 1998 et 2018 (Silver, 2018, para.10). Aussi, les biais moyens constatés lors des élections présidentielles sont encore plus faibles que ceux constatés pour tous types d'élections. Ainsi, on constate qu'entre 1972 et 2016, l'erreur moyenne était de 4,6%, soit que 0.2% de différence avec les dernières élections de Donald Trump (Silver, 2018, para.18). Les sondages auraient également correctement estimé le vainqueur (tous types d'élections confondus) dans près de 80% des cas (Silver, 2018, para.31). Les sondages ne seraient donc pas aussi inefficaces que le prétendraient certains !

CYCLE	NATIONAL	STATE	COMBINED
2016	**3.1**	**5.2**	**4.8**
2012	3.3	3.7	3.6
2008	2.3	3.9	3.6
2004	2.2	3.5	3.2
2000	3.9	4.6	4.4
1996	6.4	4.8	5.3
1992	4.6	5.2	5.1
1988	3.5	5.0	4.6
1984	5.4	4.5	4.7
1980	8.9	8.6	8.6
1976	2.5	3.8	3.4
1972	2.6	4.6	4.3
Average	**4.1**	**4.8**	**4.6**

Tableau 6 : Taux d'erreur constaté entre les sondages et la réalité entre 1972 et 2016

Source : Silver, N. (2018). *The Polls Are All Right*. Récupéré le 10 août 2018 de https://fivethirtyeight.com/features/the-polls-are-all-right

Dans le même ordre d'idée, qui consiste à assembler une multitude de sondages réalisés par différents instituts en vue d'obtenir des résultats plus précis et, au passage, démontrer l'efficacité des sondages traditionnels, nous retrouvons l'agrégation établie par HuffPost Pollster (2017). Celle-ci présente, comme nous pouvons le constater dans le graphique (cf. infra p.81), des résultats extrêmement proches de la réalité. Ainsi, l'estimation des sondages agrégés à deux jours du premier tour des présidentielles françaises de 2017 indiquait :

- 23,80% pour le candidat Macron (HuffPost Pollster, 2017), contre 24,01% dans la réalité (Ministère de l'Intérieur, 2017) ;
- 22,20% pour Le Pen (HuffPost Pollster, 2017), contre 21,30% dans la réalité (Ministère de l'Intérieur, 2017) ;

- 19,80% pour Fillon (HuffPost Pollster, 2017), contre 20,01% dans la réalité (Ministère de l'Intérieur, 2017) ;
- 19,40% pour Mélenchon (HuffPost Pollster, 2017), contre 19,58% dans la réalité (Ministère de l'Intérieur, 2017).

Ce niveau de précision est extrêmement élevé en comparaison de l'étude américaine alors que seules 88 enquêtes ont dans le cas français été agrégées (HuffPost Pollster, 2017), contre plus de 8.500 pour l'étude américaine (Silver, 2018, para.7). Si les sondages américains éprouvent plus de difficultés à déterminer le futur président, c'est en partie en raison du mode de scrutin. Sans trop rentrer dans les détails, ce sont les « grands électeurs » (choisis par les citoyens) qui accordent leur vote au candidat à la présidentielle. Si le candidat sort vainqueur d'un État, le candidat reçoit le vote de l'ensemble des « grands électeurs » de cet État, même si l'écart en termes de voix entre les deux candidats est infime (Baron, 2016). En guise d'exemple, imaginons le cas de figure où 50,1% de la population d'un État voteraient pour le candidat A et 49,9% pour le candidat B. Le candidat A ayant plus de voix (des citoyens) remporterait l'ensemble des votes des « grands électeurs » qui détiennent le dernier mot au niveau des élections.

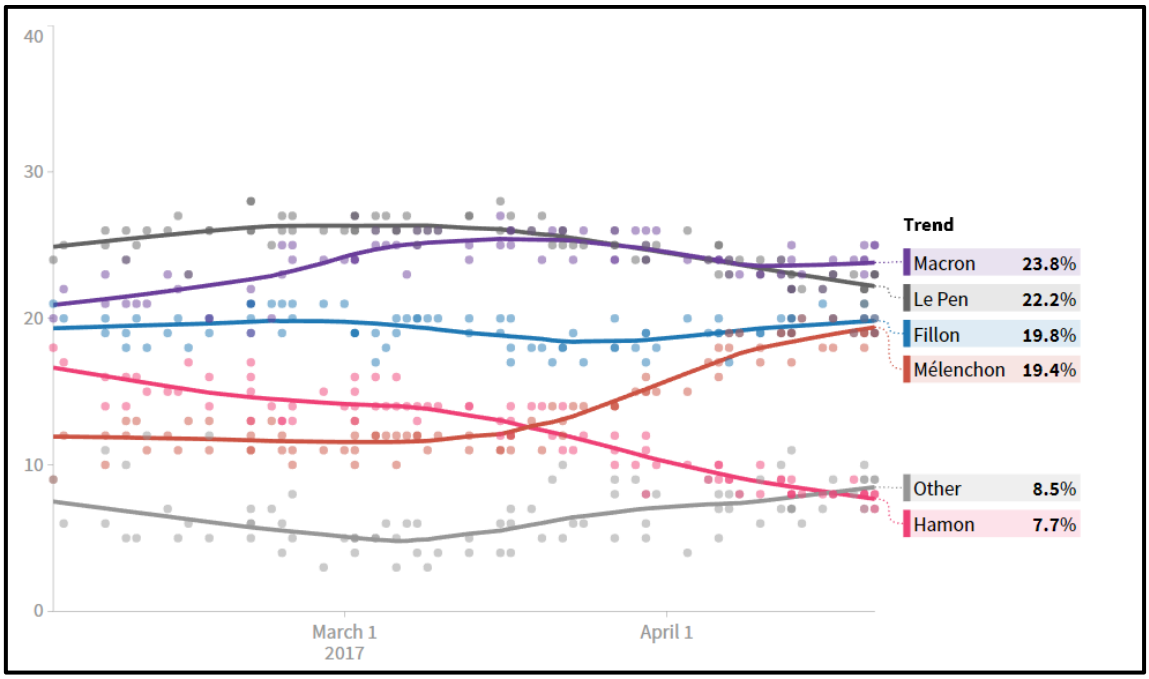


Figure 16 : Agrégation de sondages français en vue des présidentielles

Source : HuffPost Pollster. (2017). *France Presidential Election.* Récupéré le 10 août 2018 de https://elections.huffingtonpost.com/pollster/france-presidential-election-round-1

L'industrie du sondage traditionnel, malgré ses quelques imprécisions, resterait bel et bien parmi les outils auxquels devrait faire appel la société de demain, à l'instar du Big Data qui, comme nous l'avons évoqué un peu plus tôt (cf. supra p.76), devrait voir ses revenus doublés en l'espace de dix ans. La question qu'il nous reste à combler étant dès

lors de savoir si une forme de collaboration est possible entre d'une part l'industrie du sondage traditionnel, et d'autre part, le Big Data. Selon Hing Lo (2016), les spécialistes de l'industrie du sondage évoqueraient déjà cette possibilité. Le Big Data serait utilisé pour mettre en évidence des corrélations nouvelles, établir des faits concrets sur l'activité de la population. Autrement dit, cette technologie s'attarderait plutôt sur ce que font les individus et combien ils sont à le faire. Le sondage traditionnel aurait pour rôle, quant à lui, de définir pour quelles raisons les individus adoptent tel ou tel comportement. Cette pensée rejoint également celle de Crawford (2013) qui indiquait qu'il était essentiel de combiner les deux outils afin d'améliorer nos connaissances, et de ne pas uniquement se contenter d'un nombre volumineux de données, car cela aurait pour effet d'explorer l'être humain d'une manière superficielle. Une idée qui, semble-t-il, ferait déjà sa route auprès des instituts de sondages BVA et Opinion Way qui utiliseraient les réseaux sociaux pour obtenir des informations complémentaires sur l'opinion publique, sans pour autant que celles-ci ne transparaissent dans les résultats de leurs études (Rolland, 2017).

Pour Callegaro et Yang (2017), la combinaison des deux outils nous permettrait de sortir de certaines impasses intrinsèquement liées à leur mode de fonctionnement. Pour ces experts, lorsque le répondant est confronté à des questions précises concernant par exemple ses habitudes d'achat, il ne peut de toute évidence se souvenir des moindres détails des articles achetés, surtout si ces achats remontent à quelques semaines, voire des mois. Il serait dès lors préférable à ce stade de tendre le relais au Big Data qui, pour ce type de détail, est plus performant. Toutefois, ces spécialistes n'estiment pas pour autant que cette technologie, si sophistiquée soit-elle, soit en mesure de répondre à toutes les demandes des études comportementales. Par ailleurs Zhang et Chen (cités par Callegaro et Yang, 2017) nous signalent que le réseau social Facebook a régulièrement recours au sondage traditionnel pour évaluer leurs services. Les questionnaires représenteraient un outil complémentaire (au-delà des apports du Big Data via les likes, commentaires, etc.) pour améliorer l'expérience des utilisateurs du réseau social qui passe bien entendu par une meilleure gestion de leur fil d'actualité. Enfin, Facebook ne serait pas le seul à faire usage des sondages traditionnels, Google serait également adepte de telles pratiques dont les finalités seraient identiques à celles de Facebook, à savoir : comprendre pour quelles raisons les internautes sont amenés à visiter tel ou tel site, pour quelles raisons ils n'y sont pas restés plus longtemps, etc. (Eduardo Kokoyachuk, 2017).

Enfin, Slessareva (2016) indique quelques-uns des bénéfices que nous pourrions tirer de la combinaison des deux outils. Tout d'abord, cela devrait permettre d'obtenir une vision plus globale et complète de la clientèle ; le Big Data irait collecter toutes les données disponibles sur les clients d'une entreprise, tandis que le sondage permettrait par exemple d'établir les raisons qui poussent les consommateurs à se rendre auprès de la concurrence. Ensuite, le sondage permettrait de contextualiser les raisons d'achat (ce qui rejoint Callegaro et Yang), servir lors de prétests de campagnes promotionnelles, établir le parcours client sur les sites web des entreprises, etc. Aussi, le Big Data serait plutôt habilité à évaluer l'efficacité d'une campagne promotionnelle uniquement après que celle-ci ait été diffusée au grand public, alors que le sondage traditionnel peut en amont

évaluer la bonne compréhension du message publicitaire (d'une campagne encore jamais diffusée). En résumé, cette spécialiste évoluant dans le monde du sondage, vice-présidente du département technologie au sein de GfK, reconnaît d'une part les forces du Big Data en matière d'acquisition des données, mais pointe du doigt ses lacunes telles que les difficultés à exprimer les motivations des individus, établir leurs intentions d'achat, etc. D'où l'utilité d'une collaboration avec le sondage traditionnel qui viendrait combler les manquements du Big Data, et inversement.

3.2. Étude qualitative

Pour peaufiner et actualiser quelque peu les informations recueillies lors de notre desk research, nous avons décidé de réaliser une étude qualitative auprès d'experts et de professionnels ayant une expérience dans le domaine du Big Data (obligatoire) et de l'industrie du sondage (facultatif).

Ces répondants devaient au minimum avoir deux années d'expérience. Il aurait bien entendu été préférable de ne recueillir des informations qu'auprès d'individus ayant au minimum 7 à 10 ans d'expérience, mais les difficultés rencontrées ne nous ont pas permis d'atteindre un nombre suffisant de personnes répondant à ce critère. Cependant, nous avons décidé de répartir nos différents participants selon leur niveau d'expérience (comme cela est indiqué plus bas) afin que le lecteur puisse nuancer certains des avis recueillis. Enfin, par années d'expérience, nous prenions en compte les années d'activité professionnelle ainsi que les années d'études consacrées au sujet (Big Data et/ou sondage traditionnel) dans le cadre d'un doctorat ou de recherches professionnelles par exemple.

Si nous avions au préalable souhaité réaliser ces enquêtes qualitatives en face à face (de préférence) ou par téléphone, nous n'avons malheureusement pas pu obtenir de rendez-vous avec des experts malgré l'envoi de mails et les appels téléphoniques. Nous avons dès lors décidé d'administrer les questionnaires via un outil d'enquêtes en ligne professionnel intitulé Net-Survey. Ce logiciel est celui que nous avons pour habitude d'utiliser dans le cadre de notre emploi au sein de l'institut de sondages Dedicated. Outre les diverses fonctions supplémentaires qu'offre ce logiciel comparé aux versions gratuites que nous trouvons sur la toile, Net-Survey facilite la tâche de programmation des questionnaires, de leur traduction et de la gestion de base de données. Aussi, Dedicated a mis à notre disposition ses serveurs sécurisés afin que nous puissions avoir le parfait contrôle de nos données et veiller à leur sécurité.

Le questionnaire a été réalisé en français et traduit en anglais afin d'atteindre une cible beaucoup plus large (les questionnaires dans les deux versions de langue sont disponibles en annexe p.112 et p.119) ; c'est d'ailleurs cette traduction qui nous a permis de recueillir 8 de nos 10 enquêtes. Le questionnaire comportait 4 parties distinctes (profil du répondant, connaissances générales sur le Big Data [pour mieux évaluer la grille de lecture du répondant], domaines d'application du Big Data, et comparaison entre le Big Data et l'industrie du sondage traditionnel).

La durée moyenne de complétion de ce questionnaire fut initialement estimé à 45-60 minutes (30 minutes par Internet), elle fut de 43 minutes. La durée totale du terrain fut d'environ 3 semaines ; l'enquête s'est déroulée entre le 16 juillet et le 08 août 2018.

Pour recruter nos experts, nous avons dans un premier temps utilisé les réseaux sociaux LinkedIn et Facebook ; nous avons intégré des groupes dédiés spécifiquement à notre thématique d'étude qui comportaient parfois près de 100.000 individus. Aussi, après avoir lu certains articles d'experts, nous avons décidé de contacter par mail ces derniers (en leur faisant directement parvenir le lien de l'enquête). Enfin, le bouche-à-oreille nous a permis de rentrer en contact avec des personnes qui soit travaillaient dans le Big Data, soit avaient des collègues, des amis, des connaissances dans ce domaine.

Au total, ce sont 13 répondants qui ont participé et achevé la complétion de notre étude. Cependant, 3 répondants avaient une expérience insuffisante ou n'avaient pas formulé de réponses consistantes (ce qui nous a obligé à les supprimer). Après nettoyage de notre base de données, il ne nous restait donc plus que 10 enquêtes sur lesquelles nous pouvions effectuer nos analyses. Aussi, nous tenons à préciser que parmi ces dix enquêtes, deux répondants ayant une forte expérience dans le domaine du Big Data n'ont pas complété les deux dernières thématiques de notre enquête. Nous avons tout de même décidé de comptabiliser ces répondants (spécialisés dans le Big Data), car leurs apports au sein des éléments de définition du Big Data étaient totalement pertinentes.

a. Profil des répondants

Du point de vue du profil des répondants, nous pouvons catégoriser les profils en trois groupes distincts :

1. un premier groupe comportant des experts de l'industrie du sondage et/ou de l'univers du Big Data ; nous les nommerons « les spécialistes », ils possèdent tous (à une exception près) de 10 à 35 ans d'expérience (par année d'expérience, nous entendons aussi bien des années de travail, que d'études ou autres),
2. un groupe ayant une moins grande expérience dans le domaine du Big Data et aucune expérience dans l'industrie du sondage ; nous les nommerons « les connaisseurs », ils possèdent 5 ans d'expérience *[ce groupe n'a répondu qu'aux questions de profil et de définition du Big Data]*,
3. et un dernier groupe composé de personnes ayant très peu d'expérience au sein du Big Data et de l'industrie du sondage ; nous les nommerons « les débutants », ils possèdent deux années d'expérience.

Groupes	Années d'expérience		Nombre d'individus
	Big Data	Sondages	
Les spécialistes	10 à 15 ans	10 à 35 ans	5
Les connaisseurs	5 ans	Aucune	2
Les débutants	2 ans	2 ans	3

Tableau 7 : Profils des répondants

Étant donné la dispersion de notre échantillon, nous préciserons lorsque nous le jugerons utile (c'est-à-dire en cas de (forte) divergence à une question donnée) le groupe auquel appartient le répondant. Cela devrait permettre au lecteur de pouvoir nuancer certaines

des affirmations, pouvant paraître très pertinentes, mais qui, par exemple, proviendraient d'un répondant n'ayant que très peu d'expérience.

Nous noterons que l'ensemble des répondants a indiqué travailler dans le domaine du Big Data et/ou de l'industrie du sondage. Certains répondants sont ou ont été chefs d'entreprise dans l'un des domaines. D'autres réalisent des recherches dans l'univers du Big Data que ce soit dans un cadre professionnel ou académique (doctorat, post-doctorat).

Une telle variété de répondants devrait nous permettre d'obtenir des regards croisés sur notre thème de recherche, différentes grilles de lecture selon les expériences auxquelles ont été confrontées les répondants. Aussi, hormis les connaisseurs, nous devrions en tout état de cause éviter les raisonnements binaires ; les répondants ont de manière générale une assez grande expérience dans les deux domaines. Enfin, cette vision plurielle est aussi renforcée par la dimension multinationale que revêt notre étude, qui a permis à des individus de divers pays de pouvoir participer à notre enquête.

b. Connaissances générales de l'univers du Big Data

Spontanément, il apparait que la majorité des répondants (9 sur 10) a fait mention d'une des caractéristiques les plus évidentes de cette technologie, à savoir le volume. Une masse de données que cette technologie permet de traiter assez facilement et à moindre coût (selon un connaisseur). 5 répondants sur 10 ont indiqué que cette technologie autorisait l'utilisation de données variées structurées ou non, provenant par exemple de transactions (commerciales), d'un service clientèle, d'un réseau social. Enfin, la vélocité figure parmi les caractéristiques du Big Data pour 3 individus ; ils spécifient que cet outil permet de gérer très rapidement d'importants volumes de données.

Après avoir indiqué de manière spontanée ce que représentait le Big Data, nous avons proposé aux répondants divers éléments pouvant faire partie d'une définition plus globale de cette technologie qui sont tous issus de la définition émise par Kitchin (cf. supra p.26). Ces éléments ont été résumés comme suit :

- le volume de données : le Big Data est un outil permettant l'exploitation de masses considérables de données ;
- l'exhaustivité : le Big Data est un outil permettant l'exploitation de la totalité (ou de la quasi-totalité) des données de l'univers de référence étudié ;
- la granularité : le Big Data est un outil permettant le traitement des données de manière distincte, c'est-à-dire que chaque donnée serait totalement unique et identifiable ;
- le degré de relation entre les données : le Big Data est un outil permettant de croiser différentes données (quelles que soient leur structure) afin de répondre à des questions nouvelles ;
- la vélocité : le Big Data est un outil permettant d'enregistrer des données en temps réel, de manière instantanée ;

- la variété : le Big Data est un outil permettant de collecter des données hétérogènes, c'est-à-dire de formats différents (vidéos, images, textes…) et de sources différentes (réseaux sociaux, recensements…) ;
- la flexibilité : le Big Data est un outil permettant d'ajouter de nouveaux éléments à une base de données déjà existante (en vue de croiser ces nouvelles données) et peut s'adapter à l'explosion de la demande de données.

Ainsi, nous apprenons que pour :

- le volume de données : l'ensemble des répondants a estimé que cet élément de définition représentait « tout à fait » ou « plutôt » bien l'idée qu'ils se faisaient du Big Data. Par ailleurs, ils considèrent tous que cet élément est « assez » ou « très » important pour le bon fonctionnement de cette technologie.

  Le volume de données est une caractéristique inséparable de cette technologie qui permet aux analystes d'avoir une vision plus globale de l'objet d'étude, et d'acquérir plus de connaissances. La masse de données permettrait également de traiter des phénomènes complexes,

- l'exhaustivité : cet élément divise les répondants. Les spécialistes ont indiqué ne pas vraiment être d'accord, alors que les deux autres groupes étaient plutôt d'accord. Néanmoins, la plupart (6 répondants) estime que l'exhaustivité est « très » ou « assez » importante ; parmi ces répondants 5 figurent parmi les spécialistes.

  Les répondants reconnaissent qu'il est difficile, voire illusoire d'obtenir un échantillon exhaustif comportant l'intégralité d'une population étudiée, mais tendre vers l'exhaustivité accorderait forcément plus de précision. De plus, pour un connaisseur, cette notion d'exhaustivité n'a pas de sens en raison de certaines limitations relatives à l'imagination du data analyst, à l'accessibilité des données ainsi qu'à l'infrastructure matérielle nécessaire à l'exploitation d'une telle masse de données,

- la granularité : si les spécialistes acceptent majoritairement (5 répondants sur 6) cet élément de définition, les connaisseurs le rejettent. Pour autant, à l'exception d'un répondant, tous considèrent que l'identification individuelle est nécessaire au bon fonctionnement du Big Data.

  L'identification des données de manière distincte revêt une assez grande importance pour les spécialistes. Cette particularité permet de mieux cibler les individus, d'identifier ce qui les distingue des autres. De plus, indique un spécialiste, la connaissance passe par l'analyse de petits détails et de tendances qui jusqu'alors ne pouvaient être étudiés à l'aide d'outils traditionnels,

- le degré de relation entre les données : À l'instar du volume de données, la possibilité de mettre en relation des données pour répondre à des questions

nouvelles est un élément de définition largement accepté par les répondants, et qui est nécessaire pour le bon fonctionnement de cette technologie.

L'interconnexion de base de données conduit indubitablement à une plus-value en termes de connaissance. Par ailleurs, ajoute un spécialiste, c'est cette particularité qui rend le Big Data si intéressant pour ses utilisateurs. Néanmoins, précise un connaisseur, cette caractéristique n'est pas propre au Big Data, elle existerait bien avant la venue de cette technologie (un point de vue que nous ne pensons pas impossible, mais nous doutons que cette capacité fut aussi performante qu'elle ne l'est actuellement au sein du Big Data),

– la vélocité : la vitesse avec laquelle les données sont collectées figure parmi les signes distinctifs de cette technologie pour 8 répondants, soit la quasi-totalité de ceux-ci. Elle est d'ailleurs considérée comme (très) importante pour le bon fonctionnement du Big Data.

  Si l'étude de certains cas n'exige pas réellement de vélocité (ex. : lors de l'analyse d'historiques de données), dans d'autres cette collecte en temps réel serait très utile (ex. : l'étude de données volatiles qui disparaissent aussi rapidement qu'elles apparaissent) et permettrait aux chercheurs d'améliorer leurs analyses.

– la variété : l'ensemble des répondants, sans exception, estime que cet élément fait partie de leur représentation du Big Data. L'agrégation de différents types de données (ex. : formats différents) représente en toute logique une valeur essentielle pour le bon fonctionnement de cette technologie pour l'ensemble des personnes interrogées (hormis un répondant qui indique pour sa part que les données structurées sont source de beaucoup de savoirs, et qu'il faudrait dès lors s'affairer à les exploiter au mieux avant de se tourner vers la combinaison de données différentes).

  Aujourd'hui, les individus s'expriment (ou laissent des traces digitales) de différents formats sur la toile (images, commentaires, etc.). Il devient de ce fait important d'agréger ces différentes sources d'informations sous un format identique afin d'obtenir une meilleure vision des sujets d'études. Enfin, nous indique un débutant, la variété autorise l'analyse des individus sous différents angles et permet d'éclaircir certaines zones d'ombre concernant ces individus (qu'une seule source d'informations n'aurait pu être en mesure de réaliser),

– la flexibilité : si cet élément de définition ne fait pas l'unanimité auprès des répondants, la majorité de ceux-ci considère ce critère comme « assez » ou « très » important pour le bon fonctionnement du Big Data.

  Si la difficulté d'implémenter une nouvelle base de données à une base déjà existante semble assez difficile (un connaisseur qualifie d'ailleurs la flexibilité comme étant un mythe du Big Data), cette possibilité reste intéressante. Les individus et la société changeant constamment, il est nécessaire pour le Big

Data de s'adapter aux nouvelles sources de données (possédant un format différent) ainsi qu'à l'augmentation de la quantité de données récoltées, et de pouvoir ainsi intégrer celles-ci aux données existantes.

En résumé, les répondants ne reconnaissent pas tous les attributs du Big Data que propose Kitchin dans sa définition (cf. supra p.26). Mais de manière générale, il apparaît que tous ces éléments sont intéressants pour le bon fonctionnement de cette technologie, à l'exception de l'exhaustivité pour laquelle les répondants ont un regard mitigé sur l'utilité et la possibilité du Big Data à s'approprier l'ensemble des données d'un univers de référence.

c. Avantages et faiblesses du Big Data

[À ce stade, nous souhaitons rappeler que seuls 8 répondants ont apporté des réponses, à savoir les 5 spécialistes et les 3 débutants ; les 2 connaisseurs n'apparaîtront plus dans le cadre de cette enquête]

Comparativement au sondage traditionnel, le Big Data possède les avantages suivants :

- le nombre de données pouvant être traité et analysé est théoriquement infini ;
- plusieurs hypothèses peuvent être infirmées ou confirmées en même temps, étant donné que, contrairement au sondage traditionnel, le Big Data ne se limite pas à répondre à une seule interrogation ;
- facilite l'interconnexion, la corrélation des données ;
- permet d'apporter des réponses de manière plus rapide ;
- les coûts seraient moins importants (si on ne tient pas compte des coûts de l'investissement initial) ;
- la granularité (qui permet de traiter les individus de manière distincte et non par groupe) ;
- une vision globale du sujet d'étude ;
- améliore la prise de décision par l'intermédiaire de cette capacité de prédiction que possède cette technologie ;
- la prise de décision est basée sur des faits, des chiffres plutôt que sur des ressentis.

Comparativement au sondage traditionnel, le Big Data possède les faiblesses suivantes :

- ne permet pas de savoir pour quelles raisons les individus pensent, ressentent ou agissent de telle ou telle façon ;
- si une solution (un produit ou un service) n'existe pas encore, il n'est pas possible pour le Big Data d'établir si celle-ci sera acceptée par la population ;

la capacité de prédiction ne servirait que si le produit ou le service sont déjà consommés par une population ;

- certains cas particuliers pourraient malencontreusement pousser l'analyste à déboucher sur de mauvaises conclusions ;
- une focalisation trop importante sur la quantité de données amassée qui fait oublier aux analystes le caractère imprévisible du comportement humain.

d. Les domaines d'application du Big Data

De manière spontanée, les répondants ont indiqué que l'outil Big Data pourrait être utilisé dans les domaines du politique, du monde de la finance, au sein du secteur commercial (que ce soit dans la grande distribution ou pour les sites commerciaux afin de mieux identifier le profil des clients, s'assurer qu'une campagne marketing a fonctionné, évaluer le taux d'attrition [qui mesure la perte de clientèle]), au sein du secteur de la santé ainsi qu'au sein du monde des assurances (gestion des risques : pour permettre d'identifier les personnes les plus susceptibles de commettre un accident de voiture).

De manière assistée, il apparait pour les différents domaines faisant objet de notre étude que :

- au niveau du domaine politique, le Big Data devrait permettre d'indiquer :
  - pour quel parti un individu a voté par le passé ;
  - pour quel parti un individu compte voter aux prochaines élections ;
  - quel est le profil sociodémographique (âge, niveau d'éducation, genre, religion…) de cet individu ;
  - le ressenti envers certains sujets d'actualité (afin de permettre aux politiques de mettre en place des mesures appropriées pour tenter de convaincre les électeurs) ;
  - [néanmoins, pour un spécialiste, tous les changements n'apparaissent pas directement au sein des radars du Big Data. Il y a un délai pour que ceux-ci puissent être détectés, et ce, aux dépens des politiques] ;
- au niveau du secteur public : le Big Data aurait une meilleure vision de la population et des difficultés que celle-ci rencontre au quotidien. Par exemple, en termes de déplacement, cette technologie pourrait permettre d'identifier où se situent les zones à forte concentration pour pouvoir mettre en place des modes de transport supplémentaires (ou alternatifs) et, ainsi, répondre plus efficacement aux besoins de la population ;
- au niveau du secteur de la santé : cette technologie pourrait aider le monde médical à améliorer le diagnostic des patients. Cependant, même si le potentiel au sein de ce domaine est grand, le Big Data serait freiné par toutes les lois sécuritaires qui visent à protéger les données des patients ;

- au niveau du secteur commercial : le Big Data devrait permettre d'améliorer l'expérience client (proposer des produits ciblés, adapter le message selon le profil du consommateur, etc.).

e. Perceptions actuelles de la valeur ajoutée du Big Data sur le sondage traditionnel

Les répondants indiquent qu'actuellement le Big Data démontre une réelle capacité à apporter une plus-value dans les différents domaines de prédilection de l'industrie du sondage. Cela est notamment rendu possible du fait que cette technologie permet de traiter des systèmes complexes, des bases de données très volumineuses, de faire apparaître des informations qui n'auraient pu être décelées via la méthode traditionnelle. Aussi, cet outil permettrait de réduire les coûts qu'engendrent les études.

Cependant, le Big Data comporte certaines lacunes. En effet, certaines études ne seraient envisageables que via le sondage traditionnel qui permet entre autres de déterminer pour quelles raisons les individus se comportent de telle ou telle manière. De plus, la méthode traditionnelle autoriserait la création de questionnaires faits sur mesure afin de répondre à des demandes spécifiques.

f. Perceptions futures du Big Data et du sondage traditionnel

Interrogés sur l'avenir de ces deux outils, les répondants sont pratiquement unanimes. Le futur devrait voir l'émergence d'une collaboration du Big Data et du sondage traditionnel. Chacun des instruments serait utilisé dans la partie de l'étude où il est le plus performant. Le Big Data se chargerait d'indiquer ce que font les individus et le sondage traditionnel se pencherait plutôt sur les raisons du comportement de ceux-ci. En définitive, ces deux outils se complèteraient ; il faudrait par ailleurs, indiquent des répondants, mener des recherches pour déterminer, lors d'une étude, la répartition des tâches entre le Big Data et le sondage traditionnel.

Toutefois, cette future (éventuelle) collaboration ne se ferait pas sans encombre. Le Big Data devra selon la plupart des répondants trouver une issue face aux lois qui visent à mieux protéger les internautes ; la gestion et l'utilisation des données devront dès lors être sécurisées pour éviter que ces données ne tombent entre de mauvaises mains. Aussi, on évoque un défi d'ordre technologique ; le Big Data devra être bien plus performant pour pouvoir intégrer de grands volumes de données et traiter des données complexes.

II. Conclusion

Comme nous l'évoquions dans l'introduction de cette seconde section, le lecteur dispose à présent d'un regard pratique sur l'objet de notre étude. Notre question de recherche était focalisée sur les domaines d'application du Big Data et les conséquences de cette technologie sur l'industrie du sondage traditionnel. Ainsi, des cas pratiques, des analyses d'experts et de professionnels sont venus répondre à la plupart de nos interrogations.

Le mode de fonctionnement, les domaines d'application ainsi que les forces et faiblesses du Big Data étant établis, tant de manière théorique que pratique, il ne nous reste plus qu'à rassembler l'ensemble des informations de manière synthétique afin d'apporter une réponse croisée à notre question de recherche.

Troisième partie : Synthèse

I. Introduction

Nous voilà arrivés à la fin de notre développement. Il convient pour conclure ce travail de rassembler les pièces du puzzle que nous avons pu récolter tout au long de notre processus de recherches. L'objectif étant de formuler une réponse à notre question de recherche qui, pour rappel, était la suivante : « Quel est l'impact du Big Data au sein des domaines de prédilection de l'industrie du sondage traditionnel, à savoir : le monde politique, le secteur privé, le secteur de la santé et le secteur public, et, dès lors, quelles sont les conséquences des apports du Big Data sur les activités de l'industrie du sondage traditionnel ? ».

Cette troisième et dernière partie de notre travail est également pour nous l'occasion de communiquer au lecteur les limites de cette étude (les freins qui nous ont empêchés d'aller plus loin dans notre recherche) et les perspectives de recherche futures (les prochaines questions de recherche qui gravitent autour de notre thématique).

1. Analyse critique et mise en perspective

Pour répondre à notre questionnement, nous avons décidé de scinder les différents éléments de notre question de recherche afin de nous concentrer uniquement sur des points spécifiques l'un après l'autre. Nous commencerons dès lors cette analyse critique par une synthèse de tous les éléments récoltés sur l'impact du Big Data au sein des domaines de prédilection de l'industrie du sondage, pour ensuite évaluer les conséquences des apports de cette nouvelle technologie sur cette industrie.

1.1. Impacts du Big Data sur les domaines de prédilection du sondage

Lorsque pour la première fois (cf. supra p.40) nous évoquions les domaines d'application possibles du Big Data, nous avions offert au lecteur une vision théorique des différents apports de cette technologie. Cette vision, pour le moins idyllique, mettait en exergue les impressionnantes contributions que le Big Data pouvait apporter au monde politique, au secteur privé, au secteur de la santé et au secteur public. Par ailleurs, nous terminions ces différents sous-points en nous interrogeant de manière inquiète sur l'avenir du sondage traditionnel qui, selon nous, se voyait véritablement empiéter sur son terrain ; en d'autres mots, ses zones de prédilection.

Pour évaluer plus concrètement ces différents apports (et infirmer ou confirmer ceux-ci), nous sommes allé à la recherche de cas pratiques (cf. supra p.57) afin de mieux appréhender le mode de fonctionnement du Big Data. Notre travail de prospection a été fructueux, mais non sans mal. Il paraît en effet logique que si cette technologie ait été utilisée dans un domaine ou un autre, les bénéficiaires ne souhaitent pas forcément révéler les avantages retirés de cet outil pour leur entreprise ; ce qui tendrait dans un sens à donner quelques indices sur l'état du marché, la position concurrentielle, l'état d'esprit de la population envers l'un ou l'autre projet politique, etc.

Au niveau du secteur privé (cf. supra p.57), nous avons pu évaluer l'efficacité du Big Data à l'aide d'un exemple concret provenant de la grande distribution. Dans cet univers

en perpétuel changement, qui ne donne que peu de places à l'erreur, il est important pour l'entreprise de prendre très rapidement des décisions concernant les produits et services qu'elle propose ou souhaiterait proposer aux consommateurs. Par ailleurs, il est important pour une société telle que Walmart d'identifier de manière minutieuse les comportements des consommateurs et l'évolution de leurs tendances en matière d'achat. Ce besoin de rapidité dans la prise de décision et de suivi constant des consommateurs peut en regard de nos analyses correspondre à la technologie Big Data. Cet outil, comme nous l'avons indiqué plus tôt (cf. supra p.30), permet à ses utilisateurs de déterminer ce que font les individus en temps réel et permet aussi d'apporter des informations pertinentes qui aideront les entreprises à réaliser des prédictions quant à la demande future.

Au niveau du secteur médical (cf. supra p.60), nous avons pu constater l'utilité du Big Data pour le traitement d'une base de données assez conséquente. Bien entendu, le cas que nous avons évoqué ne comportait en définitive que très peu de variables et d'individus à analyser en comparaison des capacités de cette technologie. Cependant, la base de données traitée était exhaustive et les variables à étudier étaient présentes sous différents formats, ce qui rendait l'exercice d'autant plus difficile. Dans un premier temps, le Big Data avait permis à l'hôpital saoudien d'agréger l'intégralité des données stockées à divers endroits afin de pouvoir chiffrer la problématique du service d'urgence. Ensuite, une fois le problème identifié et les améliorations proposées et mises en place, l'outil avait permis de suivre en continu l'impact des améliorations sur le quotidien des premiers bénéficiaires, les patients.

Au niveau du secteur public (cf. supra p.66), le cas de la société de transport londonien nous a démontré le potentiel de la technologie Big Data. Par l'intermédiaire de cet outil, les gestionnaires du réseau ont désormais une vision tant globale que granulaire de leurs millions d'usagers. De plus, le Big Data permet de suivre en temps réel le quotidien des voyageurs, les chemins utilisés, les types de transport pris, etc. Aussi, les données sont récoltées sans faire appel à la mémoire des individus ; ce sont pour rappel les cartes Oyster et les bornes wifi auxquelles se connectent les usagers (qui s'avèrent par ailleurs plus précises) qui transmettent la position de ces derniers. Les apports du Big Data dans ce secteur en particulier ont permis de décongestionner certaines stations, proposer de nouveaux itinéraires aux voyageurs et devraient encore montrer de nouveaux bénéfices dans le futur.

Au niveau du monde politique (cf. supra p.70), les traces digitales (que nous laissons tous à travers la toile) ont permis à des chercheurs d'identifier le profil des utilisateurs du réseau social Facebook. Ils ont pu, entre autres, démontrer que la simple utilisation de likes était déjà une source d'informations pouvant indiquer nos convictions politiques et religieuses, nos origines, nos critères sociodémographiques, etc. Certes, le profilage manquait de précision, mais celui-ci n'était réalisé qu'à l'aide de likes. En imaginant que bien plus de critères soient pris en considération par les analystes, nous pourrions imaginer que le Big Data parvienne à déterminer de manière plus précise le profil des internautes (âge, genre, etc.) et croiser ces informations avec leurs convictions politiques. Par ailleurs, des sociétés comme Contemporary Analysis se targuent déjà de pouvoir

utiliser le Big Data pour aider le monde politique. D'autres, comme Cambridge Analytica, sont allées trop loin et remettent sur le devant de la scène le débat sur la protection de la vie privée.

À la suite de l'analyse de ces différents cas, nous ne pouvons qu'accepter les apports indéniables de la technologie Big Data dans les différents domaines qui ont fait l'objet de notre recherche. Cependant, cette analyse nous permet également de mieux comprendre le mode de fonctionnement de cette technologie. Celle-ci ne s'attarde pas, comme nous l'avions évoqué plus tôt (cf. supra p.82), à questionner le consommateur, le patient, l'usager, le citoyen sur son comportement, mais s'intéresse surtout à ce que celui-ci fait. Cette source d'informations est suffisante pour certains analystes qui estiment que les raisons d'une action n'ont plus réellement d'utilité (cf. supra p.75). Ce mode de réflexion est selon nous préjudiciable pour les chercheurs (et les entreprises) qui déboucheront à maintes reprises sur de fausses conclusions, avec les conséquences que celles-ci auront entrainées. Les informations recueillies par le Big Data doivent dès lors être considérées comme utiles et nécessaires pour la société, mais ne pourront en l'occurrence pas remplacer les apports du sondage traditionnel qui ont la particularité de pouvoir répondre au pourquoi d'une action. Ce qui nous amène à la conclusion suivante : les apports de ces deux outils sont complémentaires, car ils permettent aux chercheurs d'enrichir leurs savoirs et non de s'approprier deux fois le même type d'informations.

1.2. Conséquences du Big Data sur l'industrie du sondage traditionnel

L'impact du Big Data étant réel sur les divers domaines de prédilection de l'industrie du sondage traditionnel, mais non quantifiable (aucun chiffre n'étant à notre disposition pour évaluer les pertes de marchés de l'industrie du sondage au profit du Big Data), il nous restait à déterminer les répercussions de cet outil sur l'univers du sondage traditionnel.

Nous avions dès la formulation de notre question de recherche émis l'hypothèse que nous doutions fortement de voir le Big Data mettre fin au règne du sondage traditionnel. Nous avions avancé comme arguments les différentes faiblesses identifiées lors de la première partie de notre travail (cf. supra p.50). Celles-ci concernaient entre autres les normes de protection de la vie privée, l'erreur d'estimation de Google Flu, l'échec lors des élections présidentielles américaines de 2016, etc. Notre desk research est venu conforter cette idée. Elle a démontré d'autres cas (cf. supra p.77) où la seule visualisation des données recueillies entrainait des erreurs d'interprétation comme avec le cas Twitter et la mesure du taux d'anxiété d'une population à la suite d'un acte terroriste tels ceux de Bruxelles ou Paris. Par ailleurs, l'enquête que nous avons menée auprès de professionnels nous a permis de prendre un recul critique vis-à-vis de cette technologie. Les spécialistes du monde du sondage nous ont en effet rappelé certaines faiblesses du Big Data telles que l'impossibilité de répondre au pourquoi d'une action, l'oubli du caractère imprévisible de l'être humain qui ne se comporte pas toujours de manière rationnelle (cf. supra p.89), etc.

L'analyse des études de cas (cf. supra p.57) ainsi que les données relatives aux chiffres d'affaires futures du Big Data nous ont permis d'exclure le scénario inverse, à savoir que le Big Data disparaîtrait des écrans. Cette technologie ne doit pas être considérée comme

une tendance passagère, elle fera partie de notre futur (Hing Lo, 2016). Dès lors, si ces deux outils doivent cohabiter, il nous paraissait évident à ce stade de déterminer si le Big Data et l'industrie du sondage pouvaient collaborer pour l'intérêt des parties prenantes, et plus particulièrement de leurs clients communs issus des différents domaines que nous avons précédemment cités.

La collaboration existe déjà ! Celle-ci reste tout de même très timide. Nous n'avons trouvé que deux exemples, à savoir celui de Google et de Facebook, qui utiliseraient le système d'enquêtes traditionnel, en plus du Big Data, pour améliorer leurs connaissances sur leurs utilisateurs (Eduardo Kokoyachuk, 2017). Toutefois, lors de notre desk research, nous avions évoqué quelques-unes des formes de collaboration possibles entre les deux outils. Il était question, par exemple, d'utiliser la technologie Big Data pour déterminer les mouvements du marché (voir ce qui s'y passe) et d'agrémenter ces connaissances du pourquoi (cela se passe). Ainsi, les manquements d'un outil seraient comblés par les avantages de l'autre outil, et inversement (Slessareva, 2016). Enfin, les spécialistes que nous avons interrogés lors de notre enquête qualitative nous ont également fait part de leurs avis sur cette question. Pour ces derniers, le futur se fera également sous la forme d'une collaboration entre le Big Data et l'industrie du sondage. Une association qui, nous le rappelons, ne se ferait pas sans encombre. Les spécialistes ont évoqué deux défis à relever principalement, à savoir : la protection de la vie privée et les défis technologiques.

2. Les limites de notre étude

Nous souhaitons par le biais de ce point informer le lecteur quant aux deux principales limites que nous avons rencontrées lors de nos recherches ; les deux ayant un point commun, à savoir le temps qui nous était imparti pour ce travail.

Notre sujet d'étude étant très complexe, nous étions dans l'obligation d'interroger des experts ayant une certaine expérience dans le domaine du Big Data (obligatoire) et du sondage traditionnel (facultatif). Dès lors, nous avions pour objectif de n'interroger que des individus ayant au minimum 7 à 10 ans d'expérience que ce soit dans le domaine du Big Data ou dans le domaine du sondage traditionnel (mais avec quoi qu'il en soit un minimum de connaissance du Big Data). Toutefois, comme indiqué plus tôt, il fut très compliqué d'obtenir des rendez-vous durant la période estivale, celle où nous commencions notre phase de terrain. Nous avions également l'intention de nous rendre dans des salons spécialisés, mais malheureusement, le salon qui nous a semblé le plus intéressant et le plus accessible, celui de Paris, n'avait lieu qu'une fois par an aux alentours du mois de mars ; le prochain a d'ailleurs lieu les 11 et 12 mars 2019, et devrait réunir près de 250 exposants et 100 orateurs (Big Data Corp, s.d.). Durant un tel salon, nous pensons raisonnablement qu'il nous aurait été plus pratique de trouver parmi les exposants et orateurs des personnes susceptibles de correspondre à l'objet de notre étude et ayant l'amabilité de nous accorder un entretien. Ce concours de circonstances nous a forcé à adopter une stratégie d'approche différente pour cette étude.

Nous avons dû accepter des personnes avec moins d'expérience qu'escompté (il s'agit pour rappel des connaisseurs et des débutants qui avaient entre 2 et 5 ans d'expérience

maximum). De plus, toutes les enquêtes ont été réalisées par Internet alors que celles-ci étaient prévues pour des entretiens en face à face. Si le taux moyen de complétion fut d'environ 45 minutes, nous imaginons que certains répondants avaient peut-être plus de mal à écrire via un clavier d'ordinateur. De plus, quelle que soit la vitesse à laquelle un répondant puisse utiliser un clavier, la communication verbale reste le moyen le plus rapide pour communiquer. Nous pensons dès lors que nous aurions eu des réponses plus concises d'une part et, d'autre part, nous aurions pu relancer les répondants sur une réponse (exemple : « Quels types d'études seraient selon vous plus appropriés pour le Big Data ? Et pour quelles raisons spécifiquement pensez-vous que le sondage traditionnel ne pourrait pas répondre aussi efficacement ? Y a-t-il d'autres raisons que celles que vous avez évoquées ?»).

Enfin, nous avions souhaité dans un premier temps réaliser cette étude à l'aide de la méthode Delphi qui est très intéressante dans le cas de figure qui nous concerne. Sans que nous rentrions dans les détails, celle-ci consiste à interroger des experts de manière individuelle dans un premier temps. Ensuite, après avoir dépouillé les questionnaires, il est demandé de renvoyer aux experts les réponses anonymisées de tous les répondants pour établir si ces experts souhaitent rester sur leur position, s'ils souhaitent changer d'avis, mais on leur demande également de critiquer les avis qui leur sont opposés, etc. (Chirouze, 2007). Il va de soi qu'un tel procédé nous aurait demandé un temps conséquent dont nous ne disposions pas.

3. Perspectives de recherche futures

Nous avons dans le cadre de notre réponse à notre question de recherche établi que le Big Data et l'industrie du sondage traditionnel devraient plus que probablement coexister dans le futur, comme ils le font actuellement, mais devraient en plus aboutir à une forme de collaboration. Dès lors, nous pensons plus qu'intéressant d'essayer d'imaginer comment la technologie Big Data pourrait par exemple intégrer des instituts de sondages comme celui dans lequel nous travaillons. Nous rappelons que de grandes sociétés tentent déjà d'implémenter cette nouvelle solution (cf. supra p.82).

L'heure est selon nous venue d'imaginer un plan d'intégration concret pour ne pas être pris de court par la concurrence. Une telle intégration, si elle est possible, devrait permettre d'apporter un argument de poids aux clients d'instituts de sondages. Elle permettrait d'offrir un package complet aux clients qui auraient une vision tant globale que granulaire de son marché, mais comporterait également les raisons qui poussent le marché à évoluer dans telle ou telle direction.

Au cas où cette intégration n'est pas possible, car elle susciterait des investissements supplémentaires et des recrutements de data scientists (etc.), nous pourrions également investir de notre temps pour déterminer les formes de collaboration possibles et profitables pour toutes les parties prenantes.

Conclusion générale

À l'entame de notre étude, nous faisions part au lecteur de nos craintes vis-à-vis du Big Data. Une technologie qui, dans un premier temps, nous paraissait infaillible et étant en mesure de détrôner l'industrie du sondage traditionnel. Cette pensée, qui traversait notre esprit au début de notre processus de recherche, a très rapidement été balayée par nos premières lectures, nos premiers pas au sein de cet univers teinté de zones d'ombre, mais promettant monts et merveilles à ses utilisateurs. Dans un sens, nous pouvons admettre que le marketing promotionnel a fonctionné sur notre personne, ce qui pour le moins est peu flatteur pour un bachelier dans ce domaine.

Cependant, ce n'est pas pour autant que nous avons décidé de rester sur nos acquis et de nous contenter de quelques lectures, de quelques bribes d'informations disséminées ici et là par des chercheurs, spécialistes, experts, professionnels, etc. Nous avions une soif de connaissance suscitée par cette curiosité qui est la nôtre et qui souhaitait investiguer davantage sur cette thématique du Big Data. Ce choix de thème nous paraissait par ailleurs évident du fait que cette technologie représentait tout de même une forme de menace qui pesait sur notre secteur d'activité, l'industrie du sondage traditionnel.

En toute logique, nous formulions dès notre introduction notre question de départ qui, pour rappel, était la suivante : « Quel est l'impact du Big Data sur l'industrie du sondage traditionnel ? ». Ce questionnement nous a conduit à nous pencher dans un premier temps sur une description exhaustive de ces deux instruments. Le lecteur a pu, entre autres, établir quelles étaient les qualités et faiblesses de ces deux outils et quels étaient les apports (d'un point de vue théorique) du Big Data dans les domaines de prédilection de l'industrie du sondage traditionnel. Cette première étape nous a permis de disposer de tous les éléments théoriques nécessaires pour la suite de notre réflexion.

Après avoir posé les jalons de cette étude, il nous restait à préciser notre questionnement. Dès lors, nous avions opté pour la question de recherche suivante : « Quel est l'impact du Big Data au sein des domaines de prédilection de l'industrie du sondage traditionnel, à savoir : le monde politique, le secteur privé, le secteur de la santé et le secteur public, et, dès lors, quelles sont les conséquences des apports du Big Data sur les activités de l'industrie du sondage traditionnel ?». Une question que nous avons bien entendu scindée en deux parties afin de ne pas nous emmêler les pinceaux en cours de route et perdre le lecteur par la même occasion.

En début de notre seconde section, nous présentions donc au lecteur notre desk research qui démontrait des cas concrets d'application du Big Data, entrant en concurrence avec l'industrie du sondage traditionnel. Ensuite, nous laissions la parole aux professionnels, experts, etc. qui argumentaient preuve à l'appui sur la disparition prochaine de l'industrie qui est la nôtre ou d'une possible collaboration entre les deux outils. Enfin, nous avions agrémenté notre desk research d'une étude qualitative réalisée auprès d'« experts » qui avait pour objectif d'actualiser les informations que nous possédions déjà, à la suite de notre desk research. Cette étude supplémentaire nous a permis de recueillir des renseignements confirmant ceux que nous avions.

Notre développement nous a permis d'aboutir à un raisonnement que nous pensons avoir pris avec un recul critique suffisant, ce qui nous empêche ainsi que le lecteur d'adopter une position sectaire sur notre problématique. Le Big Data figure certainement parmi les outils auxquels fera appel la société de demain. Cette technologie qui, à l'instar du monde informatique, devrait fortement se perfectionner d'ici quelques années, permet déjà de collecter des données pertinentes pour le secteur privé, le secteur public, le secteur médical ainsi que le monde politique. Cependant, cet outil ne peut à lui seul apporter une réponse à tous les questionnements des différents domaines précités. Le Big Data est surtout en mesure de présenter les faits tels qu'ils se passent et où ceux-ci se passent. Cette récolte de données peut par ailleurs dans certains cas se faire en temps réel, ce qui devrait profiter à bon nombre de sociétés telles celles évoluant dans la grande distribution.

L'industrie du sondage traditionnel, quant à elle, s'avère être plus bénéfique pour les sociétés lorsque celles-ci s'intéressent aux raisons d'une action. Car, en effet, il nous est apparu peu probable que la technologie Big Data puisse dans tous les cas de figure nous renseigner sur le « pourquoi » d'une activité humaine. Les données récoltées par cet outil présentent une multitude d'actions, mais aucunement ne les accompagnent-elles d'une multitude de raisons.

Une disparition du Big Data et du sondage traditionnel nous est en l'occurrence apparu improbable. La question qui restait alors à combler était de savoir si une collaboration était possible entre les deux outils. Nous apprenions dans le cadre de nos recherches que celle-ci, assez timide, était déjà présente auprès des mastodontes de la donnée : Facebook et Google. Nous pouvons en déduire que les autres (Amazon, Microsoft…) ne tarderont pas à emprunter cette voie, si ce n'est pas encore le cas (ce que nous pensons être). Aussi, les instituts de sondages BVA et Opinion Way feraient déjà usage du Big Data sans pour autant se lancer totalement dans l'aventure (Rolland, 2017). L'ensemble des informations recueillies nous amène donc à penser qu'une collaboration reste la meilleure option pour ces deux outils auxquels nous souhaitons pour terminer cette conclusion rappeler la devise belge : l'union fait la force !

Bibliographie

Livres et ouvrages

- Amer-Yahia, S., Ganascia, J-G. et Ogier, J-M. (2017). Les sciences participatives et les sciences du numérique se complètent-elles ? In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.192-193). Paris : Édtions CNRS.
- Antoine, J. (2005). *Histoire des sondages.* Paris : Éditions Odile Jacob.
- Bidoit-Tollu, N. et Doucet, A. (2017). La diversité des données. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.24-25). Paris : Édtions CNRS.
- Buléon, P. (2017). Big Data, individu et société. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.36-37). Paris : Édtions CNRS.
- Chirouze, Y. (2007). *Le marketing. Études et stratégies* (2e éd). Paris : Éditions Ellipses.
- Corvol, P. (2017). Introduction. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (p.247). Paris : Édtions CNRS.
- Daydé, M. et Veynante, D. (2017). Les grands sites de calcul et de stockage. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.80-81). Paris : Édtions CNRS.
- Dehon, C., Droesbeke, J-J. et Vermandele, C. (2015). *Éléments de statistique.* Bruxelles : Éditions de l'Université de Bruxelles.
- Eveno, E. (2017). Big Data et Gouvernance. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.282-283). Paris : Édtions CNRS.
- Fanet, H. et Duranton, M. (2017). Les technologies au service du Big Data. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.26-27). Paris : Édtions CNRS.
- Frega, R. et Tsoukiàs, A. (2017). L'apport des Big Data dans les décisions publiques. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.292-293). Paris : Édtions CNRS.
- Gilbert, A. et Malcorps, C. (2012). *Techniques d'enquêtes de marchés par sondages.* Bruxelles : J-G Lahaye.
- Kitchin, R. (2014). *The Data revolution. Big Data, Open Data, Data Infrastructures & Their Consequences.* Londres : Sage Publishing
- Laudon, K., Laudon, J., Fimbel, É., Costa, S. et Canevet-Lehoux, S. (2013*). Management des systèmes d'information* (13e éd.). Montreuil : Éditions Pearson France.
- Marr, B. (2016). *Big data in practice. How 45 successful companies used big data analytics to deliver extraordinary results.* Chichester : Wiley Publishing.
- Meynaud, H.Y. et Duclos, D. (2007). *Les sondages d'opinion* (4e éd.). Paris : Éditions La Découverte.
- Mooi, E., Sarstedt, M. et Mooi-Reci, I. (2017). *Market Research. The Process, Data, and Methods Using Stata.* Singapour : Springer Nature Publishing.

- Pinker, S. (2017). Foreword. In Stephens-Davidowitz, S. *Everybody lies.* (pp.ix-xi). Londres : Bloomsbury Publishing.
- Rosenberg, D. (2013). Data before the fact. In Gitelman, L. *Raw Data is an Oxymoron*. (pp.15-40). Cambridge : MIT Press.
- Schafer, V. (2017). La déferlante des données : une courte mise en perspective d'une longue histoire. In Bouzeghoub, M. et Mosseri, R. *Les Big Data à découvert* (pp.22-23). Paris : Édtions CNRS.
- Stephens-Davidowitz, S. (2017). *Everybody lies.* Londres : Bloomsbury Publishing.
- Strong, C. (2015). *Humanizing Big Data. Marketing at the meeting of data, social science & consumer Insight*. Londres : Kogan Page Publishing.
- Vandercammen, M. et Gauthy-Sinéchal, M. (2014). *Études de Marchés. Méthodes et outils* (4e éd.). Louvain-la-Neuve : Éditions De Boeck Supérieur.

Articles de revue ou de journal

- Beauté, B. (2015, 6 septembre). Google Flu Trends vaincu par ses poussées de fièvre. *Tribune de Genève.* Récupéré de https://www.tdg.ch/sante/sante/Google-Flu-Trends-vaincu-par-ses-poussees-de-fievre/story/28554893
- Belga. (2017, 6 octobre). Les données de patients vendues par des hôpitaux: "Elles doivent être protégées", réagit De Block. *La Libre.be.* Récupéré de http://www.lalibre.be/actu/belgique/les-donnees-de-patients-vendues-par-des-hopitaux-elles-doivent-etre-protegees-reagit-de-block-59d71006cd70be70bcd3ef3b
- Berry, P. (2012, 8 novembre). Election américaine: «Big data», l'arme secrète d'Obama. *20 minutes.* Récupéré de https://www.20minutes.fr/monde/1038034-20121108-election-americaine-big-data-arme-secrete-obama
- Blondiaux, L. (1991). L'invention des sondages d'opinion : expériences, critiques et interrogations méthodologiques (1935-1950). *Revue française de science politique*, 6, 756-780. Récupéré de http://www.persee.fr/doc/rfsp_0035-2950_1991_num_41_6_394600
- Callegaro, M. et Yang, Y. (2017). The Role of Surveys in the Era of "Big Data". *The Palgrave Handbook of Survey Research* (pp.175-192). doi : 10.1007/978-3-319-54395-6_23
- Campbell, B. (1979, 10 février). Daniel Starch, Ad Analyzer, at 95. *The New York Times.* Récupéré de https://www.nytimes.com/1979/02/10/archives/daniel-starch-ad-analyzer-at-95.html
- Caulier, S. (2015). Les data centers, clés de voûte du réseau. *Le Monde.fr*. Récupéré de https://www.lemonde.fr/economie/article/2015/05/26/les-data-centers-cles-de-voute-du-reseau_4640752_3234.html
- Chamay, A. (2014). Des logiciels pour gérer sa campagne comme un produit commercial. *01net.com.* Récupéré de https://www.01net.com/actualites/des-logiciels-pour-gerer-sa-campagne-comme-un-produit-commercial-616428.html

- Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review*. Récupéré de https://hbr.org/2013/04/the-hidden-biases-in-big-data
- Daboll, P. (2013, 3 décembre). 5 Reasons Why Big Data Will Crush Big Research. *Forbes.* Récupéré de https://www.forbes.com/sites/onmarketing/2013/12/03/5-reasons-why-big-data-will-crush-big-research/#1cb099f65d0f
- De Fournas, M. (2018, 7 juin). Comment utiliser Reddit, le site qui a détrôné Facebook aux Etats-Unis. *20 minutes. Récupéré de https://www.20minutes.fr/high-tech/2285891-20180607-comment-utiliser-reddit-site-detrone-facebook-etats-unis*
- De Montcheuil, Y. (2014, 25 mars). Quand le big data révolutionne l'automobile. *La Tribune.fr*. Récupéré de https://www.latribune.fr/opinions/tribunes/20140325trib000821791/quand-le-big-data-revolutionne-l-automobile.html
- Dèbes, F. (2018, 19 février). Données personnelles : le RGPD en six questions. *Les Echos.fr*. Récupéré de https://www.lesechos.fr/19/02/2018/lesechos.fr/0301285400549_donnees-personnelles---le-rgpd-en-six-questions.htm
- Dening, P.J. (1990). Saving All the Bits. *American Scientist*, 78, 402-405. Récupéré de http://denninginstitute.com/pjd/PUBS/AmSci-1990-5-savingbits.pdf
- Diebold, F.X. (2000). "*Big Data" Dynamic Factor Models for Macroeconomic and Forecasting.* Récupéré de https://www.sas.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF
- Georis, V. (2017, 3 juin). "Le meilleur conseil en politique, c'est le big data". *L'Écho.* Récupéré de https://www.lecho.be/dossiers/presidentielles-francaises-2017/le-meilleur-conseil-en-politique-c-est-le-big-data/9900886.html
- Gershenfeld, N., Krikorian, R. et Cohen, D. (2004). The Internet of Things. *Scientific American.* Récupéré de http://fab.cba.mit.edu/classes/S62.12/docs/Cohen_Internet.pdf
- Groves, R.M. (2011). Three Eras of Survey Research. *Public Opinion Quarterly*, 75 (5), 861-871. doi : 10.1093/poq/nfr057.
- Hays, C.L. (2004, 14 novembre). What Wal-Mart Knows About Customers' Habits. *The New York Times.* Récupéré de https://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html?mtrref=www.google.be
- Khalifa, M. et Zabani, I. (2016). Utilizing health analytics in improving the performance of healthcare services: A case study on a tertiary care hospital. *Journal of Infection and Public Health,* 9 (6), 757-765. doi : 10.1016/j.jiph.2016.08.016
- Kitchin, R. et McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3, 1-10. doi : 10.1177/2053951716631130

- Kosinski, M., Stillwell, D. et Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110 (15), 5802-5805. doi : 10.1073/pnas.1218772110
- L'Express.fr. (2018, 18 mars). Élection de Trump: le hold-up de Cambridge Analytica sur les usagers de Facebook. *L'Express.fr.* Récupéré de https://www.lexpress.fr/actualite/monde/amerique-nord/election-de-trump-le-hold-up-de-cambridge-analytica-sur-les-usagers-de-facebook_1993257.html
- Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies.* Récupéré de https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
- Le Monde.fr. (2013, 26 octobre). New York, un an après l'ouragan Sandy, en images. *Le Monde.fr*. Récupéré de https://www.lemonde.fr/planete/article/2013/10/26/sandy-new-york-un-an-apres-l-ouragan-en-images_3503348_3244.html
- Lohr, S. (2013, 1 février). The Origins of 'Big Data': An Etymological Detective Story. *The New York Times.* Récupéré de https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/
- Lohr, S. et Singer, N. (2016, 10 novembre). How Data Failed Us in Calling an Election. *The New York Times.* Récupéré de https://www.nytimes.com/2016/11/10/technology/the-data-said-clinton-would-win-why-you-shouldnt-have-believed-it.html?rref=collection%2Fbyline%2Fsteve-lohr&action=click&contentCollection=undefined&region=stream&module=stream_unit&version=latest&contentPlacement=81&pgtype=collection
- Mann, S., Nolan, J. et Wellman, B. (2003). Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments. *Surveillance & Society*, 1(3), 331-355. Récupéré de https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/3344/3306
- Nickerson, D.W. et Rogers, T. (2013). Political Campaigns and Big Data. *HKS Faculty Research Working Paper Series.* Récupéré de https://scholar.harvard.edu/files/todd_rogers/files/political_campaigns_and_big_data_0.pdf
- Ollion, É. et Boelaert, J. (2015). Au-delà des big data. Les sciences sociales et la multiplication des données numériques. *Sociologie*, 6, 295-310. Récupéré de https://www.cairn.info/revue-sociologie-2015-3-page-295.htm
- Pouillon, J. (1951). Les sondages et la science politique. *Revue française de science politique*, 1 et 2, 83-106. Récupéré de https://www.persee.fr/doc/rfsp_0035-2950_1951_num_1_1_392074
- Press, G. (2013, 9 mai). A Very Short History Of Big Data. *Forbes.* Récupéré de https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#10f1156265a1

- Rolland, S. (2017). Présidentielle : le big data éliminé dès le premier tour, revanche des sondages. *La Tribune.fr*. Récupéré de https://www.latribune.fr/technos-medias/presidentielle-le-big-data-elimine-des-le-premier-tour-revanche-des-sondages-695888.html
- Signoret, P. (2015, 22 avril). Le crowdsourcing, ou l'art de faire travailler gratuitement le client. *Trends-Tendances.* Récupéré de http://trends.levif.be/economie/entreprises/le-crowdsourcing-ou-l-art-de-faire-travailler-gratuitement-le-client/article-normal-390463.html
- The Economist. (2010, 25 février). All too much. Monstrous amounts of data. *The Economist.* Récupéré de https://www.economist.com/node/15557421

Sites internet et pages web

- Accountlearning.com. (2018). *Advantages and Disadvantages of Sampling*. Récupéré le 3 juin 2018 de https://accountlearning.com/advantages-and-disadvantages-of-sampling/
- AdAge. (2003). *Testing Methods.* Récupéré le 14 août 2018 de http://adage.com/article/adage-encyclopedia/testing-methods/98903/
- Angeles, S. (2013). *Cloud vs. Data Center: What's the difference?* Récupéré le 10 mai 2018 de https://www.businessnewsdaily.com/4982-cloud-vs-data-center.html
- Balagué, C. (2017). *Opportunités du Big Data et des données issues des réseaux sociaux.* Récupéré le 9 mai 2018 de http://www.soft-concept.com/surveymagazine/opportunites-big-data-et-donnees-reseaux-sociaux/
- Baron, Léa. (2016). *Election présidentielle américaine : comment ça marche ?* Récupéré le 11 août 2018 de https://information.tv5monde.com/info/election-presidentielle-americaine-comment-ca-marche-3739
- Big Data Corp. (s.d.). *Accelerate The Future!*. Récupéré le 11 août 2018 de https://www.bigdataparis.com/2019/
- Bresnick, J. (2017). *56% of Hospitals Lack Big Data Governance, Analytics Plans*. Récupéré le 6 juillet 2018 de https://healthitanalytics.com/news/56-of-hospitals-lack-big-data-governance-analytics-plans
- Commission Européenne. (s.d.). *Quelles données peut-on traiter et sous quelles conditions?* Récupéré le 25 juin 2018 de https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/what-data-can-we-process-and-under-which-conditions_fr
- Contemporary Analysis. (2018). *Political Campaigns & Predictive Analytics: Changing how to campaign.* Récupéré le 8 juillet 2018 de http://canworksmart.com/case-study/political-campaigns-and-big-data/
- Dimensional Insight. (2017). *Dimensional Insight Dives into the Data Governance Challenges Plaguing U.S. Hospitals*. Récupéré le 6 juillet 2018 de https://www.dimins.com/press-releases/yr2017/data-governance-challenges-plague-hospitals/

- Dontha, R. (2017). *The Origins of Big Data*. Récupéré le 27 avril 2018 de https://www.kdnuggets.com/2017/02/origins-big-data.html
- Eduardo Kokoyachuk, R. (2017). *How Does Market Research Stay Relevant In A Big Data World?* Récupéré le 12 août 2018 de https://thinknowresearch.com/blog/how-does-market-research-stay-relevant-in-a-big-data-world/
- Encyclopædia Universalis. (2018). *NEYMAN JERZY (1894-1981)*. Récupéré le 23 mai 2018 de https://www.universalis.fr/encyclopedie/jerzy-neyman/
- Floridi, C. (2018). *How FMCG companies can harness the power of Big Data to drive growth in challenging times*. Récupéré le 16 juin 2018 de https://www.datalab-crm.de/big-data-for-fmcg/?lang=en
- Foursquare. (2018). *À propos de nous.* Récupéré le 5 août 2018 de https://fr.foursquare.com/about
- Frintz, A. (2015). *Les pays en voie de développement sont-ils «connectés»?* Récupéré le 30 avril 2018 de http://www.rfi.fr/hebdo/20150410-pays-voie-developpement-technologies-internet-telephonie-mobile-vie-quotidienne-connexion
- Futura. (2017). *Internet des objets.* Récupéré le 10 mai 2018 de https://www.futura-sciences.com/tech/definitions/internet-internet-objets-15158/
- GDPR Associates. (2018). *Understanding GDPR Fines.* Récupéré le 16 juin 2018 de https://www.gdpr.associates/what-is-gdpr/understanding-gdpr-fines/
- Giacometti, P. (2001). *Intentions de vote : que mesurent les instituts de sondage ?* Récupéré le 13 mai 2018 de https://www.ipsos.com/fr-fr/intentions-de-vote-que-mesurent-les-instituts-de-sondage
- Giezendanner, F.D. (2012). *Taille d'un échantillon aléatoire et Marge d'erreur.* Récupéré le 30 mai 2018 de http://icietla-ge.ch/voir4/IMG/pdf/taille-d_un-echantillon-aleatoire-et-marge-d_erreur-cms-spip.pdf
- Gouvernement du Québec. (2018). *Thésaurus de l'activité gouvernementale.* Récupéré le 5 mai 2018 de http://www.thesaurus.gouv.qc.ca/tag/terme.do?id=6837
- Grinberg, N., Naaman, M., Shaw, B. et Lotan, G. (2013). *Extracting Diurnal Patterns of Real World Activity from Social Media.* Récupéré le 8 août 2018 de https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6087/6359
- Hing Lo, S. (2016). *The impact of big data and technology on the market research industry: where are we at?* Récupéré le 25 juillet 2018 de https://www.accent-mr.com/sr/the-impact-of-big-data-and-technology-on-the-market-research-industry-where-are-we-at/
- Hon Ho, D.Y. (2016). *How Market Research Has Evolved Over the Past Decade.* Récupéré le 30 avril 2018 de https://www.marketstrategies.com/blog/2016/07/how-market-research-has-evolved-over-the-past-decade/
- HuffPost Pollster. (2017). *France Presidential Election.* Récupéré le 10 août 2018 de https://elections.huffingtonpost.com/pollster/france-presidential-election-round-1

- King Faisal Specialist Hospital and Research Center. (2018). *Organization Structure*. Récupéré le 6 juillet 2018 de https://www.kfshrc.edu.sa/en/home/aboutus/organizationstructure
- Kobielus, J. (2018). *Wikibon's 2018 Big Data Analytics.* Récupéré le 25 juillet 2018 de https://wikibon.com/wikibons-2018-big-data-analytics-trends-forecast
- Lucchese, V. (2017). *Top 10 des projets de science participative.* Récupéré le 13 mai 2018 de https://usbeketrica.com/article/top-10-des-projets-de-science-participative
- Ministère de l'Intérieur. (2017). *Résultats de l'élection présidentielle 2017.* Récupéré le 10 août 2018 de https://www.interieur.gouv.fr/Elections/Les-resultats/Presidentielles/elecresult__presidentielle-2017/(path)/presidentielle-2017/FE.html
- Roux, T. (2017). *Ces data-brokers qui font commerce de nos données personnelles.* Récupéré le 29 avril 2018 de https://atelier.bnpparibas/smart-city/article/data-brokers-commerce-donnees-personnelles
- Sager Weinstein, L. (2017). *Women in big data: Why business intelligence and data strategy are the future of transport.* Récupéré le 7 juillet 2018 de http://www.womanthology.co.uk/women-big-data-business-intelligence-data-strategy-future-transport-lauren-sager-weinstein-chief-data-officer-transport-london/
- Silver, N. (2018). The *Polls Are All Right.* Récupéré le 10 août 2018 de https://fivethirtyeight.com/features/the-polls-are-all-right/
- Slessareva, L. (2016). *Why We Need Market Research to Get the Most Out of Big Data.* Récupéré le 5 août 2018 de https://www.aaaa.org/need-market-research-get-big-data/
- Transport for London. (2017). *Review of the TfL WiFi pilot*. Récupéré le 7 juillet 2018 de http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf
- Transport for London. (s.d.). *What we do.* Récupéré le 7 juillet 2018 de https://tfl.gov.uk/corporate/about-tfl/what-we-do?intcmp=2582
- Tronchet, S. (2017). *Sondages : peut-on (encore) leur faire confiance ?* Récupéré le 5 juin 2018 de https://www.franceinter.fr/politique/sondages-peut-on-encore-leur-faire-confiance
- Van Rijmenam, M. (2018). *Walmart Is Making Big Data Part Of Its DNA*. Récupéré le 5 juillet 2018 de https://datafloq.com/read/walmart-making-big-data-part-dna/509
- Van Rijmenam, M. (2018). *Will Big Data Mark The End Of The Market Research Industry?* Récupéré le 20 juillet 2018 de https://datafloq.com/read/will-big-data-mark-the-end-of-the-market-research-/210
- Walmart Staff. (2017). *5 Ways Walmart Uses Big Data to Help Customers.* Récupéré le 5 juillet 2018 de https://blog.walmart.com/innovation/20170807/5-ways-walmart-uses-big-data-to-help-customers
- Walmart. (2018). *Our Business.* Récupéré le 5 juillet 2018 de https://corporate.walmart.com/our-story/our-business

- Wyner, G. (2017). *Does Market Research Need a Makeover?* Récupéré le 2 juillet 2018 de https://www.ama.org/publications/MarketingNews/Pages/market-research-uncertain-future.aspx

Rapports

- Dumoulin, M. et Sterckmans, W. (2017). *Baromètre politique trimestriel.* Bruxelles : Dedicated.
- Esomar. (2014). *Global Market Research 2014.* Amsterdam : Esomar.
- Esomar. (2017). *Global Market Research 2017.* Amsterdam : Esomar.
- Greater London Authority. (2018). *Mayor's Transport Strategy*. Londres : Greater London Authority. Récupéré de https://www.london.gov.uk/sites/default/files/mayors-transport-strategy-2018.pdf
- Portelli, H. et Sueur, J-P. (2010). *Sondages et démocratie : pour une législation plus respectueuse de la sincérité du débat politique.* Paris : Sénat. Récupéré de http://www.senat.fr/rap/r10-054/r10-054_mono.html#toc216

Sources orales

- Hilton, S., Huffman, S. et Rabois. K. (2016, 10 novembre). Why Big Data Failed to Predict the U.S. Election [TV Shows]. *Bloomberg Technology*. Récupéré de https://www.bloomberg.com/news/videos/2016-11-10/why-big-data-failed-to-predict-the-u-s-election
- Wylie, C. (2018, 17 mars). How Cambridge Analytica turned Facebook 'likes' into a lucrative political too*l* [Interview video]. *The Guardian*. Récupéré de https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm