Données individuelles et Société connectée : le thème de la confiance

Philippe Tassi (1972), Directeur général Adjoint de Médiamétrie, Vice président d'ENSAE Alumni

La révolution récente due à la convergence numérique, et demain celle des objets connectés, a permis de mettre sous forme homogène des informations que l'histoire considérait comme de nature différente : données numériques, textes, son, images fixes, images mobiles. Ceci favorise l'émergence des big data, bases de données d'importante volumétrie. Comme toute innovation, les données massives posent de nouvelles questions aux spécialistes de multiples domaines.
Une thématique dominante est celle de la confiance. Comment conforter ou accroître la confiance des individus dans une société de la donnée, individus souvent à l'origine, consciente ou non, de ces mêmes données.
Cette confiance procède de plusieurs approches : la première est l'utilité de ces données nombreuses et le progrès qu'elles permettent ; la deuxième repose le respect de la vie privée, par le cadre législatif, réglementaire ou déontologique ou par les techniques scientifiques de défense de l'espace de l'intimité.
C'est cette vision transversale des big data qui est à l'origine du colloque « Individu, données et société connectée : opportunités, risques et confiance » organisé par ENSAE Alumni le 18 juin prochain.

Des données...

Des premières tablettes cunéiformes du pays de Sumer à la fin du XIXème siècle, le monde a vécu sous le règne quasi-exclusif de « l'exhaustivité ». En dépit, au XVIIème siècle, de tentatives de l'école de l'arithmétique politique de John Graunt et William Petty en Angleterre, et des avancées de Sébastien Le Prestre de Vauban et Jean-Baptiste Colbert en France, la norme reste la vision exhaustive : les pays développés sur le plan statistique gèrent des recensements, qu'ils soient de population, d'entreprises ou d'établissements industriels, ou de terres agricoles.

Le XXème siècle est marqué par le recul de l'exhaustif et la montée de l'échantillonnage, dont la naissance est datée par la communication « Observations et expériences concernant des dénombrements représentatifs » d'Anders N. Kiaer, directeur du Bureau Central de Statistique de Norvège, au Congrès de Berne de l'Institut International de Statistique d'août 1895.

En 1925 l'IIS confirme l'intérêt scientifique de l'approche de Kiaer. En 1934 paraît l'article de

référence de la théorie des sondages de Jerzy Neyman. Les applications opérationnelles sont immédiates : dès 1935, premiers panels de consommateurs et de distributeurs, premières études d'opinion dont le père est G. Gallup, qui crée aux Etats-Unis l'*American Institute for Public Opinion*. Après-guerre, l'échantillon devient la référence : rapidité d'exploitation, réduction des coûts, dans un contexte de progrès des théories des probabilités et statistique, de l'informatique, avec en outre une généralisation des domaines d'application : statistique publique, économie, démographie, santé, marketing, sociologie, audience, politique, etc.

... aux big data

Depuis la fin du XXème siècle, la convergence numérique a favorisé le recueil automatique de données observées sur des populations de plus en plus grandes, créant des bases de données contenant une masse croissante d'informations, et annonçant le potentiel retour du paradigme de l'exhaustif.

En outre, le passage au digital a permis de mettre sous la même forme des informations historiquement distinctes et hétérogènes comme des fichiers de données quantitatives, de textes, de sons (l'audio), d'images fixes ou d'images mobiles (la vidéo), ce qui est une révolution aussi fondamentale que si tous les êtres humains se mettaient à parler la même langue.

Sont ainsi apparues ce qu'on appelle les big data, ou données massives, ou données nombreuses, ou encore données à grande dimension pour les mathématiciens, et pour les puristes les méga-données puisque c'est le nom que recommande depuis août 2014 la Commission Nationale de Terminologie.

Les big data possèdent deux paramètres majeurs qui aident à définir leur volumétrie : quantité et fréquence d'acquisition. On peut écrire « Volumétrie = Quantité x Fréquence », la quantité pouvant aller jusqu'à l'exhaustivité, et la fréquence pouvant aller jusqu'au temps réel.

Ces big data sont une réalité : en 2014, par minute, le monde a créé 1,7 million de milliards d'octets de données, c'est-à-dire l'équivalent du contenu de 360 000 DVD. Et ce n'est qu'un début, puisque tout annonce la prochaine évolution : l'internet des objets, données engendrées par des capteurs insérés dans des objets du quotidien.

Les dimensions de la confiance

Comme toute innovation, les données massives soulèvent de nouvelles questions aux spécialistes de multiples domaines, en termes d'avantages mais aussi de points de surveillance.

Une thématique dominante, bien que non unique, est celle de la confiance : comment

maintenir, conforter ou accroître la confiance des individus dans une société de la donnée, individus souvent à l'origine, consciente ou non, de ces mêmes données.

Cette confiance procède de l'utilité de ces données nombreuses et du progrès qu'elles permettent, mais aussi du respect de la vie privée, par le cadre législatif, réglementaire ou déontologique ou par les techniques scientifiques de défense de la sphère intime.

Confiance en l'utilité sociale

Sans cesse apparaissent de nouvelles utilisations de ces données nombreuses dans des secteurs divers touchant l'individu dans son quotidien : médecine, épidémiologie et santé, assurances, sport, marketing, culture, ressources humaines…

Certes, le grand public ou les médias peuvent considérer que la ligne blanche est parfois dépassée dans certaines approches de ciblage marketing, mais à côté de ces errements, combien de progrès potentiels ? A-t-on supprimé le système bancaire à cause de banquiers parfois véreux, de produits financiers toxiques ou de l'existence de hold-up ?

La note intitulée « Analyse des Big Data. Quels usages, quels défis ? », publiée en novembre 2013 par France Stratégie, fournit une bonne synthèse des « plus » attendus par la société.

Les avantages sont tels qu'au niveau politique, globalement, les orientations sont clairement annoncées.

Pour ne parler que de la France, les data sont le point commun des 34 propositions pour relancer l'industrialisation en France (François Hollande, septembre 2013), du rapport de la Commission Innovation 2030 présidée par Anne Lauvergeon (« 7 ambitions pour la France », octobre 2013), de la feuille de route gouvernementale sur le numérique (février 2013), ou bien plus récemment du schéma « Stratégie Nationale de Recherche », publié le 5 mars 2015.

Ce dernier définit des orientations et des programmes d'actions. Parmi les orientations : objets connectés, exploitation des grandes masses de données, collaboration homme-machine, disponibilité des données et extraction de connaissances, résilience des systèmes de sécurité… Et le premier des cinq programmes d'action s'intitule : big data.

Notons que dans son rapport, Anne Lauvergeon met particulièrement en avant la qualité reconnue internationalement des formations mathématiques et statistiques françaises.

Mentionnons également le rapport « La nouvelle grammaire du succès, la transformation numérique de l'économie française », issu d'une mission confiée à Philippe Lemoine, remis au

gouvernement en novembre 2014 et, en avril 2015, la publication par l'Institut Montaigne du rapport « Big Data et Objets connectés : faire de la France un champion de la révolution numérique ».

Au plus haut niveau de l'Etat, le gouvernement a nommé le 18 septembre 2014 Henri Verdier comme *chief data officer*, c'est-à-dire administrateur général des données, une première en Europe. Il est en charge de l'*open data*.

Enfin, un dernier avantage : les données massives et leurs applications sont un espoir pour l'économie et l'emploi : on estime à au moins 300 000 le nombre de postes de scientifiques de la donnée créés d'ici à 2022 en Europe.

Ce qui précède ne se réduit pas à la France, comme en témoigne le rapport « *Big Data : Seizing Opportunities, Preserving Value* » remis au Président Obama en mai 2014.

Confiance en la confidentialité

Les données ou statistiques, détenues ou élaborées par des administrations ou des entreprises, ont en général été construites à partir d'informations individuelles. Assurer au mieux la protection des sources, c'est-à-dire de la vie privée, est une question sociétale. Comment établir et maintenir la confiance du grand public, partie prenante numéro un, et respecter l'équilibre entre promesse de confidentialité et utilisation des données recueillies ?

Pour y répondre, deux voies : l'une est réglementaire, et montre que les Etats ont pris conscience depuis longtemps de la nécessité d'établir des garde-fous juridiques ; l'autre vise à s'appuyer sur la technologie en mettant en place des obstacles techniques empêchant la diffusion de données contre le gré de leur auteur.

Le cadre réglementaire et législatif de la vie privée

La première loi fut celle du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. Elle définit le « secret statistique », et un concept appelé « impossibilité d'identification » dans le cadre des dispositifs de recueil de la statistique publique (recensements, enquêtes).

La loi la plus connue est la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, entrée dans le langage commun sous le nom de « Loi Informatique et Libertés ». Elle précise les règles de protection de la vie privée. Cette loi, remarquablement écrite et qui a servi de modèle à de nombreux pays, définit les données à caractère personnel, une catégorie particulière étant les données sensibles, et la notion de « traitement ».

Ceci est important. En effet, si l'inférence « fisherienne » de la statistique classique va de l'échantillon à la population (apprendre sur un tout en observant une partie), les bases de données et les données à caractère personnel (DCP) peuvent engendrer une inférence horizontale, de la base de données vers l'individu lui-même, puisque les modèles, les algorithmes, bref le *data mining*, peuvent permettre d'accéder à des éléments personnels estimés.

A ce cadre s'ajoute le Code des Postes et Télécommunications électroniques (loi du 23 octobre 1984, modifiée), qui aborde le traitement des DCP pour la fourniture au public de services de communications électroniques, notamment via les réseaux qui prennent en charge les dispositifs de collecte de données et d'identification.

A côté de ces lois, il y a des codes de déontologie professionnels, dont, par exemple, celui des administrateurs de l'Insee ou celui d'ESOMAR (*European Society for Opinion and Market Research*). Né en 1948, régulièrement mis à jour, il précise dès son origine les « bonnes pratiques » pour les études de marché et d'opinion, son objectif étant d'établir la confiance de la partie prenante majeure qu'est le public envers ce type d'étude.

De nombreuses réflexions sont en cours sur données, vie privée et traitements.

Le Conseil d'Etat a publié mi-septembre 2014 un livre intitulé « Le numérique et les droits fondamentaux », contenant 50 propositions pour mettre le numérique au service des droits individuels et de l'intérêt général. En septembre 2014, le Forum d'Avignon – groupe transversal réunissant des personnalités des univers de la culture, des médias, de l'économie, et du numérique – a dévoilé une une première version d'une Déclaration des huit Droits de l'Homme Numérique.

Le Groupe de l'Article 29, réunissant les entités ad hoc européennes, a officialisé une « Déclaration commune des autorités européennes de protection des données » le 25 novembre 2014. Le 8 décembre 2014, un Forum de

l'UNESCO a été consacré à « *The European Data Governance* ».

Et pour conclure, la future loi sur la République numérique, prévue courant 2015, à laquelle le public pouvait contribuer sur quatre thèmes : croissance, innovation et disruption ; la loyauté dans l'environnement numérique ; la transformation numérique de l'action publique ; et enfin la société face à la métamorphose numérique.

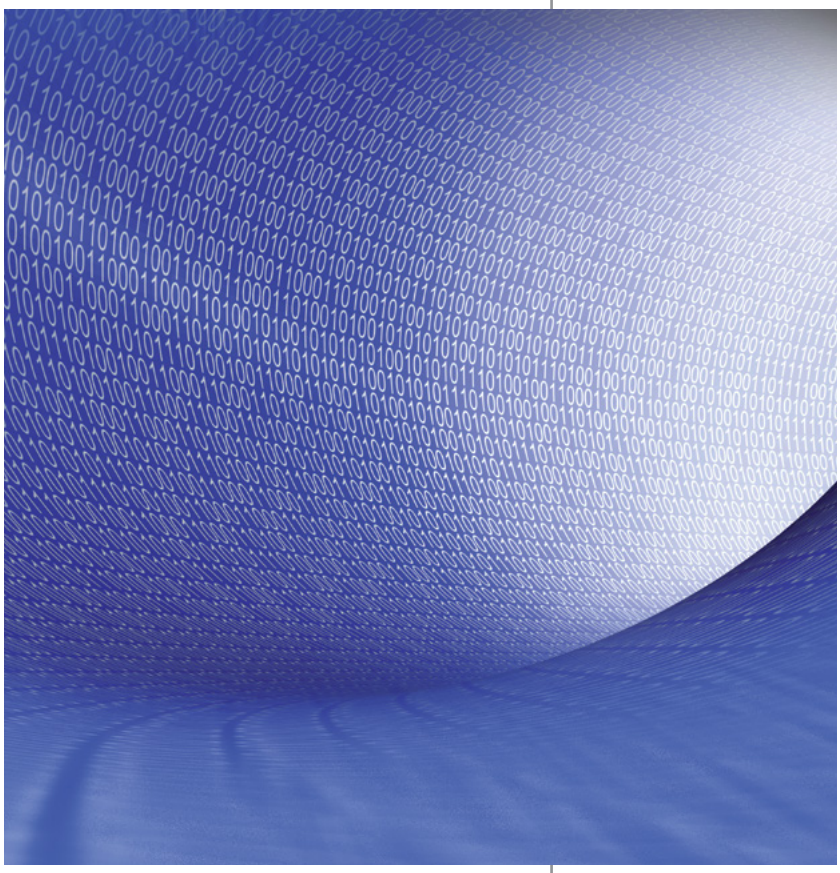
La technique au service de la confiance

Les rapports entre la vie privée et les bases de données sont un champ de recherche abordé depuis les années 1970. Comment assurer le respect de la vie privée sur un plan technique, en plus du cadre réglementaire existant ? Comment concilier promesse de confidentialité et utilisation des données recueillies ?

Les méthodes de cryptage – la cryptographie, ou science du chiffrement, est l'apanage des gendarmes là où la cryptanalyse, science du déchiffrement est le terrain de jeu des voleurs – ont bien évolué depuis leurs origines il y a plus de trois millénaires. Jules César cryptait les messages qu'il envoyait à ses généraux, le « Grand Chiffre » du Cabinet Noir de Louis XIV acquiert au XVIIème siècle une célébrité mondiale.

Ces méthodes permettent de rendre illisible, c'est-à-dire incompréhensible, un document – au sens large – à tout acteur ne possédant pas la clé de cryptage. Les travaux les plus élaborés, somme toute récents, sont dus à Madame Shafi Goldwasser et Silvio Micali, professeurs au MIT, publiés en 1984 dans *Journal of Computer and Systems Sciences*, et ont valu à leurs auteurs le bien nommé *Turing Award* décerné par l'*Association for Computing Machinery*.

Ces protocoles de cryptage sont renforcés par diverses techniques. L'obscurcissement des données consiste à altérer les données, par exemple en les plongeant dans des espaces de dimension plus élevée, principe de dilution, soit directement en les transformant. Dans la première voie, on peut créer de nouvelles variables additionnelles augmentant la dimension initiale du vecteur recueilli et créer du « brouillard » masquant. Dans la deuxième voie, on distingue les techniques non perturbatrices (masquer la valeur de cellules dans un tableau de résultats, enlever des variables pour certains individus, partager seulement un échantillon de données extrait de la



base générale, combiner des catégories pour des variables à modalités) et des méthodes directement interventionnistes sur les données qui vont engendrer du bruit, au sens large. Ainsi, il est possible de modifier des variables numériques en les arrondissant, ou en les tronquant, ou, plus généralement en les transformant par homomorphisme.

Une autre piste d'investigation porte le nom de données synthétiques (« *synthetic data* »), née des travaux sur les données manquantes de J. Little et D. Rubin.

Une approche innovante est la confidentialité différentielle, apparue au milieu des années 2000 avec les travaux de Cynthia Dwork, chercheuse de Microsoft. Sa philosophie est la suivante « la probabilité d'une conséquence négative quelconque pour l'individu i (par exemple le fait que i se voie refuser un crédit, ou une assurance) ne peut augmenter significativement en raison de la participation de i à une base de données. »

Ceci repose sur des fondements probabilistes et statistiques, puisqu'on admet l'existence du risque de faille dans la confidentialité mais on cherche à en minimiser la probabilité de survenance. ■