BigData et protection des données personnelles : quels enjeux ?
Éléments de réflexion

Sophie VULLIET-TAVERNIER
Directeur des relations avec les publics et la recherche
Commission nationale de l'Informatique et des Libertés

Bon nombre des applications du BigData touchent à des activités ou comportements humains, et mobilisent donc des données personnelles. Les spécificités du BigData sont souvent présentées comme susceptibles de remettre en cause les principes cardinaux de la protection de ces données, ou l'applicabilité des dispositions légales prohibant certains usages des algorithmes. Dans le passé, la Commission informatique et libertés, se prononçant sur des applications de datamining, a su trouver des solutions pour permettre une application adaptée de la législation. En concertation avec l'ensemble des acteurs concernés, elle se donne aujourd'hui pour priorité de rechercher de la même façon des solutions d'accompagnement aux projets de BigData.

Sous un intitulé un peu « attrape-tout », l'expression BigData permet en réalité de prendre conscience des capacités nouvelles de traitement de données apparues ces dernières années. Au-delà des développements techniques spécifiques[1], le BigData est couramment appréhendé comme concernant des traitements d'ensembles de données dont les trois caractéristiques principales (volume, vélocité et variété[2]) conduisent à s'interroger sur l'applicabilité des règles de protection des données personnelles.

Certes, le BigData n'implique pas nécessairement des traitements de données personnelles : le concept est beaucoup plus large[3]. Mais dans la réalité, bon nombre des applications concrètes du BigData touchent directement ou indirectement à des activités ou comportements humains (que ce soit dans le domaine du commerce, de la santé, des transports, des assurances…).

En effet, le BigData appliqué aux données personnelles offre notamment la possibilité d'une connaissance plus fine de populations ciblées et le cas échéant la construction de modèles prédictifs de comportements (voire de prise de décision) grâce au traitement de masse de données structurées comme non structurées (et issues de multiples sources dont le web social et les objets connectés) et à des algorithmes d'analyse sophistiqués.

1. Par exemple HADOOP, le NoSQL, MapReduce, …
2. La vélocité renvoie aux capacités d'analyse et de traitement lesquelles évoluent de manière exponentielle. La variété renvoie à la diversité des formats de données désormais susceptibles d'être traitées : bases non structurées, qui peuvent prendre la forme de fichiers audios ou vidéos, de données collectées issues du web social, de capteurs… Enfin, le volume est le résultat de l'évolution des 2 premières caractéristiques qui amènent les entreprises à traiter de quantités très importantes de données.
3. Et intéresse par exemple l'exploitation de données dans des domaines aussi divers que la météorologie, la géologie…

Ces types de traitements ne sont pas nouveaux pour la Commission Nationale de l'Informatique et des Libertés (CNIL). La Commission a ainsi déjà eu à connaitre d'applications en infocentre ou de datamining reposant sur l'exploitation statistique de bases de données internes, par exemple, pour avoir une meilleure connaissance de catégories de populations (ex : demandeurs d'emploi, contribuables, assurés sociaux, demandeurs de crédits...), déterminer des profils de personnes vulnérables (par exemple dans le domaine sanitaire et social) ou encore détecter des comportements à risques ou anormaux (notamment en matière de lutte contre la fraude sociale, fiscale ou encore dans le cadre du crédit scoring)[4].

Le BigData apporte cependant une dimension nouvelle à ces objectifs de connaissance et de profilage, du fait de l'explosion du volume des données (et notamment du développement des capteurs et objets connectés), et de la mise à disposition d'outils d'analyse toujours plus puissants. Mais s'agit-il là d'un changement d'échelle dans la fouille de données ou d'un véritable changement de nature ?

Les spécificités du Big Data sont souvent présentées comme susceptibles de remettre en cause ou en tout cas de questionner certaines notions clés ou principes cardinaux de la protection des données. Qu'en est-il en réalité? Notre droit et en particulier notre droit de la protection des données est-il adapté pour répondre à ces enjeux ?

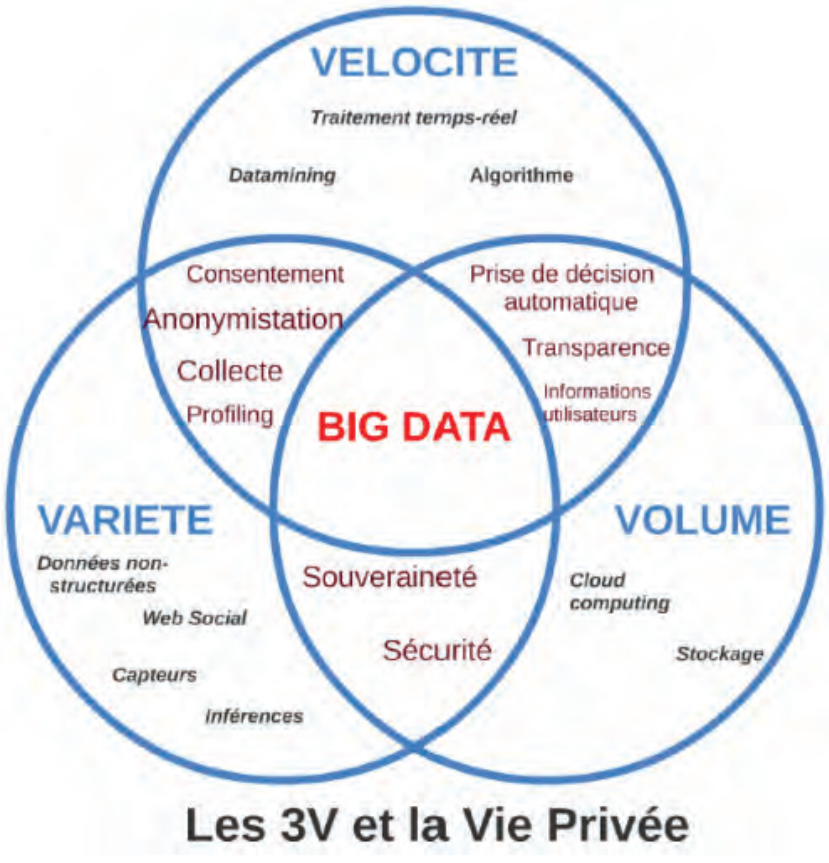

Figure 1 : Les mots-clés du BigData (source CNIL)

4. Cf par exemple, les avis rendus sur la segmentation comportementale dans le domaine bancaire,(1993) sur des systèmes d'aide à la décision dans le domaine social (ex SIAM et SNIIRAM pour l'assurance maladie) , aide à la sélection et au contrôle fiscal des particuliers (SIRIUS) , outil statistique d'aide à la connaissance des demandeurs d'emploi SIAD...

Le concept de donnée personnelle : toutes les données deviennent-elles identifiantes ?[5]

Outre le fait que les outils du BigData peuvent porter sur des données directement identifiantes, ils peuvent aussi conduire à ce que des données anonymes à l'origine, par recoupement avec d'autres données, permettent de déduire plus d'informations sur les personnes, voire de les identifier ou de les ré-identifier . Peut-on alors en déduire que toute donnée ou « trace » devrait être qualifiée de potentiellement identifiante, voire de donnée personnelle ? Comment assurer dès lors une anonymisation efficace et protectrice des individus sans faire perdre toute valeur scientifique aux analyses de données ?

La CNIL a toujours donné une interprétation large du concept de donnée personnelle en prenant en compte notamment :

- la nature des données : ex. initiales des noms et prénoms, date et lieu de naissance, commune de résidence, lieu de travail, nature de l'emploi, indications de dates (d'examens, d'hospitalisation, etc.), métadonnées (adresse IP, données de géolocalisation…)
- l'importance relative de l'échantillon de population concernée ;
- le type de traitement effectué : ex. datamining.

Au plan européen, le Groupe européen des autorités de protection des données (G29), dans son avis du 10 avril 2014[6] sur les techniques d'anonymisation apporte un éclairage intéressant sur les différentes méthodes d'anonymisation.

Finalité et pertinence : des principes remis en question ?

En application de ces principes, si l'on résume, seules peuvent être collectées et traitées les données strictement nécessaires à la poursuite de finalités déterminées, explicites et légitimes (ce qui suppose qu'elles aient été préalablement définies). Or, il pourrait être considéré que la logique du BigData est inverse : on recueille le maximum de données et on définit ensuite quel(s) usage(s) on peut en faire. Cependant, une démarche BigData bien construite n'implique-t-elle pas en amont de réfléchir, à tout le moins, sur l'objectif poursuivi (ex : lutte contre la fraude, connaissance des pratiques de consommation d'un segment de population…) et sur les catégories de bases de données disponibles ? Est-on si loin finalement de la démarche d'analyse « informatique et libertés » ?

En outre, les techniques de BigData appliquées concrètement par les professionnels d'un secteur d'activité donné (ex : assurances, banques…) sont normalement sous-tendues par la poursuite de finalités s'inscrivant a priori dans le cadre de leurs activités.
La question est plus délicate lorsqu'il s'agit pour un acteur, par exemple un opérateur de télécommunications, d'exploiter ses bases de données de gestion – ses données de trafic - pour des objectifs de meilleure connaissance des déplacements de populations (par exemple trajectoires de transports d'une population donnée), de répartition de foyers épidémiques, de fréquentation de zones de chalandage, de comptage de manifestants…

Il en est de même pour des applications de BigData reposant sur l'exploitation de données provenant de sources diverses (bases de données internes d'operateurs, données du web social) et ce pour des finalités parfois fort éloignées des finalités initiales.

5. Cf article 2 loi informatique et libertés : constitue une donnée à caractère personnel **toute information relative à une personne physique identifiée ou qui peut être identifiée**, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.
6. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Dans le passé, la CNIL se prononçant sur des applications de datamining a su trouver des solutions pour permettre une application adaptée des principes de protection des données : solutions d'anonymisation (hachage des identités, restriction sur certaines requêtes, interdiction de certains croisements de données…), mesures de sécurité (traçabilité des accès, nombre d'utilisateurs limité…), catégories de familles d'usages jugées compatibles.

Par ailleurs, les traitements BigData ont vocation à réutiliser des jeux de données personnelles initialement recueillies pour une autre fin : à cet égard, l'article 6 de la loi ouvre des possibilités de réutilisation notamment à des fins statistique et scientifique (les termes mériteraient sans doute d'être précisés) ou encore, en application de l'article 36, pour d'autres finalités sous réserve du consentement de la personne ou encore avec l'autorisation de la CNIL. Ces possibilités mériteraient d'être explorées plus avant dans le cadre de la problématique BigData.

La « gouvernementalité algorithmique »

L'article 10 de la loi informatique et libertés prohibe toute décision produisant des effets juridiques, prise à l'égard d'une personne sur le seul fondement d'un traitement destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité. Or, le vrai moteur du BigData tient dans les algorithmes qui ont souvent pour but d'être « prédictifs », c'est-à-dire de détecter des corrélations et d'anticiper des situations voire de prendre des décisions collectives voire individuelles par le biais des analyses statistiques (la « gouvernementalité algorithmique[7] »). Se pose dès lors la question de l'applicabilité éventuelle de l'article 10. De façon corollaire, se pose aussi la question de l'application effective du principe selon lequel toute personne se voit reconnaître le droit d'obtenir les informations permettant de connaitre et de contester la logique qui sous-tend le traitement automatisé en cas de décision prise sur le fondement de celui-ci et produisant des effets juridiques à l'égard de l'intéressé[8] (l'exercice de ce droit a ainsi pu être invoqué par exemple en matière de credit scoring).

Loyauté de la collecte et respect des droits des personnes

Comment informer les personnes sur les conditions d'exploitation de leurs données et sur les droits, comme le prescrit la loi, si l'on ne connait pas a priori la finalité de cette exploitation, ou encore s'il s'agit de données très indirectement nominatives ?

La loi apporte des éléments de réponse. En effet, en cas de réutilisation de données, l'obligation d'information ne s'applique pas, aux termes de l'article 32 III : « quand son information se révèle impossible ou exige des efforts disproportionnés par rapport à l'intérêt de la démarche » , ce qui vise notamment le cas de collectes de données très indirectement identifiantes ou de personnes perdues de vue.

Cette dérogation pourrait sans doute être invoquée pour les cas dans lesquels des traitements de BigData porteraient sur des données qui au départ ne sont pas directement identifiantes mais le deviendraient par croisement.

Elle est à rapprocher de la dérogation au droit d'accès, prévue à l'article 39 II lorsqu'il s'agit de données « conservées sous une forme excluant manifestement tout risque d'atteinte à la vie privée des personnes concernées et pendant une durée n'excédant pas celle nécessaire aux seules finalités d'établissement des statistiques ou de recherche scientifique ou historique ».

De façon connexe, se pose la question de l'utilisation des données issues du web : toutes les

7. A ce sujet, voir les travaux d'Antoinette Rouvroy : *Gouvernementalité algorithmique et perspectives d'émancipation*
8. Art 39 5°

données du web social sont-elle utilisables librement ?

Quel statut accorder à ces informations publiquement accessibles, « à portée de main », sur internet et donc a priori facilement réutilisables ? Ces données peuvent-elles être ré-exploitées par des tiers, sans que les personnes concernées en aient été informées et aient pu faire valoir leur point de vue et exercer leurs droits? A l'inverse serait-il réaliste d'exiger une information ou a fortiori un accord systématique ? Pour quelles finalités peuvent-elles être réutilisées? Autant de questions qui méritent assurément une réflexion concertée avec l'ensemble des acteurs concernés et une réponse adaptée aux spécificités du web social.

Quelles mesures particulières de sécurité ?

Compte tenu de l'ampleur des bases de données constituées et des capacités de traitement des données, des mesures de sécurité adaptées doivent pouvoir être prises pour à la fois assurer la traçabilité des opérations de traitements effectués et contrôler les accès à ces bases, points sur lesquels la CNIL a toujours insisté lorsqu'elle s'est prononcée sur des applications de datamining.

Au delà, le recours à des méthodes d'anonymisation des données devrait être largement promu et faire partie intégrante de la conception des projets BigData impliquant des croisements de bases de données provenant d'acteurs tiers.

Enfin, en termes de stockage des données, les traitements BigData reposeront généralement sur le *cloud computing*, seul à même de fournir espace, souplesse et vitesse nécessaires. En matière de protection des données personnelles, les enjeux du cloud computing, bien connus, sont à la fois juridiques (notamment en ce qui concerne la qualification des parties, la responsabilité du prestataire et les transferts) et techniques (niveau de sécurité du prestataire, risque de perte de gouvernance sur le traitement, dépendance technologique vis-à-vis du fournisseur de cloud, absence d'information sur ce que fait réellement le prestataire...). Enfin, le *cloud* est devenu accessible à tous en raison de sa capacité à passer à l'échelle : les entreprises adaptent le besoin à l'usage quasiment en temps réel, sans mobiliser inutilement de la puissance de calcul ou de stockage en permanence et payent à l'usage sur des plateformes tierces. Comment dans ce cas garantir le lieu et la sécurité du stockage des données ? Quel droit appliquer à des données qui peuvent passer des frontières selon les besoins techniques de l'opérateur de *cloud*, souvent sans que leur propriétaire ne le sache ? Sur ces différents points la CNIL à la suite d'une large concertation publique, a adopté un certain nombre de recommandations disponibles sur son site[9].

En résumé, il apparaît sans doute nécessaire de dresser une typologie des problématiques et des réponses à apporter en fonction des traitements BigData et des acteurs concernés : il est évident que ces problématiques et les solutions ne sont pas du même ordre selon que l'on pratique du datamining sur ses propres bases pour un objectif de meilleure connaissance de la population que l'on gère, ou que l'on souhaite procéder à des croisements de jeux de données issues de sources externes[10] pour une mise en commun dans un « puits de données » et pour en tirer des usages et services nouveaux.

User de pédagogie, engager la concertation et rechercher des solutions d'accompagnement aux projets de BigData (notamment sous la forme de préconisations en matière d'anonymisation) sont aujourd'hui la priorité pour la CNIL.

9. http://www.cnil.fr/linstitution/actualite/article/article/cloud-computing-les-conseils-de-la-cnil-pour-les-entreprises-qui-utilisent-ces-nouveaux-services/
10. Avec les conséquences qui en résultent sur le plan du régime de formalités préalables applicables : demande d'autorisation en cas d'interconnexions de traitements présentant des finalités différentes.