SOCIETE FRANÇAISE DE STATISTIQUE

Les cafés de la statistique

"La statistique éclaire-t-elle les questions de société" ?

Soirée du 27 mars 2017 – Centième Café de la statistique

Big Data : Big Science ? Big Brother ?

Synthèse des débats [*]

L'exploitation des données massives ouvre-t-elle de nouvelles portes pour la compréhension des grands problèmes de société et pour l'action ? Santé, environnement, villes, agriculture… quel rôle peuvent jouer les données massives ? Avec quelles limites ?

Invitée :
Valérie Peugeot
Chercheuse en « Digital studies » à Orange Labs, Commissaire à la Cnil en charge des données de santé, Présidente de l'association Vecam-Citoyenneté dans la société numérique.

Exposé introductif :

L'invitée se déclare heureuse de participer à ce Café de la statistique. Elle se présente comme chercheuse dans un laboratoire de sciences sociales et humaines fonctionnant - ce qui n'est pas si fréquent - au sein d'une entreprise, en l'occurrence Orange. L'association qu'elle préside mène depuis plus de vingt ans une réflexion sur les enjeux des technologies de l'information et, depuis dix ans, sur les biens communs de la connaissance. Toutes les personnes qui y contribuent sont des bénévoles.

---

[*] Tant l'exposé liminaire que le contenu des échanges sont structurés en quelques thèmes, sans suivre l'ordre chronologique. Par ailleurs, l'identité des intervenants n'était pas toujours connue et l'on a choisi de ne pas attribuer nominativement les propos. Au reste, ceux-ci ont été reconstitués à partir des notes du secrétariat sans reprendre leur formulation détaillée. Pour retracer le débat, les thèmes sont souvent introduits sous forme d'une question : ce qui vient ensuite n'est pas la seule réponse de l'invitée, mais l'ensemble des contributions des participants.

Depuis 2010, le Big Data[1] occupe une place croissante dans les débats publics et l'espace médiatique, le plus souvent associée à un vocabulaire inflationniste : Data déluge, Nouvel or noir, Pétrole du 21e siècle ! Il faut mettre ces métaphores à distance. La sociologie montre que la plupart des innovations s'accompagne d'un discours sur les promesses, destiné à attirer les financeurs, les médias, les chercheurs, etc. Il faut garder la tête froide et se demander jusqu'à quel point le déploiement des mégadonnées est porteur de véritables ruptures. L'invitée refuse pour sa part le déterminisme technologique. L'introduction d'une technologie dans la société n'est pas neutre, ne serait-ce que parce que la technologie est porteuse d'une certaine vision de ses concepteurs. Pour autant, c'est l'aller-retour avec les usages, les appropriations, qui dessinera les contours des transformations sociétales effectives. Il n'y a en la matière ni prédestination, ni neutralité.
L'initiale «V » signe les propriétés prêtées au monde des données massives : Vélocité, Variété, Volume, mais aussi Véracité, Valeur, Visibilité… Pourtant, il n'y a rien de fondamentalement nouveau dans les pratiques sociales en matière de collecte et de manipulation de grands ensembles de données. En Chine, la dynastie des Han, commencée 206 ans avant Jésus-Christ, a su recenser plus de 58 millions d'habitants en l'an 2 ! Les nouveautés résident en premier lieu dans ce que l'on peut appeler « la mise en données du monde ». La plupart de nos actions, nos déplacements, nos communications, nos traitements médicaux, nos démarches administratives, génèrent aujourd'hui en se dématérialisant une masse de données inédite. A cela s'ajoutent les données produites par l'Internet des objets connectés qui produiront des informations sur nos agissements, que nous en soyons conscients ou pas. D'autres ruptures tiennent à la croissance des capacités de stockage liée à la baisse des coûts et à la création de nouveaux outils d'analyse, ces derniers permettant entre autres des croisements inédits de données, notamment de données non structurées.

Certains ont pu affirmer il y a quelques années que le Big Data signait la fin de la science reposant sur des hypothèses et des théories, les données livrant en quelque sorte d'elles-mêmes les clés de la compréhension des phénomènes. Ce discours a depuis largement été battu en brèche. Pour autant, l'analyse de données massives participe en profondeur à un changement du rapport à la connaissance. Elle est aujourd'hui convoquée dans la plupart des champs de la recherche, qu'il s'agisse d'astronomie, de sismologie, d'océanographie, d'écologie, de recherche médicale, etc.

Dans ce dernier domaine, l'accès aux données est modifié et les croisements se multiplient. Le SNDS, système national de données de santé[2], va élargir l'accès à ses données à de nouveaux acteurs, publics et privés. Les établissements hospitaliers regroupent leurs bases de données, à l'instar de l'APHP[3] qui a réuni les informations produites par ses 39 établissements. Grâce à ces bases massives, des cohortes de malades peuvent être rapidement constituées et étudiées, de nouvelles thérapeutiques plus commodément entreprises, les temps d'expérimentation raccourcis, etc. En outre, le croisement avec des données environnementales des patients est facilité : on peut prendre en compte, par exemple, l'exposition de ces derniers aux pesticides pour comprendre s'il y a un lien de causalité, ce qui ouvre la voie à de nouvelles politiques publiques en matière de santé, dans des logiques de prévention. Les séquençages génétiques massifs peuvent s'accompagner de croisements avec des données cliniques, d'imagerie médicale, ou encore de déterminants sociaux-économiques, et déboucher sur une prometteuse médecine individualisée.

---

[1] NDR On a choisi, dans ce compte rendu, de conserver des locutions anglaises ou américaines mais en fournissant en principe toutes les traductions. Ainsi, Big Data est traduit dans certaines phrases par « données massives » ou par « mégadonnées ». Par ailleurs, quand on a opté pour le masculin (**le** Big Data), c'est pour désigner, soit l'ensemble constitué par les données et les traitements qui leur sont appliqués, soit le phénomène de société qui se développe autour de cet ensemble.

[2] Instauré par l'article 193 de la loi du 26 janvier 2016 de modernisation de notre système de santé, le SNDS sera géré par la Caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS). Il permettra de chaîner les données de l'assurance maladie (base Sniiram : voir note 14), les données des hôpitaux, les causes médicales de décès, les données relatives au handicap et un échantillon de données en provenance des organismes complémentaires.

[3] Assistance publique – Hôpitaux de Paris, centre hospitalier universitaire à dimension européenne regroupant 39 hôpitaux.

Dans les sciences humaines et sociales, le sociologue et le géographe bénéficient de nouvelles sources de données, comme les données de téléphonie. Ils utilisent des mégadonnées pour mieux comprendre les mouvements de population, par exemple en cartographiant les mouvements des taxis dans les villes ou les mouvements de populations liés à un événement traumatique (cyclone ou autre). La recherche en littérature est tout aussi concernée, alors que des corpus patrimoniaux de millions de documents sont numérisés, à l'image du fonds de 53 millions de documents de la bibliothèque Europeana. Ces sources multiples permettent d'étudier par exemple la diffusion des mouvements culturels dans une société, le contexte social et historique d'émergence d'une œuvre ou d'un mouvement littéraire.

Ces multiples usages participent d'une rupture épistémologique, entendue ici comme la transformation substantielle de la façon dont l'humanité produit de la connaissance. Cette transformation croise d'autres bouleversements de la science :
- la résilience des anciennes méthodes scientifiques, qui se réinventent avec l'utilisation des données massives ;
- l'irruption citoyenne dans une science participative : voir Open Street Map[4], service gratuit et contributif de cartographie en ligne, équivalent ouvert du Google Map ;
- la manière dont la science circule : alors qu'historiquement les résultats scientifiques circulaient, notamment par le biais des sociétés savantes et de leurs publications, au cours du 20e siècle les savoirs scientifiques ont été victimes de formes d'enclosure liées à la systématisation des brevets et aux politiques tarifaires des publications scientifiques oligopolistiques. De récentes évolutions législatives, notamment en France avec la loi pour une République numérique, encouragent l' « Open Access » et le « Text and Data Mining[5] » scientifique (la fouille de données massive sur de grands corpus scientifiques).

Mais, plus que la recherche, ce qui, dans le monde du Big Data, attire le plus les regards et les capitaux, c'est la transformation de l'économie et des marchés. Dans l'organisation des entreprises, la relation-client, l'implantation de nouveaux établissements, la gestion des ressources humaines (l'usage de ce qu'on appelle les « Data RH » pour la mobilité des agents ou les recrutements) comme dans le marketing personnalisé voire prédictif, l'exploitation des mégadonnées débouche sur de nouvelles pratiques. Ce développement est à la fois rapide et en partie opaque. Si certains acteurs, notamment les fournisseurs de solutions de stockage et d'analyse, sont bien connus, d'autres le sont moins, à l'image des « Data Brokers », les courtiers de données qui déposent des « cookies » dans nos terminaux pour enregistrer nos actions en ligne et ainsi nous profiler sans que nous en ayons une conscience claire. Les pouvoirs publics commencent à prêter attention à cette activité. Le marché publicitaire en ligne s'accompagne de systèmes d'enchères qui se déroulent à des vitesses extrêmes sans qu'on en ait conscience et sans qu'on ait, en tant qu'utilisateurs, la moindre prise sur eux.

---

[4] Wikipédia : OpenStreetMap (OSM) est un projet qui a pour but de constituer une base libre de données géographiques du monde (permettant par exemple de créer des cartes sous licence libre), en utilisant le système GPS et d'autres données libres. Il a été initié en 2004 au University College de Londres. L'utilisation de moyens informatiques reposant sur Internet permet l'intervention et la collaboration de tout utilisateur volontaire.

[5] Wikipédia : L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, *data mining*, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.
Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances.
L'utilisation industrielle ou opérationnelle de ce savoir dans le monde professionnel permet de résoudre des problèmes très divers, allant de la gestion de la relation client à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation de sites web. C'est aussi le mode de travail du journalisme de données.

Le monde politique est touché lui aussi par le Big Data : lors de sa première campagne électorale, Barack Obama a utilisé le logiciel Catalyst pour prendre contact avec 220 millions d'Américains et collecter environ 600 informations concernant chaque citoyen. De multiples logiciels (Nation Builder, Digital Box, etc.) sont désormais utilisés par les candidats à des fonctions électives, toutes tendances politiques confondues, qu'il s'agisse de Donald Trump ou de Jean-Luc Mélenchon. Ces logiciels permettent – entre autres - de mieux cibler les démarches de porte-à-porte. Ils visent à optimiser la connaissance des personnes mais ouvrent de sérieux risque d'excès.

La ville connectée constitue un autre secteur de mobilisation des données massives. Avec les « Smart Data[6] », l'objectif consiste à rendre la gestion de la ville plus efficace, par exemple par une analyse des mobilités dans le but de rendre celles-ci moins énergivores. Ce qui n'est pas sans poser des problèmes de gouvernance et de modèle de société pour la ville de demain, mais c'est un autre débat.

Bien d'autres utilisations des données massives existent. L'agriculteur au volant de son tracteur équipé de capteurs analyse la composition chimique des sols qu'il travaille. Le « Data Journalism »[7], dont les « Panama Papers[8] » constituent l'illustration la plus magistrale, ou la lutte contre la fraude fiscale pour profiler les particuliers ou les entreprises les plus à même de se livrer à cette activité répréhensible, constituent deux exemples parmi bien d'autres.

Ces multiples utilisations entraînent de nombreuses controverses :
- on s'interroge en premier lieu sur la qualité des données. On a pu croire un temps que l'Open Data[9] était une entreprise aisée et qu'il suffisait d'ouvrir l'accès aux fichiers pour que d'autres puissent s'en servir. Mais la plupart des bases de données ne sont pas prêtes à être utilisées pour d'autres usages que ceux pour lesquels elles ont été conçues. Il y a un véritable travail de la donnée pour que celle-ci soit réutilisable, un travail qui a un coût ;

- autre problème, celui de l'empreinte énergétique et écologique des données. La courbe d'évolution des consommations énergétiques associées aux données (notamment dans les réseaux et les « *Data Centers*) est de loin supérieure à celle de l'amélioration de l'impact énergétique résultant de l'application du Big Data à divers domaines énergivores, et cela malgré les nombreuses recherches menées à des fins d'économie. On peut citer aussi l'usage croissant et le mauvais recyclage par l'industrie du numérique des métaux que l'on appelle « les terres rares ». Peut-on imaginer dans certains secteurs une forme de frugalité dans la production et l'usage de données ? La question n'est jamais posée ;

---

[6] Smart Data ou « Données intelligentes » ; c'est la pratique consistant à extraire les informations les plus pertinentes d'un très grand ensemble de données.

[7] Le journalisme de données (*Data Journalism* en anglais), ou journalisme de bases de données (*Database Journalism*), repose sur l'exploitation de données (statistiques ou autres) et la mise à la disposition de celles-ci au public. Il est lié à la libre disponibilité des données : de plus en plus de données sont diffusées par les institutions et les gouvernements, et un journaliste d'investigation sachant les analyser peut mettre en lumière des faits importants. La présentation graphique des données est une composante de ce journalisme.
Le journalisme de données a fait l'objet d'un Café de la statistique le 14 avril 2015. L'invité était Alexandre Léchenet, journaliste-web à Libération. Le compte rendu de ce Café est à l'adresse suivante :
http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=2147

[8] Wikipedia : Les Panama Papers désignent la fuite de plus de 11,5 millions de documents confidentiels issus du cabinet d'avocats panaméen Mossack Fonseca, détaillant des informations sur plus de 214 000 sociétés offshore ainsi que les noms des actionnaires de ces sociétés. Les documents fournis par un lanceur d'alerte anonyme et non rémunéré représentaient un total de 2,6 téraoctets de données. Celles-ci ont été partagées avec les rédactions de médias dans plus de 80 pays par l'intermédiaire de l'International Consortium of Investigative Journalists (ICIJ).

[9] Possibilité d'accéder à et de réutiliser gratuitement des données détenues par les institutions publiques ou par des entités privées.

- les usages des données à des fins d'hyper individualisation soulèvent d'autres problèmes. Cela peut être positif, comme on l'a vu pour la médecine, mais l'est beaucoup moins lorsqu'on touche aux questions d'"assurance. A terme, notamment avec l'Internet des objets, les assureurs peuvent disposer de données extrêmement précises sur nos comportements ou notre état de santé[10]. Ce qui les conduit à moduler leurs offres tarifaires, une pratique déjà en place avec les voitures connectées. C'est potentiellement grave pour plusieurs raisons : d'une part cela place des acteurs privés en position de décréter des normes sociales - à partir de quel moment une personne peut être considérée comme malade par exemple – au lieu que ce soit le fruit d'un consensus général ; par ailleurs, cela porte atteinte au principe de solidarité qui sous-tend nos sociétés occidentales et repose sur un risque partagé et traditionnellement assumé par l'assureur ;

- l'usage des données massives peut par ailleurs conduire à des formes de discrimination, liées entre autres à la manière dont sont construites et recueillies les données. Contrairement à une idée répandue, les données, même brutes, ne sont pas neutres mais sont le fruit de choix. Par exemple, une application installée sur smartphone pour détecter automatiquement l'état des rues et informer les autorités de l'existence de nids de poule va favoriser la remise en état des quartiers où passent les personnes équipées de smartphone, au détriment des quartiers sous équipés, instaurant *de facto* une forme de segmentation sociale ;

- les pratiques de profilage à des fins de marketing peuvent aller jusqu'à recueillir des informations très intimes (orientations sexuelles, politiques, religieuses…). Au-delà de ces risques d'atteinte à notre vie privée, notamment lorsque ces systèmes sont mal sécurisés comme des scandales réguliers nous le montrent, ce profilage tend à nous proposer des contenus et des publicités qui nous enferment dans un univers supposé nous correspondre. C'est ce qu'on appelle des « bulles de filtres », pour reprendre l'expression du militant américain Eli Pariser : un univers clos finit par nous enfermer sur la Toile, dont la sérendipité s'effrite du même coup, avec la perte de diversité qui en résulte ;

- dernière grande problématique mais qui n'est pas la moindre, les infrastructures qui collectent massivement des données, comme les informations issues de réseaux sociaux, rendent techniquement et économiquement possible la surveillance de masse par les pouvoirs publics : il leur « suffit » de se connecter à ces infrastructures pour y chercher des informations sur Monsieur et Madame tout le monde. L'affaire Snowden[11] a été révélatrice à cet égard. Au nom des peurs, on multiplie dans de nombreux pays les lois permettant à l'État d'exploiter les bases privées. C'est là le plus gros danger mais il existe sans doute des réponses possibles.

Débat :

Quoi de nouveau sous le soleil ?

Pourquoi le Big Data, au sens des technologies mises en œuvre, suscite-t-il un tel engouement, demande un participant ? Est-ce lié à la baisse des coûts de stockage ? Aux yeux de l'invitée, l'engouement pour le Big Data tient d'une part aux bénéfices collectifs qu'on peut en attendre et d'autre part au fait qu'il existe un secteur économique qui pousse à la roue. Les politiques ne sont pas en reste (tels par exemple Viviane Reding[12]), qui sont à la recherche d'un nouveau segment de

[10] Demain, mon assureur ayant capté mon activité physique au travers d'objets connectés pourrait m'imposer de fait une surprime pour excessive sédentarité.

[11] Du nom de Edward Joseph Snowden, informaticien américain, ancien employé de la Central Intelligence Agency (CIA) et de la National Security Agency (NSA), qui a révélé en 2013 les détails de plusieurs programmes américains et britanniques de surveillance de masse.

[12] Wikipédia : Viviane Reding, journaliste et femme politique luxembourgeoise, a été nommée en 1999 à la Commission européenne, où elle était chargée de l'éducation, la culture, la jeunesse, les médias et les sports. En 2004, dans la

croissance. Au demeurant, les prédictions sont parfois auto réalisantes... Sur le plan technique, il y a en effet baisse des coûts de stockage et surtout mise en œuvre de technologies puissantes et accessibles dont Hadoop constitue un exemple[13].
Au titre des bénéfices collectifs, une personne souligne l'intérêt du Big Data pour la santé publique en France et cite deux exemples : le Sniiram[14], qui permet davantage d'enquêtes et une surveillance très améliorée de la santé de la population, et l'application « Au secours » qui enregistre tous les passages aux urgences depuis la canicule de 2003. Il faut donc distinguer les objectifs commerciaux des autres. Entièrement d'accord sur l'intérêt de l'exploitation des mégadonnées pour la santé publique, l'invitée rappelle que le Sniiram va faire place à un système national sur les données de santé (SNDS) fort utile pour l'analyse et l'amélioration de la santé de la population ; mais elle observe que des acteurs comme IBM ou Google ont une telle puissance de calcul qu'ils risquent de capter la valeur correspondante. Des questions analogues se posent pour les données pharmaceutiques.

Les ruptures épistémologiques n'apparaissent pas évidentes à certains participants. L'invitée les conforte dans l'idée que le Big Data ne fait pas disparaître le besoin d'élaborer des hypothèses et qu'une corrélation ne signe pas nécessairement une causalité. Les problèmes qu'on se pose n'ont pas changé de nature, ni les données de signification par la seule vertu de leur nombre. Mais, une fois formulées les hypothèses, la disponibilité d'une grande masse de données permet de multiplier les tests et les allers et retours entre les hypothèses et les données.
A condition, note un participant, que la qualité de celles-ci soit bonne, faute de quoi leur quantité même peut être source de graves erreurs dans les conclusions tirées de leur exploitation. Ce qui amène une autre personne à dire que cette qualité peut être altérée par des comportements pervers ou simplement malicieux des producteurs des données. Par exemple, pour avoir plus de chances d'être au calme dans un restaurant, d'aucuns ne seront-ils pas tentés de mettre en ligne des commentaires dissuasifs concernant la cuisine qu'on y sert ou son ambiance lumineuse ?

Indépendamment d'altérations volontaires des données, observe un participant, il peut y avoir des biais importants liés à la nature de la collecte, qui ignore des zones aveugles. Par exemple, en matière d'utilisation de la téléphonie mobile, les personnes en grande difficulté économique sont sans doute sous-représentées dans les bases de données puisées à cette source ; a contrario, les jeunes sont certainement surreprésentés si on utilise des informations issues de Facebook. Sait-on corriger cela ? A quoi l'invitée ajoute, tout en confirmant l'existence de zones aveugles, que la donnée est un construit social. La donnée brute n'est pas neutre. Il existe des dérives possibles. On en a un bon exemple avec l'ouragan Sandy (NDR : en 2012, dans l'Atlantique nord), dont l'impact mesuré sur les réseaux sociaux pouvait donner l'impression que l'ouragan n'avait frappé que les zones de population aisée et pratiquement pas les banlieues ! Cela dit, même constituées avec soin et exemptes de zones aveugles, les bases de données ne sont pas d'une utilisation facile pour tous.

Commission Barroso I, elle devient commissaire chargée de la société de l'information et des médias. Dans la Commission Barroso II, elle remplit une troisième fonction en devenant vice-présidente et commissaire chargée de la justice, des droits fondamentaux et de la citoyenneté.

[13] Wikipédia : Hadoop est une structure logicielle libre écrite en Java destinée à faciliter la création d'applications distribuées entre plusieurs machines (au niveau du stockage des données et de leur traitement) et adaptables au volume de la demande. Ces applications peuvent travailler avec des milliers de nœuds et de très grandes quantités de données. Ainsi chaque nœud est constitué de machines standard regroupées en grappe. Tous les modules de Hadoop sont conçus dans l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par la structure.

[14] Le Sniiram est le système national d'information inter-régimes de l'Assurance maladie. Il a pour finalités l'amélioration de la qualité des soins, la connaissance des dépenses de l'ensemble des régimes d'assurance maladie, la contribution à une meilleure gestion des politiques de santé et la transmission aux prestataires de soins des informations pertinentes relatives à leur activité, à leurs recettes et, s'il y a lieu, à leurs prescriptions. Les données collectées sont décrites à l'adresse suivante :
http://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/sniiram/description-des-donnees-et-wiki-sniiram.php

Ainsi, dans le cas des données de l'opérateur Orange, la réutilisation par des tiers n'est en rien immédiate. En fin de compte, rien de tout cela n'est vraiment nouveau pour des professionnels de la statistique.

Ce qui l'est davantage est ce qui touche à la gouvernance et à la circulation de la donnée. Qui contrôle l'usage des données ? Pour l'invitée, les données doivent être contrôlées et valorisées sur le double plan de leur valeur d'usage et de leur valeur monétaire. Cela soulève un débat de fond. En effet, la donnée est coproduite par nous tous et par le collecteur en ligne. Or, le collecteur, qu'on appelle aussi responsable de traitement, dispose de fait de l'intégralité de la valeur alors qu'il serait légitime de penser partage de la valeur. Des réflexions ont suggéré la mise en place d'une nouvelle assiette fiscale, qui tiendrait compte de l'endroit où les données ont été produites. Une autre piste - explorée par l'invitée avec la FING (Fondation Internet Nouvelle Génération) et un consortium d'entreprises, dont la MAIF, EDF et la Société générale - est celle du « *Self Data* » : on restitue les données à l'utilisateur et c'est lui qui décide de leur utilisation.

Sont inédites aussi certaines questions que le Big Data pose à l'enseignement. Selon un participant, les impacts sont déjà pris en compte sous plusieurs formes : les MOOC[15] sont devenus un standard et des certificats sont obtenus par ce moyen. Il existe par ailleurs des plates-formes de compétition (les *Challenge Data*[16]) et l'évaluation des élèves est faite sur cette base.
Les formations ont beaucoup évolué depuis dix ans. Dans le mastère de sciences des données, la statistique est enseignée avec le « *Machine Learning*[17] » et d'autres disciplines et cet enseignement est orienté vers le prédictif. L'intervenant se demande à quel moment du cursus scolaire faire évoluer les enseignements : dès le lycée ? plus tard ? Une autre personne s'interrogera sur les moyens informatiques mis à la disposition des étudiants universitaires.
De l'avis de l'invitée, l'enseignement est un point essentiel et sensible. Elle cite le cas d'un accord passé (sans appel d'offres) entre Microsoft et l'Éducation nationale aux termes duquel des données sur les élèves étaient collectées pour analyser leur comportement, avec des algorithmes opaques. Cette affaire a suscité un tel tollé qu'une charte d'autorégulation est en cours de rédaction. Il faut prendre conscience que de plus en plus de données sur les élèves vont être générées par les dispositifs numériques, par exemple avec les classes inversées[18]. Si de tels fichiers pouvaient être revendus, des discriminations entre élèves au moment du passage au niveau universitaire deviendraient possibles.

---

[15] NDR : Les MOOC (*Massive Open Online Courses*, ou cours en ligne ouverts et massifs) sont nés dans les universités américaines. Ils permettent aux étudiants - et à tous ceux qui se connectent - de se former en ligne et d'interagir avec d'autres personnes, même à l'autre bout de la planète, le tout gratuitement. Un MOOC repose sur une plateforme Internet. On y trouve des cours sous forme de vidéos, de diaporamas, de cours rédigés, ainsi que des exercices et des espaces d'échanges interactifs. Les utilisateurs peuvent participer à l'élaboration des cours, partager entre eux des fichiers, les commenter ou en proposer de nouveaux.

[16] Il s'agit par exemple de compétitions d'apprentissage statistique destinées à des étudiants en master et doctorat, avec la participation de chercheurs. Elles sont proposées par des entreprises et sont issues de problématiques concrètes que ces entreprises rencontrent dans leur activité. Elles s'inscrivent dans un esprit d'échange scientifique, avec un partage des données et résultats : les données mises à disposition par les entreprises doivent être non-confidentielles et les rapports algorithmiques des élèves et des chercheurs sont mis à la disposition des entreprises. Une plate-forme web dédiée assure les échanges de données et l'évaluation automatique des résultats des participants.

[17] Wikipédia : L'apprentissage automatique ou apprentissage statistique (*Machine Learning* en anglais), champ d'étude de l'intelligence artificielle, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou que des moyens algorithmiques plus classiques peineraient à remplir.

[18] Dans les classes inversées, les élèves doivent étudier leurs cours chez eux, afin que les activités en classe deviennent plus concrètes. Durant les heures d' « apprentissage », ils font des exercices d'application et de découverte. Ce n'est plus l'enseignant qui apporte des connaissances d'un nouveau chapitre, mais il aide l'élève pour la compréhension des notions importantes et a davantage de temps pour le suivre au cas par cas. L'enseignant joue un rôle de guide dans les apprentissages de l'élève.

Comment toucher les jeunes ? Au Conseil national du numérique[19], auquel l'invitée a appartenu, deux rapports ont été produits en 2014/2015, l'un sur « Citoyenneté et numérique », l'autre sur « Éducation et numérique », qui développent le concept de « littératie numérique », à savoir le bouquet de connaissances nécessaires pour être un citoyen actif et éclairé : aptitudes à comprendre les principes de fonctionnement des algorithmes, à contrôler ses données personnelles, à produire, publier et partager des contenus en ligne, à décrypter l'environnement économique et social des technologies, les modes de financement de l'économie numérique et plus globalement à acquérir une culture de la science informatique…). Avec les jeunes, le mode de l'interdiction ou de la mise en garde défensive est inefficace ; mieux vaut privilégier le mode constructif et du partage, de manière à faire émerger des acteurs pro-actifs.

Statistique et Big Bata

Le Big Data condamne-t-il, à terme, les sondages traditionnels, demande un participant ? L'invitée estime qu'en matière d'enquêtes et de sondages, l'exploitation des données massives n'est pas la panacée. Il suffit pour s'en convaincre de lire ce que dana boyd[20] a écrit sur les sondages électoraux aux États-Unis[21] : les méthodes en sont décalées par rapport aux réalités sociales. On rencontre là les mêmes problématiques que pour la statistique.

Un participant précise que les enquêtes statistiques se veulent représentatives des populations étudiées alors que les échantillons du Big Data sont très différents. Les enquêtes traditionnelles concernent peu de personnes à qui on pose beaucoup de questions, parfois compliquées. Dans les recensements, au contraire, on pose peu de questions – et des questions simples - à beaucoup de personnes. On peut tenter de combiner les sources de données, et notamment utiliser des informations issues des mégadonnées. Des données commerciales sont parfois accessibles et, dans ce cas, on peut déboucher sur une heureuse complémentarité des méthodes grâce au Big Data. S'en tenir à ce dernier trouvera rapidement des limites. Par exemple, en matière de tourisme, on pourra mesurer des flux en temps réel mais on ne saura rien des pratiques culturelles des personnes composant ces flux (à moins de tracer leurs téléphones cellulaires ?). Il faudrait mettre au point des protocoles sérieux permettant, en particulier, la reproductibilité des études. Où en est-on sur ce point ?
Cette intervention est l'occasion pour l'invitée de préciser que les données n'appartiennent pas au collecteur. On vit dans un régime « personnaliste », c'est-à-dire qu'on préserve les droits de la personne. Une approche de ces questions misant sur une régulation par le seul marché serait dangereuse.

Un autre participant rappelle que les statistiques publiques sont l'objet d'une certaine désaffection, imputable sans doute à la méfiance grandissante dans beaucoup de pays envers les institutions, mais imputable aussi, vraisemblablement, à la confusion découlant de la prolifération des informations. De ce point de vue, on pourrait imaginer que la statistique publique, caractérisée par la rigueur de ses procédures et de ses méthodes, se consacre à l'observation périodique des faits structurant la société et aux nomenclatures, le tout étant éventuellement enrichi grâce aux données massives (avec des problèmes analogues à ceux que les statisticiens publics ont rencontrés quand ils ont commencé à exploiter des fichiers administratifs). On serait ici dans la durée longue.
A contrario, la statistique privée sera toujours le domaine des défricheurs et de la créativité, dans la durée courte. On peut esquisser une typologie rapide :

---

[19] https://cnnumerique.fr/

[20] NDR : Danah Michele Mattas, alias Danah Boyd, nom qu'elle écrit danah boyd pour des raisons personnelles et politiques, est née en 1977 à Altoona (Pennsylvanie). Elle est chercheuse en sciences humaines et sociales, spécialisée dans l'étude des médias sociaux et de leur utilisation par la jeunesse.

[21] Traduction française « danah boyd : J'en veux aux médias. Il est temps d'atterrir. Arrêtez avec les sondages. » http://vecam.org/danah-boyd-J-en-veux-aux-medias-Il-est-temps-d-atterrir-arretez-avec

- dans le cas des assureurs détenteurs de nombreuses données individuelles, on peut espérer que des dispositifs institutionnels éloigneront les dangers signalés par l'invitée ;
- les entreprises utilisant pour leurs besoins propres les données dont elles disposent légalement conduiront l'ensemble de leurs études selon leurs normes ;
- le danger vient plutôt des personnes physiques ou morales utilisant les résultats de leurs études pour prendre des positions publiques susceptibles d'influencer l'opinion. On pourrait exiger d'elles qu'elles se réfèrent à une charte. La loi a pu obtenir, en matière de sondages électoraux, la conformité à des règles précises, sous le contrôle de la Commission des sondages. On pourrait imaginer une charte nationale, voire de l'Union européenne, ou - pourquoi pas ? - une charte universelle, faisant obligation à ces personnes diffusant de l'information d'indiquer les objectifs de leurs études, l'identité des financeurs, l'origine des données, les méthodes et traitements appliqués, le tout avec déclaration de conformité à la charte, celle-ci étant assortie de possibilités de contrôle et de sanction[22].

Présent aux débats, le président de l'Autorité de la statistique publique[23] rappelle que la statistique publique est élaborée selon les règles de l'art, qui ne sont pas incompatibles en effet avec un enrichissement par les mégadonnées. Il y a là des opportunités pour la statistique publique : l'exploitation des données de caisse du grand commerce de distribution, des transactions par carte bancaire ou encore de la téléphonie apparaît prometteuse. Pour les données de caisse, on est dans la continuité sur la méthodologie de calcul de l'indice national des prix à la consommation. Des réflexions au sein de l'Insee éclairent ces sujets qui soulèvent beaucoup de questions, par exemple en matière de secret statistique. La loi sur le numérique cadre bien les choses pour le recours aux données détenues par les entreprises : elle exige un intérêt public (contrôlé par le Conseil national de l'information statistique) et la confidentialité est garantie aux fournisseurs des données. La qualité des données doit être examinée cas par cas. Il convient donc que la statistique publique continue de rechercher les opportunités d'utiliser les potentialités du Big Data. Au passage, on peut signaler que la valorisation des données serait à prendre en compte dans le PIB. La statistique publique elle-même apparaît de plus en plus comme un bien public et son utilité doit être justifiée car elle est financée par l'impôt.

Quelle société voulons-nous ?

Plusieurs participants s'inquiètent de l'usage qui est fait des données individuelles captées par de multiples voies et de leur circulation. Quand les candidats à des élections utilisent les techniques du Big Data, ont-ils accès à des données individuelles ? En principe, non. Assurément oui en ce qui concerne les adhérents d'un parti : les responsables de ce parti utilisent ces données personnelles des adhérents afin de mobiliser ces derniers de manière individualisée. Lors de sa première campagne électorale, Barack Obama a su utiliser son expérience de travailleur social pour organiser les tournées de porte-à-porte de ses partisans. Une fois qu'il est arrivé au pouvoir, les mêmes méthodes ont échoué pour faire vivre une démocratie véritablement contributive. On est loin encore, par ces procédures d'exploitation de fichiers en fin de compte assez opaques, de la démocratie participative. *A contrario*, un exemple de ce que celle-ci pourrait être a été donné par l'élaboration de la loi sur le numérique. Chargée de ce dossier, Axelle Lemaire, secrétaire d'État chargé du numérique, a conduit un long processus d'élaboration du projet de loi, puis a soumis ce

[22] Cette idée rejoint la question d'un participant : Ira-t-on vers un "label rouge" distinguant les traitements Big Data / Data Analytics irréprochables sur le plan éthique ?

[23] http://www.autorite-statistique-publique.fr/asp/. L'Autorité de la statistique publique veille à l'indépendance professionnelle dans la conception, la production et la diffusion de statistiques publiques. Elle assure également une vigilance quant au respect des principes d'objectivité, d'impartialité, de pertinence et de qualité des données produites, en référence aux recommandations européennes en matière de bonnes pratiques statistiques.

dernier à discussion publique. C'est une première intéressante car on est sorti de l'entre-soi et on a ainsi mis en lumière le travail des lobbies, ce qui a été une véritable révolution !

Au moins le régime juridique de l'Union européenne est-il plus protecteur des données individuelles que celui des États-Unis. Aux USA, la circulation de ces données se fait uniquement sur une base contractuelle. Le règlement européen de protection des données personnelles (RGPD) dont l'Union européenne vient de se doter[24] et qui entrera en application en mai 2018 renforce le droit de la personne avec, entre autres, la notion de « portabilité » : l'individu producteur de la donnée personnelle pourra la récupérer au bénéfice d'un autre détenteur ou de lui-même (« *Self Data* »). Cela entraîne une tension avec les régimes juridiques américain et chinois. Il est à noter que le RGPD repose sur trois pieds : juridique, technologique et organisationnel. Ce dernier appui doit être pris dès le stade de la conception des applications informatiques, c'est ce qu'on appelle le « *Privacy by Design* ».
Peu à peu, ajoute l'invitée, le contrôle du citoyen sur ses propres données se renforce, mais c'est un processus de longue haleine. Un citoyen autrichien a pu attaquer Facebook au motif que la protection des données individuelles le concernant était insuffisante aux États-Unis. Il a eu gain de cause et cela a débouché sur l'annulation d'un accord entre l'Union européenne et les États-Unis (le *Safe Harbour*) et la passation d'un nouvel accord (le *Privacy Shield*), qui est d'ailleurs encore imparfait. D'où la nécessité d'une possibilité d'action collective, ce qu'on appelle la *Class Action* en anglais. En France, elle n'était possible depuis la loi Hamon que pour des dommages sur des biens matériels. La loi de modernisation de la justice du 21e siècle adoptée en novembre 2016 l'a étendue entre autres aux préjudices sur les données personnelles.

Où en est le marketing prédictif[25], demande un participant ? Le marketing prédictif (« *Targeting* » ou ciblage) n'est pas encore excellent mais il fait appel, d'ores et déjà, à des techniques raffinées, qui se perfectionnent de jour en jour. Si je m'intéresse à un voyage dans tel pays et si je suis lecteur de tel quotidien, je recevrai des publicités ciblées, avec un algorithme d'enchères autour des données concernant ma personne, même si mon nom reste inconnu du système. Certaines personnes se satisfont de ce qu'elles considèrent comme un service gratuit qui leur est rendu : une publicité adaptée à leurs centres d'intérêt ; d'autres se soucient de cette collecte d'informations personnelles et estiment perdre en plaisir de la découverte dès lors que le système les enferme dans ce qu'il croit savoir d'elles.
Plus profondément, l'inquiétude est aussi liée à ce qu'on appelle « la loyauté des algorithmes » : quand un site de e-commerce me recommande en ligne des ouvrages, le fait-il à la lumière de mes préférences décryptées au travers de mes requêtes précédentes ou bien agit-il au nom d'un accord qu'il a passé avec l'éditeur de l'ouvrage qu'il me recommande ? Qu'y a-t-il dans les algorithmes

[24] Remplaçant la directive actuelle sur la protection des données, le règlement général de l'Union européenne sur la protection des données (RGPD) adopté le 14 avril 2016 étend la portée de la protection pour couvrir les données détenues non seulement par les personnes morales ou physiques européennes, mais également par les entreprises ou organismes non européens qui traitent les données de citoyens européens.
L'objectif de ce règlement est de redonner aux citoyens le contrôle de leurs données personnelles, tout en unifiant les réglementations relatives à la protection de la vie privée dans l'Union européenne. Ses dispositions vont entraîner de profonds changements en matière de collecte et de traitement des données et auront probablement un impact global sur le fonctionnement des entreprises.
La définition des données à caractère personnel couvre désormais toute une série de détails comprenant les informations personnelles habituelles, mais aussi des éléments tels que les photographies et les données des réseaux sociaux. D'autres défis liés au « droit à l'oubli » et au droit du citoyen à demander l'accès à ses données obligeront toutes les organisations à examiner de près leurs politiques relatives aux données des clients.
Le règlement impose, par exemple, d'obtenir un consentement explicite pour la collecte et l'exploitation de ces données. Des amendes ou sanctions sévères sont prévues en cas d'infraction. L'amende maximale s'élève désormais à 20 millions d'euros ou 4 % du chiffre d'affaires mondial annuel (la valeur la plus élevée étant retenue), ce qui représente un coût considérable pour la plupart des organisations. Le RGPD confère de facto au responsable de la protection des données un rôle beaucoup plus important en raison de l'impact des amendes ou sanctions potentielles en cas d'infraction.

[25] Le marketing prédictif regroupe les techniques de traitement et de modélisation des comportements des clients et prospects qui permettent d'anticiper leurs actions futures à partir du comportement présent et passé.

utilisés ? Faudra-t-il demain des « commissaires aux algorithmes » habilités à vérifier leur architecture ? La transparence en la matière ne va pas de soi car on touche au cœur de la valeur de l'entreprise et du secret de la marque. Et si on connaît le cheminement logique de l'algorithme, alors on peut tricher ! Au moins une vigilance collective pourrait-elle être organisée pour signaler sur une plate-forme les anomalies que l'on croit discerner. On jouerait ce faisant sur la réputation des entreprises, à laquelle elles sont très sensibles.
On trouve sur le site de la Cnil beaucoup de réflexions intéressantes sur ces sujets, indique un participant, qui note que certains systèmes ont même créé la catégorie des curieux présumés ! Assurément, conclut l'invitée, la publicité devient plus intelligente grâce à l'exploitation des données massives, mais acceptons-nous d'être ainsi catalogués de manière occulte au vu de nos comportements, à l'heure où il nous faudrait réfléchir à nos orientations collectives ?
Quant au marketing politique, largement utilisé par le candidat Trump aux États-Unis, un récent papier de *Libération*, reposant sur l'analyse des médias et des blogs alternatifs, défendait la thèse que ces derniers sont très propagateurs de rumeurs et d'informations déformées ou fausses, ce qui donne une viralité extrême à ces « *fake news* » susceptibles d'influencer les électeurs à travers de véritables manipulations collectives[26].
À propos d'élections, un participant signalera les travaux de la société canadienne Filteris, qui estime les chances de succès des candidats aux grandes élections à partir de leur « poids politique » sur Internet et sur les réseaux sociaux.

Comment assurer la sécurité de mes propres données ?

Que penser de la sécurité des données et de celle des systèmes, interrogent plusieurs participants ? En tant que citoyens, que pouvons-nous faire pour nous prémunir alors que 98 indicateurs sont collectés par Facebook sur chaque utilisateur de ce réseau social ?
En effet, des questions de sécurité se posent ; il suffit de voir combien il y a de rackets et de demandes de rançon sur le Web, par exemple auprès d'utilisateurs de sites de rencontres adultères.
Les protections individuelles existent, précise l'invitée. Ainsi, on peut utiliser le moteur de recherche Qwant plutôt que Google, même si ses performances ne sont pas encore du même niveau. On peut refuser les cookies et se montrer prudent sur les services proposés, comme on peut utiliser les fonctions de blocage des accès publicitaires, etc. (Voir en bibliographie l'ouvrage de Tristan Nitot).

On s'inquiète de l'émergence des données massives, mais le danger ne viendrait-il pas plutôt de l'opacité des algorithmes qui exploitent ces données et permettent des profilages sans que l'usager final en ait toujours conscience, demande un participant ? Un autre s'inquiète de la place future de l'humain dans un environnement où le Big Data rend possible l'automatisation des décisions : qu'automatiser, selon quels types de raisonnements et sous quels contrôles ?
L'invitée confirme que ce sont bien les algorithmes qui font tourner le Big Data. La loi sur le numérique a pointé leur importance. Dans le cas d'un logiciel d'orientation automatisée vers l'enseignement supérieur, il a été demandé que le code source soit accessible au titre de la loi sur l'accès aux documents administratifs (loi Cada). La Cnil vient de lancer un vaste chantier de réflexion éthique sur le thème des algorithmes, thème à la fois bouillant et compliqué[27]. Enfin, note l'invitée, on est obsédé en Occident par la puissance et la dangerosité potentielle des GAFA (Google, Amazon, Facebook, Apple) quant à l'utilisation des données individuelles et on néglige les grands acteurs équivalents de la Chine. Heureusement, face à toutes ces entreprises puissantes et face aux tentations inquisitoriales des États, on constate des initiatives civiles de contournement des

[26] Facebook, un mois dans la machine à infos http://www.liberation.fr/futurs/2017/03/12/facebook-un-mois-dans-la-machine-a-infos_1555220

[27] La Cnil organise un débat public décentralisé sur le thème "Ethique et numérique : les algorithmes en débat". Son objectif est d'initier un processus de discussion collectif que feront vivre tous ceux – institutions publiques, société civile, entreprises – qui souhaitent y prendre part en organisant des débats et manifestations multiformes. Elle assurera la coordination et la cohérence de ces diverses manifestations.
Voir https://www.cnil.fr/fr/ethique-et-numerique-les-algorithmes-en-debat-0

censures faisant preuve d'une belle inventivité sociale (cf. le Mouvement des parapluies à Hong Kong en septembre-octobre 2014. Ou encore les outils comme TOR utilisés par les résistants aux régimes autoritaires dans l'Internet souterrain[28]).

La Cnil a-t-elle les moyens de traiter tous les problèmes évoqués plus haut, demande un participant ? A quoi l'invitée répond que la Commission, qui emploie 180 à 200 personnes, essentiellement des juristes et des ingénieurs, doit traiter des sujets de plus en plus nombreux. Des systèmes de normes uniques sont progressivement développés pour simplifier les procédures au bénéfice des responsables de traitement. Le RGPD applicable en mai 2018 devrait alléger les tâches d'instruction de dossiers. En effet, ce sont les responsables des traitements de données dans les entreprises et institutions qui auront à assurer le respect des règles, sans avoir à faire de déclaration pour chaque traitement et la Cnil pourra se concentrer sur les autorisations de traitements de données sensibles et sur la fonction de contrôle, avec un pouvoir de sanction considérablement augmenté.

Au fil de la soirée, de nombreuses questions ont émergé, que le temps disponible n'a pas permis d'approfondir : A-t-on un jour demandé aux Français s'ils sont d'accord pour qu'on utilise leurs données ? Qui peut avoir accès à quelles données ? Saurait-on "piéger" les acteurs malveillants, non éthiques ou non rigoureux du Big Data et du *Data Analytics*[29] ? Ira-t-on vers un "*testing*" du Big Data ? Existe-t-il un catalogue d'algorithmes éthiques pour certains besoins définis ?[30]

Toutes ces questions, et bien d'autres, ont autant de dimensions collectives que de résonances individuelles. Comme toujours, l'avancée des sciences et des techniques soulève autant de problèmes qu'elle n'aide à en résoudre. L'élément nouveau est peut-être que jamais les choix sociétaux nécessaires en matière d'information n'ont rendu si souhaitable une prise de conscience citoyenne des enjeux.

Petite bibliographie :

- « Terra Data - Qu'allons-nous faire des données numériques ? » livre de Serge Abiteboul et Valérie Peugeot publié en mars 2017 à l'occasion de l'exposition du même nom à la Cité des sciences de La Villette – éditions Le Pommier – 13 €

[28] NDR : L'Internet est organisé en trois couches : l'Internet de surface (celui d'utilisation courante, indexé par les moteurs de recherche, et qui représenterait moins de 5 % des contenus globaux ; l'Internet profond (*deep Web*), auxquels on peut accéder via des moteurs de recherche spécifiques ; on y trouve, entre autres, de l'information universitaire et des rapports scientifiques ; enfin, l'Internet sombre (*dark Web*), lieu privilégié des informations illégales, des trafics et des activités d'opposition dans les pays non démocratiques. On peut y accéder par des logiciels comme Tor (*The Onion Router*). La navigation s'y fait de manière cryptée et elle est très difficilement traçable. Là, pas d'indexation donc pas de moteurs de recherche. Il faut avoir l'adresse de là où on veut aller… et prendre des précautions.

[29] NDR : Pratique consistant à analyser des données brutes pour en tirer des conclusions opérationnelles.

[30] À propos des "algorithmes éthiques", on peut prendre connaissance de la démarche engagée par l'Inria (Institut national de recherche en informatique et en automatique). Son projet TransAlgo ambitionne d'évaluer la responsabilité et la transparence des systèmes algorithmiques. Voir https://www.inria.fr/actualite/actualites-inria/transalgo

- « Les Big Data à découvert » sous la direction de Mokrane Bouzeghoub et Rémy Mosseri CNRS éditions 16 mars 2017 – 39 €

- « Surveillance ://Les libertés au défi du numérique : comprendre et agir » Tristan Nitot – C&F éditions – 2016 – 19 €

- La revue « Statistique et société », revue de la SFdS, publie régulièrement des articles consacrés aux Big Data. Un dossier « BigData, entre régulation et architecture » est paru dans le volume 2 n° 4 (décembre 2014). Un dossier « BigData, marketing, consommateurs, citoyens et entreprises » est paru dans le volume 4 n° 3 (décembre 2016).
Ces dossiers, et tous les articles de la revue sont téléchargeables librement sur le site
www.statistique-et-societe.fr

- Un article tout récent, parmi une myriade : « Les limites des modèles comportementaux du big data » Paul Seabright dans le journal Le Monde du 16 mars 2017.

Annexes :

Exemples de réalisations et de promesses du Big Data permettant de réfléchir sur l'efficacité des nouveaux outils de connaissance, sur leurs limites et sur leurs conséquences sociales :
- dans le domaine de la santé (annexe1) ;
- dans celui de l'agriculture (annexe 2) ;
- et dans celui des statistiques publiques (annexe 3) ;

Par ailleurs, quelques définitions sont reprises en annexe 4.

Annexe 1

Extrait de « La santé à l'ère et à l'aune du « big data » - article de Daniel Eilstein et Jérôme Pozuelos paru dans la revue Futuribles n° 412 – mai-juin 2016
[Il s'agit de la conclusion de cet article]
L'article de Futuribles peut être acheté sur le site de cette revue www.futuribles.fr (7 €)

La question du *big data* est un véritable enjeu pour la santé et est désormais prise en compte, en France, comme axe de développement stratégique de la future Agence nationale de santé publique. À moyen terme, le *big data* devient un outil au service de la veille sanitaire et de la prévention (grâce au *data mining* notamment, ou à l'ouverture à nouvelles sources de données : réseaux sociaux, objets connectés, Google, etc.), mais qui pourra aussi être mis au service de la veille prospective permettant d'anticiper les futures menaces.

Le *big data* est indissociable des questions de sécurité des systèmes d'information et un certain nombre d'obstacles doivent être franchis pour assurer une fiabilité suffisante (continuité, traçabilité, confidentialité, intégrité des données), ce qui peut aujourd'hui conduire certains à exprimer leur scepticisme. On peut néanmoins supposer sans grand risque que les intérêts économiques liés à une bonne utilisation du *big data* (assurances, *marketing,* services aux personnes, services publics avec les villes intelligentes, raccourcissement des délais de recherche-développement, détection de la fraude, etc.) conduiront l'intelligence humaine à fiabiliser encore cette technologie.

À plus long terme, nous pouvons imaginer que la convergence des NBIC (nanotechnologies, biotechnologies, technologies de l'information et sciences cognitives), des techniques d'exploitation du *big data,* de l'intelligence artificielle et de la génétique dans les projets de recherche, accélère notre capacité d'innovation dans ce domaine et qu'une révolution du contexte technique ait lieu prochainement (d'ici 2020), modifiant profondément notre rapport à la santé. La capacité des processus de recherche dans ce domaine étant augmentée par un certain nombre de tendances: loi de Moore, *open data, crowdfounding,* autres évolutions techniques ...

Les questions et enjeux seront d'autant plus importants que beaucoup de nos repères pourraient être modifiés, y compris sur le plan sociétal. Quelle place pour le médecin lorsque l'intelligence artificielle couplée aux nanotechnologies et les thérapies géniques pourront nous « réparer » sans même que nous en ayons conscience ? Qui aura accès à ces futures technologies ? Existera-t-il une rupture entre l'*Homo sapiens* et l'*Homo data ?* Quels seront nos comportements dans un contexte où le moindre écart connu pourrait nous empêcher l'accès aux services assuranciels ou au crédit ? Quel avenir pour la santé publique lorsque certaines maladies seront sous contrôle technologique et les comportements à risque maîtrisés par les enjeux économiques ? Comment seront financées les retraites et l'assurance maladie dans un tel contexte ? Qu'en sera-t-il des pratiques eugénistes ? Les couples s'uniront-ils en fonction de leur génotype afin d'optimiser le potentiel de leur progéniture ? Quelle sera la politique de natalité dans un monde aux ressources limitées où la mortalité sera sans cesse repoussée ? Comment notre identité sera-t-elle associée à notre information génétique ? Risque-t-elle d'être usurpée ?

Annexe 2

Extrait de « La data et le territoire », interview de François Houllier, président-directeur général de l'INRA, par Serge Abiteboul et Claire Mathieu, parue sur le blog « Binaire » le 11 décembre 2015
L'interview de François Houllier est consultable à l'adresse :
http://binaire.blog.lemonde.fr/2015/12/11/la-data-et-le-territoire/

Le monde agricole s'intéresse beaucoup au big data. Comme ailleurs, cela semble causer des inquiétudes, mais pourrait être aussi une belle source de progrès. Comment voyez-vous cela ?

FH : Nous voyons arriver le big data sous deux angles différents, celui de la recherche et celui de l'agriculture.

Premier angle : la recherche, pour laquelle le big data a une importance énorme. Considérons, par exemple, l'amélioration génétique classique : on cherche à utiliser de plus en plus précisément la connaissance du génome des animaux et des végétaux en repérant des « marqueurs » le long des chromosomes ; ces marqueurs permettent de baliser le génome et de le cartographier. Les caractères intéressants, comme le rendement ou la tolérance à la sécheresse, sont corrélés à de très nombreux marqueurs. On va donc faire des analyses sur les masses de données dont on dispose : beaucoup d'individus sur lesquels on identifie la présence ou l'absence de beaucoup de marqueurs qu'on corrèle avec un grand nombre de caractères. L'objectif est de trouver des combinaisons de marqueurs qui correspondent aux individus les plus performants. On sait faire cela de mieux en mieux, notamment à l'INRA. Les grands semenciers le font aussi : ils investissent entre 10 et 15 % de leurs ressources dans la R&D. Aujourd'hui, la capacité bioinformatique à analyser de grandes quantités de données devient un facteur limitant. […]

Deuxième angle : l'utilisation du big data chez les agriculteurs. Un robot de traite est équipé de capteurs qui produisent des données. Un tracteur moderne peut aussi avoir des capteurs, par exemple pour mesurer la teneur en azote des feuilles. Avec les masses de données produites, nous avons vu se développer de nouveaux outils d'analyse et d'aide à la décision pour améliorer le pilotage des exploitations. Mais ce qui inquiète le monde agricole, c'est : qui va être propriétaire de toutes ces données ? Qui va faire les analyses et proposer des conseils sur cette base ? Est-ce-que ces données vont être la propriété de grands groupes comme Monsanto, Google, ou Apple ou les fabricants de tracteurs ? En face de cela, même les grandes coopératives agricoles françaises peuvent se sentir petites. Le contrôle et le partage de toutes ces données constituent un enjeu stratégique.

Il ressort de tout cela que l'agriculteur est souvent très connecté ?

FH : Il reste bien sûr des zones dans les campagnes qui sont mal couvertes par Internet, mais ce n'est pas la faute des agriculteurs. Les agriculteurs sont plutôt technophiles. Quand les tracteurs, les robots de traite ou les drones sont arrivés, ils se sont saisis de ces innovations. Il en va de même avec le numérique. Les agriculteurs qui font de l'agriculture biologique sont eux aussi favorables au numérique. Les nouvelles technologies permettent aux agriculteurs de gagner du temps, d'améliorer leur qualité de vie, de réduire la pénibilité de certaines activités. Ils sont conscients des améliorations que les applications informatiques peuvent leur apporter.

Annexe 3

Extrait de « Comment la statistique a perdu son pouvoir – et pourquoi nous devrions craindre ce qui va suivre » - Traduction de l'article de William Davies paru dans The Guardian le 19/1/2017
L'article de William Davies est consultable en anglais sur le site du journal The Guardian :
https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-datademocracy?CMP=Share_iOSApp_Other
Une traduction de cet article paraîtra dans le prochain numéro de « Statistique et société » (avril 2017)

Depuis quelques années, une nouvelle façon de quantifier et visualiser les populations apparaît, qui potentiellement marginalise la statistique et sonne l'avènement d'une ère toute différente. La statistique, collectée et compilée par des techniciens experts, laisse la place aux données qui s'accumulent automatiquement, du fait de la numérisation envahissante. Traditionnellement, les statisticiens savaient quelles questions ils voulaient poser et quelle était la population concernée, puis ils allaient chercher les réponses. En revanche, les données se produisent d'elles-mêmes lorsque nous utilisons une carte de fidélité, laissons un commentaire sur Facebook ou cherchons quelque chose sur Google. Comme nos villes, nos voitures, nos maisons et les objets du ménage sont dorénavant connectés, la masse des données que nous laissons dans notre sillage va devenir encore plus considérable. Dans ce monde nouveau, les données sont captées d'abord : les questions de recherche viennent ensuite.
À terme, cela aura des implications probablement aussi profondes que l'invention de la statistique à la fin du 17e siècle. L'expansion des « données massives » fournit des occasions d'analyses quantitatives beaucoup plus abondantes qu'autant de sondages ou de modèles statistiques que vous voudrez. Et ce n'est pas seulement la quantité des données qui diffère. C'est un type de connaissance entièrement différent, accompagné d'un nouveau mode d'expertise.
Tout d'abord, il n'y a aucun cadre d'analyse déterminé (comme la nation) ni aucune catégorie constituée (comme « chômeur »). Ces énormes nouveaux stocks de données peuvent être explorés pour rechercher des motifs, des tendances, des corrélations et des tendances émergentes. Cela devient une façon de suivre à la trace les identités que les gens s'attribuent (comme «#JesuispourCorbyn » ou « entrepreneur ») plutôt que de leur assigner une catégorie. C'est là un mode d'agrégation approprié à une ère politique plus fluide, où tout ne peut être – comme avec les Lumières – relié de façon fiable à un quelconque idéal d'État-nation gardien de l'intérêt public.
En second lieu, la plupart d'entre nous oublions totalement ce que toutes ces données disent de nous, individuellement ou collectivement. Il n'y a aucun équivalent d'un Office national de statistique pour les données de masse enregistrées par le commerce. Nous vivons à une époque où nos sentiments, identités ou affiliations peuvent être pistés et analysés à une vitesse et avec une précision sans précédent - mais rien ne rattache cette nouvelle capacité à l'intérêt public ou au débat public. Des « data-analystes » travaillent pour Google et Facebook, mais ce ne sont pas des « experts » du même genre que ceux qui produisent la statistique et qui sont maintenant si largement condamnés. L'anonymat et le secret des nouveaux analystes les rendent potentiellement plus puissants politiquement que tout spécialiste des sciences humaines.
Une entreprise comme Facebook est capable de procéder à une analyse sociale quantitative sur des centaines de millions de gens, à très bas prix. Mais elle est très peu incitée à en révéler les résultats. En 2014, quand les chercheurs de Facebook ont publié les résultats d'une étude « de la contagion émotionnelle » qu'ils avaient effectuée sur leurs usagers – où ils avaient modifié des fils d'actualités pour voir comment cela affectait les contenus que les usagers partageaient en réponse – ce fut un tollé : les gens avaient été soumis à une expérience à leur insu. Ainsi, du point de vue de Facebook, pourquoi, en publiant, aller au-devant de tracasseries ? Pourquoi ne pas plutôt faire l'étude et se taire ?

Annexe 4

Extrait de « Analyse des big data : quels usages, quels défis ? » Marie-Pierre Hamel et David Marguerit, Commissariat général à la stratégie et à la prospective, note d'analyse n° 8 de novembre 2013
Cette étude est disponible sur le site de France Stratégie :
http://www.strategie.gouv.fr/publications/analyse-big-data-usages-defis

ENCADRÉ 1. ÉLÉMENTS DE DÉFINITION

Big data : Énormes volumes de données structurées et non structurées, difficilement gérables avec des solutions classiques de stockage et de traitement. Ces données proviennent de sources diverses et sont (pour la plupart) produites en temps réel.

Cloud computing : Désigne des prestations à distance - logiciels, stockage de données - physiquement réparties dans des *data centers* et non pas sur le terminal de l'utilisateur.

Datamining : Ensemble de techniques ayant pour objet l'extraction d'un savoir à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

Internet des objets : Désigne les objets connectés à internet qui transmettent des données numériques par le biais de puces radiofréquences (RFID). Ces objets peuvent communiquer entre eux. On les retrouve dans la grande distribution, dans les objets du quotidien (podomètres connectés, domotique, compteurs électriques intelligents), dans les avions, les voitures, dans le monde médical, etc.

Open data : Processus d'ouverture des données publiques ou privées pour les rendre disponibles à l'ensemble de la population sans restriction juridique, technique ou financière. *L'open data* contribue à l'augmentation des données disponibles à l'analyse.