LA DATA EST-ELLE ÉTHIQUE-COMPATIBLE ?

Philippe Tassi[1]

[1] *Médiamétrie, 70 rue Rivay, 92532 Levallois-Perret Cedex, ptassi@mediametrie.fr*

Résumé. La thématique générale est celle des avancées permises par des sujets à formaliser ou à résoudre dans le monde des entreprises, où il y a beaucoup de questions mal posées, ou mal résolues, et bien sûr des questions nouvelles. La révolution récente due à la convergence numérique fait du secteur des médias, au sens large, un domaine de réelle création scientifique. Dernier phénomène apparu, celui des Big Data - les « mégadonnées » -. Elles constituent un retour vers l'exhaustif, le paradigme existant jusqu'à la fin du XIXème siècle, l'exposé d'Anders Kiaer à l'IIS de Berne, le rapport de la Commission Jensen de 1925 et l'article de Neyman en 1934. Mais exhaustif ne signifie pas perfection absolue, les données qui en résultent sont parfois de granularité plus grossière que celles issues d'échantillons. Néanmoins, deux exemples de réflexions innovantes apportées par les Big Data : l'hybridation entre données d'échantillonnage et données exhaustives d'une part, l'éthique au sens du respect de l'intimité et de la vie privée (« privacy ») d'autre part. Privacy et bases de données sont un champ déjà abordé au milieu des années 70, avec un article précurseur de Tore Dalenius publié dans Statistik Tidskrift (1977), « Towards a methodology for statistical disclosure control ». Depuis, deux approches apparaissent : la première est plutôt européenne et basée sur des interdits juridiques ; la seconde, nord-américaine, dénommée « differential privacy », repose sur des fondements probabilistes et statistiques.

Mots-clés. Mégadonnées, exhaustivité, hybridation, confidentialité

Echantillonnage ou collecte peu structurée ?

Depuis le début des années 2000, on assiste à un retour vers l'exhaustif dans certains domaines, grâce aux nouvelles technologies et à la convergence numérique.

Dans certains domaines, en effet, la montée en puissance du digital entraîne le retour de l'exhaustivité ou quasi-exhaustivité (« voie de retour », « return path data »).

- *Convergence numérique, digital*
- *Télécom, Médias, Internet (smartphones, tablettes, TV connectées, box ADSL, …)*
- *Consommation (cartes de fidélité)*

Du point de vue juridique, on peut considérer que le fait d'accepter de faire partie d'un échantillon, d'un panel est un acte volontaire donnant naissance à l'établissement d'un « contrat » entre la société d'études organisant l'échantillon et le panéliste et ses données à caractère personnel (DCP).

Le recueil automatique de données sur des populations de plus en plus grandes a parfois changé la donne. Les bases de données contiennent de plus en plus d'informations, annonçant le potentiel retour du paradigme de l'exhaustif.

Sont apparues ainsi ce qu'on appelle les Big Data, ou plutôt, pour les français, les mégadonnées ainsi que le recommande depuis cet été la Commission Nationale de Terminologie, qui montre ainsi sa détermination à lutter contre le latin, puisque data n'est autre que le pluriel de la forme A du supin du verbe do, donner en latin.

Les Big Data ont deux dimensions majeures pour définir la volumétrie :

Volumétrie = Quantité x Fréquence

Quantité pouvant aller jusqu'à l'Exhaustivité, fréquence pouvant aller jusqu'au temps réel.

A ces Big Data on a coutume d'associer des V, historiquement 3, puis plus récemment 6. De quoi s'agit-il ?

Les 3 V « primitifs » sont :

- Volume : quantité de données échangées (recueil, stockage et traitement),
- Variété : diversité des formats (texte, audio, image, vidéo), des sources (sites, réseaux sociaux, téléphones, RFID, GPS, ..), des origines (données internes structurées, externes non structurées),
- Vélocité : recueil et traitement des données en « temps réel ».

Ensuite ont été ajoutés les 3 autres V, plus « marketing » :

- Véracité : confiance dans l'information recueillie,
- Visualisation : traitement optimisé des données pour l'aide à la décision,
- Valeur : création de valeur pour l'entreprise.

Les Big Data ont apporté des réflexions innovantes ; sur les méthodes d'hybridation entre données d'échantillonnage et données exhaustives d'une part, et sur « l'éthique » au sens du respect de l'intimité et de la vie privée (« privacy ») d'autre part.

Question 1 : on peut faire quoi avec les Big Data ?

Le quotidien fournit de nouveaux exemples d'utilisation de ces données nombreuses, dans des domaines d'activité en progression permanente : médecine et santé, assurances, sport, marketing, culture…

Trois illustrations : le brevet d'Amazon pour anticiper les commandes de ses clients et commencer à envoyer les colis avant la confirmation de la commande ; les prévisions de Netflix sur ce que veulent regarder les abonnés à son service de diffusion de films et de séries ; la prévision de la prochaine destination de voyage d'une personne.

Les Big Data sont « big » en deux sens :

1- En quantité et en variété de données disponibles (les 6 V)

2- Par l'étendue des analyses qui peuvent y être appliquées pour faire de l'inférence

L'inférence de la statistique classique va de l'échantillon à la population. Le respect des données à caractère personnel repose sur une inférence « perverse », une inférence inversée : de la base de données vers l'individu i.

En outre, les traitements, les modèles, les algorithmes, bref le data mining, peuvent permettre d'accéder non seulement à des données personnelles contenues dans les data, mais aussi à des données non contenues, des variables estimées. Quel est le statut de cette donnée non recueillie mais approchée : estimation, loi de probabilité (« profiling ») ?

Par exemple, le surf d'une adresse IP sur des sites S, S', S'' … dont les structures socio-démographiques des visiteurs sont connues permet assez simplement d'approcher le profil de l'utilisateur et même de le classer éventuellement dans un groupe « éphémère ».

Enfin, il nous faut bien intégrer que si les réflexions sont actuelles, les progrès de la science créeront de nouveaux modèles, de nouveaux algorithmes, conçus à une date postérieure à la constitution de la base de données.

Question 2 : faut-il opposer Big Data et échantillons ?

De même que dans les années 1990 les mégabases de données (mode déclaratif, Claritas, Acxiom) n'ont pas tué l'échantillonnage, je ne crois pas à la victoire des Big Data et à la disparition des sondages.

L'une des raisons principales est que la finesse et la granularité des données ne sont pas de même nature.

Il y a 40 ans, en statistique d'entreprise, on avait coutume de rapprocher, déjà, les donnée d'enquêtes – échantillons – des EAE des données contenues dans les diverses déclarations fiscales.

Dans le domaine des médias, ainsi que divers communications et articles l'ont montré, les données exhaustives et en temps réel sont souvent attachées à un « device ». Mais un objet n'est pas un individu. La plupart d'entre eux ont plusieurs utilisateurs. Et « exhaustif » ne signifie pas « sans erreur », ou parfait.

En médias, nous avons donc rapproché les comportements d'un panel d'internautes avec l'ensemble des visites et des pages vues des sites. Cette approche hybride, appelée « panel up », est même la référence de la mesure d'audience de l'Internet via un PC depuis l'été 2012.

En télévision, nous estimons les comportements individuels des personnes vivant dans un foyer équipé d'une box CanalSat à partir des éléments du panel TV, en utilisant un modèle, un algorithme développé par les équipes de Médiamétrie dans le cadre de la théorie de chaînes de Markov cachées.

Ce mélange de data de natures différentes est même une fabuleuse opportunité pour nos métiers, au sens large.

Là où, il y a encore quelques années, on observait en général un phénomène via un échantillon, on avait donc une « brique » d'observation avec une matière homogène, maintenant on construit un mur avec plusieurs briques, et on a besoin de ciment : ce ciment, c'est notre métier, l'observation, les sondages, la statistique mathématique, la modélisation, l'utilisation optimale des techniques mathématiques.

Mais finalement, n'est-on pas toujours dans le droit fil de ce que Carl-Eric Särndal, Jean-Claude Deville, Pascal Ardilly, et bien d'autres ont toujours proclamé : quand il existe une information auxiliaire, il faut chercher à l'utiliser ?

Question 3 : est-ce un phénomène de mode ?

Là encore, je ne le crois pas. Je disais en introduction que le numérique a donné encore plus de poids aux méthodologies, aux modélisations et aux technologies.

Vous avez probablement lu les 34 propositions pour relancer l'industrialisation en France (François Hollande, septembre 2013), le rapport de la Commission Innovation 2030 présidée par Anne Lauvergeon (7 ambitions pour la France, octobre 2013), ou encore le plan de formation de Fleur Pellerin (début 2013) : un point commun, la Data.

Dans son rapport, Anne Lauvergeon met en avant la qualité reconnue internationalement des formations mathématiques et statistiques françaises.

Plus récemment, au plus haut niveau de l'Etat, le gouvernement a nommé le 18 septembre dernier M. Henri Verdier comme Chief Data Officer, une première en Europe. Il est en charge de l'Open Data, mission qui consiste à ouvrir les données publiques : en dresser l'inventaire, s'assurer qu'elles sont bonnes, favoriser leur circulation et surtout de développer de nouvelles méthodes d'analyse des données au service des politiques publiques.

Les objets connectés, prochaine révolution, vont continuer à engendrer des bases de données multiples, dans des domaines de plus en plus diversifiés. Il est heureux de voir que le monde scientifique et statistique aura un rôle à jouer pour l'ensemble de leurs utilisations.

Opportunité pour la formation, la recherche et le développement, et l'emploi : on estime à au moins 300 000 le nombre de postes de Data Scientists créés d'ici 2022 en Europe ; et le rapport « Les métiers en 2022 » de France Stratégie avec la DARES évalue à 292 000 les postes à pourvoir en France d'ici 2022 en ingénieurs de l'informatique, étude et recherche.

Il est à espérer que les orientations de formation de Mme Fleur Pellerin seront suivies.

Question 4 : et le cadre juridique ?

Une statistique est une quantité établie à partir de données observées. Si une base de données est un échantillon représentatif d'une population, le but du respect de la vie privée et de l'intimité est d'apprendre des choses sur la population dans sa globalité, ou sur des sous-parties, dans le respect du cadre réglementaire ou législatif existant (secret statistique, loi sur les télécoms, loi de type « informatique et libertés »), et en protégeant la vie privée des individus de l'échantillon ayant donné des informations présentes dans la base. « Privacy » et bases de données sont un champ déjà abordé au milieu des années 70, avec un article précurseur de Tore Dalenius.

Le respect de la vie privée est une priorité : genre de phrase sur laquelle tout le monde est d'accord a priori, mais nous savons tous que « le diable est dans le détail ».

D'abord, un cadre de régulation ou législatif existe dans beaucoup de pays.

Ainsi, en France, nous avons la loi (modifiée) 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

Ensuite, il existe le Code des Postes et Télécommunications électroniques.

Enfin la loi 78-17 dite « Informatique et Libertés » (modifiée) du 6 janvier 1978 fixe les règles de protection de la vie privée dans l'utilisation de fichiers contenant des DCP.

L'émergence des Big Data, l'opportunité économique et de R&D qu'ils présentent, provoquent une réflexion actuelle qui frôle le grand écart. Comment concilier respect de la vie privée et potentiel de relance économique ? Aux Etats-Unis, le PCAST – instance de conseil en sciences et technologies auprès du Président – croit fermement que « les bénéfices sont incomparablement plus grands que les risques de dommages. …. »

En ce moment, tout bouge, ou va bouger. Mais dans quel sens ?

Exemple 1

Le Conseil d'Etat a publié mi-septembre un important volume intitulé « Le numérique et les droits fondamentaux », contenant 50 propositions pour mettre le numérique au service des droits individuels et de l'intérêt général.

J'en extrais là encore une courte partie concernant les algorithmes prédictifs et les cinq propositions du Conseil :

1. Pour assurer l'effectivité de l'interdiction de fonder une décision sur la seule mise en œuvre d'un traitement automatisé, confirmer que l'intervention humaine dans la décision doit être réelle et pas seulement formelle. Indiquer dans un instrument de droit souple les critères d'appréciation du caractère effectif de l'intervention humaine.

2. Imposer aux auteurs de décisions s'appuyant sur la mise en œuvre d'algorithmes une obligation de transparence sur les données personnelles utilisées par l'algorithme et le raisonnement général suivi par celui-ci. Donner à la personne faisant l'objet de la décision la possibilité de faire valoir ses observations.

3. Dans le cadre de l'article 44 de la loi du 6 janvier 1978 et dans le respect du secret industriel, développer le contrôle des algorithmes par l'observation de leurs résultats, notamment pour détecter des discriminations illicites, en renforçant à cette fin les moyens humains dont dispose la CNIL.

4. Analyser les pratiques de différenciation des prix reposant sur l'utilisation des données personnelles, mesurer leur développement et déterminer celles qui devraient être qualifiées de pratiques commerciales illicites ou déloyales, et sanctionnées comme telles.

5. Encourager la prise en compte de la diversité culturelle dans les algorithmes de recommandation utilisés par les sites internet diffusant des contenus audiovisuels ou musicaux.

Le premier point se base sur l'article 10 de la loi du 6 janvier 1978, qui indique qu'aucune « décision produisant des effets juridiques à l'égard d'une personne ne peut être prise sur le seul fondement d'un traitement automatisé de données destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité. » Sans les mentionner explicitement, cet article vise donc les algorithmes et les modèles.

Le Conseil d'État note qu'il faut donc « éviter que des systèmes présentés comme relevant de « l'aide à la décision » soient en réalité presque toujours suivis et commandent la décision, l'intervention humaine n'étant alors qu'apparente ».

La deuxième proposition vise à permettre aux personnes subissant un effet juridique, suite à une décision basée sur un algorithme, de se défendre. Afin de « bénéficier de garanties analogues à celles d'une procédure contradictoire », la victime doit donc pouvoir accéder à certaines données et explications. Mais là encore, un tel système peut-il être mis en place simplement ? Si un renforcement de la CNIL (proposition 3) ne pourra être que positif pour surveiller les algorithmes des différents sites et services en ligne, il faudra tout de même voir comment les deux premières propositions pourront être appliquées.

Exemple 2

Toujours en septembre 2014, le Forum d'Avignon (qui, en 2013, proclamait : « la donnée personnelle culturelle est une data qui vaut de l'or ») a dévoilé une version préliminaire d'une déclaration des droits de l'Homme numérique.

ADN numérique

Les données personnelles, en particulier numériques, de tout être humain traduisent ses valeurs culturelles et sa vie privée. Elles ne peuvent être réduites à une marchandise.

Ethique et équitable

L'exploitation raisonnable des données est une opportunité pour le développement de la recherche et de l'intérêt général. Elle doit être encadrée par une charte éthique universelle protégeant la dignité, la vie privée, la création de chaque être humain et le pluralisme des opinions.

Vie privée

Tout être humain a droit au respect de sa dignité, de sa vie privée et de ses créations, et ne peut faire l'objet d'aucune discrimination fondée sur l'accès à ses données personnelles et aux usages qui en sont faits. Nulle entité, publique ou privée, ne doit utiliser des données personnelles aux fins de manipuler l'accès à l'information, la liberté d'opinion ou les procédures démocratiques.

Droit de regard

Tout être humain doit disposer d'un droit de regard, de confidentialité et de contrôle sur ses données personnelles y compris sur celles produites du fait de ses comportements et des objets connectés à sa personne. Il a droit à la protection de son anonymat quand il le souhaite.

Consentement

Toute exploitation des données comme des créations de tout être humain suppose son consentement préalable, libre, éclairé, limité dans le temps et réversible.

Transparence des usages

Les utilisateurs de données personnelles, quel que soit leur niveau de responsabilité, Etats, collectivités publiques et privées, entreprises et individus, doivent faire preuve d'une totale transparence dans la collecte et l'usage des données de tout être humain et en faciliter l'accès de chacun, la traçabilité, la confidentialité et la sécurisation.

Recherche et intérêt général

La recherche et l'innovation ouvertes, s'appuyant sur le partage consenti et anonyme des données de tout être humain, dans le respect de sa dignité et de la diversité culturelle, sont favorables à l'intérêt général.

Coopération, Société aidée par les données

La coopération de la société civile et des entreprises est nécessaire pour replacer l'être humain au cœur d'une société de confiance aidée par une utilisation raisonnable des données personnelles produites et déduites.

Exemple 3

La future loi sur la République Numérique

Exemple 4 : l'approche US

Deux grands courants de pensée semblent apparaître : le premier, d'essence plutôt européenne sinon française compte tenu du rôle international de la CNIL dans le G29, est de type réglementaire, basé sur des interdits juridiques ; le second, nord-américain, est plus ouvert.

Je résume ce dernier en paraphrasant un extrait d'un rapport J. P. Holdren – E. S. Lander « Big Data and Privacy : A Technological Perspective », remis en mai 2014 au Président Obama, par le Conseil sur la Science et la Technologie auprès du Président (PCAST) :

« …. Quoiqu'il en soit, il y a des bénéfices et des risques :

Santé : des « plus » évidents, mais des infos pour les sociétés d'assurance

Déplacements par GPS : optimiser la fluidité de trafic, mais localisation de l'individu »

Exemple 5 : vers une nouvelle définition ?

Tore Dalenius énonce, en 1977, dans le contexte des bases de données telles qu'elles existent alors des principes touchant à l'éthique, au sens du respect de l'intimité et de la vie privée. Dans son article publié dans Statistik Tidskrift (1977), « Towards a methodology for statistical disclosure control », Dalenius pose le principe suivant :

Accéder à une base de données ne doit pas permettre d'apprendre plus de choses sur un individu que ce qui pourrait être appris sans accès à cette base de données.

Par parenthèse, il est à noter que ce principe sera à la base des travaux de Mme Shafi Goldwasser et Silvio Micali, professeurs au MIT, sur la notion de sécurité sémantique (pré-print en 1982, puis article dans Journal of Computer and Systems Sciences, 1984). Ils créent le premier système à chiffrement probabiliste dont on prouve la sécurité absolue. Les auteurs ont reçu le Turing Award décerné par l'ACM (Association for Computing Machinery).

Le principe de Dalenius semble simple et cohérent. Malheureusement, on peut démontrer qu'il ne peut être atteint.

La raison en est l'existence d'information auxiliaire, c'est-à-dire l'information disponible – hors de la base de données initiale – pour toute personne voulant trouver une faille et accéder à des DCP.

La démonstration est longue, mais on peut donner un exemple de l'impossibilité de ce principe.

Considérons une base de données contenant les tailles d'échantillons d'hommes adultes de K nationalités. Supposons que l'on dispose de l'information auxiliaire : « M. A mesure deux centimètres de moins que l'homme italien adulte moyen ». A partir de la base de données on déduit précisément sa taille. Et on remarque que, inversement, l'accès à la seule information additionnelle fournit relativement peu d'éléments sur M. A.

En outre, le résultat d'impossibilité s'applique que A contribue à la base de données ou pas.

Plus formellement, le Théorème d'impossibilité du principe de Dalenius peut s'écrire ainsi :

Soit S un mécanisme de protection d'une base de données, et une faille F dans cette protection.

Pour toute base de données, il existe toujours une information auxiliaire Z telle que :

a) Z seule n'a pas d'utilité pour un « hacker »

b) Z combinée avec les données protégées permet de trouver dans S une faille F avec une probabilité qui tend vers 1

Ceci donne naissance à une autre forme de « privacy » :

Un risque quelconque pour l'individu i (par exemple le fait que i se voie refuser une assurance) ne peut augmenter significativement en raison de la participation de i à une base de données.

Cette approche est appelée « differential privacy » (intimité différentielle), et repose sur des fondements probabilistes et statistiques. Je crois qu'elle va être amenée à se développer rapidement.

Pour conclure, on parle beaucoup de data scientists, personnes mêlant intérêt pour la mathématique statistique, les probabilités, les sondages ; demain, ne faudra-t-il pas hybrider les scientifiques avec les juristes ? Probablement, quand on voit se développer sur les mêmes sujets des revues de droit et des revues scientifiques.

Bibliographie

[1] Adam N., Wortmann J. (1989), Security-Control methods for statistical databases : a comparative study, *ACM Computing Surveys.*

[2] Big Data : Seizing Opportunities, Preserving Value, Executive Office of the President, mai 2014.

[3] Big Data and Privacy : A Technological Perspective , Report to the President, PCAST, mai 2014.

[4] Conseil d'Etat, Etude annuelle 2014, Le numérique et les droits fondamentaux, La Documentation Française, 2014.

[5] Dalenius T. (1977), Towards a methodology for statistical disclosure control, *Statistik Tidskrift*, Vol. 15.

[6] Domingo-Ferrer J., Torra V. (Eds) (2004), Privacy in Statistical Databases, *Proceedings CASC Project Final Conference*, Barcelone.

[7] Dwork C., Nissim K. (2004), Privacy-preserving datamining on vertically partitioned databases, in *Advances in Cryptology, Proceedings of Crypto.*

[8] Dwork C., McSherry F., Nissim K, Smith A. (2006), Calibrating noise to sensitivity in private data analysis, in *Proceedings of the 3rd Cryptography Conference*.

[9] Goldwasser S., Micali S. (1984), Probabilistic Encryption, in *Journal of Computer and Systems Sciences*, n°28.

[10] Journée "Big Data, Big Analytics", Séminaire IREP, 30 mai 2013.

[11] Journée "Concilier les enjeux business avec la protection des données", Séminaire IREP, 20 mai 2013.

[12] Provost F., Fawcett T. (2013), *Data Science for Business*, O'Reilly Ed.

[13] Santini G. (2009), Massive Modelling : a new media research challenge, *WWRS*