Statistique et société
décembre 2014

Volume 2, Numéro 4

BigData

entre régulation et architecture

www.statistique-et-societe.fr
© Société Française de Statistique (SFdS)

Sommaire
Statistique et Société

Volume 2, Numéro 4

Sommaire
Statistique et Société

Volume 2, Numéro 4

Statistique et société

Magazine trimestriel publié par la Société Française de Statistique.
Le but de Statistique et société est de montrer d'une manière attrayante et qui invite à la réflexion l'utilisation pratique de la statistique dans tous les domaines de la vie, et de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Rédaction
Rédacteur en chef : **Emmanuel Didier**, CNRS, France
Rédacteurs en chef adjoints :
Jean-Jacques Droesbeke, Université Libre de Bruxelles, Belgique
François Husson, Agrocampus Ouest, France
Jean-François Royer, SFdS - groupe Statistique et enjeux publics, France
Jean-Christophe Thalabard, Université Paris-Descartes, pôle de recherche et d'enseignement supérieur Sorbonne Paris Cité, France

Comité éditorial
Représentants des groupes spécialisés de la SFdS :
Ahmadou Alioum, groupe Biopharmacie et santé
Christophe Biernacki, groupe Data mining et apprentissage
Alain Godinot, groupe Statistique et enjeux publics
Delphine Grancher, groupe Environnment
Marthe-Aline Jutand, groupe Enseignement
Elisabeth Morand, groupe Enquêtes
Alberto Pasanisi, groupe Industrie

Autres membres :
Jean Pierre Beaud, Département de Science politique, UQAM, Canada
Corine Eyraud, Département de sociologie, Université d'Aix en Provence, France
Michael Greenacre, Department of Economics and Business, Pompeu Fabra Université de Barcelone, Espagne
François Heinderyckx, Département des sciences de l'information, Université Libre de Bruxelles, Belgique
Dirk Jacobs, Département de sociologie, Université Libre de Bruxelles, Belgique
Gaël de Peretti, INSEE, France
Theodore Porter, Département d'histoire, UCLA, États-Unis
Carla Saglietti, INSEE, France
Patrick Simon, INED, France

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial
Au-delà du consentement ?

Emmanuel DIDIER
Rédacteur en chef de Statistique et Société

Cher lecteur, bonne année.

Elle a commencé par un séisme. Les attentats ont mis en évidence des forces opposées et immenses qui pèsent sur la France, à en juger par la violence des actes autant que par la mobilisation de chefs d'Etats étrangers sans précédent, qu'ils ont suscitée. Pour faire face, les Français semblent rechercher les fondements de leur identité collective afin d'anticiper les évolutions à venir et de définir ensemble un futur désirable. Nombreux sont ceux qui voient dans la diversité de la société un de ses fondements sur lequel s'appuyer. A *Statistique et Société*, nous savons que la statistique sera, dans ce processus, un outil dont l'utilisation aura un rôle politique et social crucial. Et nous serons là pour éclairer ce rôle.

La présente livraison porte sur les BigData. Nous avions abordé le sujet à propos des données de santé (numéro 2 de cette année). Cette fois nous le traitons en publiant des articles issus d'un séminaire organisé le 22 mai 2014 par la SFdS.
Une clef de lecture de cette nouvelle collection d'articles réside dans la question du consentement. Nous sommes habitués à penser que la protection des individus dans le monde informationnel est assurée par le fait qu'ils donnent (ou non) leur consentement à certains usages des informations qu'ils fournissent.
Mais les BigData remettent profondément en cause cette protection car, comme le rappellent ici la plupart des auteurs, elles consistent à utiliser des données à des fins qui, précisément, n'étaient pas prévues au moment de leur collecte. Les BigData sont par essence des « Impredictible Data ». Comment réarmer le citoyen contre d'éventuels mésusages des données ?
On peut distinguer artificiellement trois dimensions. La première consiste à *renforcer le consentement* contre l'imprévisibilité des usages futurs. La seconde consiste à améliorer la *législation* sur la collecte des données. La troisième consiste à s'en remettre à la *technique*, et à confier aux architectures de données le rôle de protéger les individus. Bien entendu, ces dimensions ne sont pas exclusives les unes des autres. Mais la question de leur articulation n'en est pas réglée pour autant.
Il s'agit donc d'un débat extrêmement profond, qui engage jusqu'à nos conceptions de la citoyenneté et de la liberté. De fait, définitivement, nous n'avons pas fini de parler des BigData…
Ce numéro se prolonge avec un mini débat entre Yves Guiard, professeur émérite à Télécom-ParisTech et Thomas Piketty sur le libre choix des auteurs en matière de représentation graphique des chiffres. Et par une illustration de l'analyse multi-niveau sur le cas de la réussite scolaire. Enfin, le numéro se termine par un petit café convivial, comme nous en avons pris l'habitude, cette fois sur la consommation d'espace.

Emmanuel Didier

BigData : entre régulation et architecture
Introduction

Françoise DUPONT
SFdS – Groupe Statistique et Enjeux publics

Si les statisticiens sont directement concernés par le phénomène mégadonnées, celui-ci dépasse de loin le cadre de leur seule profession. Ils doivent élargir leur vision au delà des bénéfices effectifs de cette innovation pour s'interroger également sur l'impact de leurs pratiques du point de vue social. Au milieu du foisonnement d'écrits sur le sujet, ce dossier apporte une contribution en croisant la réflexion de professionnels issus de différentes disciplines tous concernés par les BigData. Il prolonge le séminaire organisé le 22 mai dernier par la SfdS[1] sous le titre « BigData : opportunités et risques ». Les opportunités, qui ne se limitent pas aux applications les plus fréquemment citées dans la presse, sont largement abordées par différentes contributions. Quant aux risques, ils ne peuvent être traités que par une réflexion pluridisciplinaire, qu'il s'agisse de la question de la protection des données personnelles ou de celle du profilage excessif des individus par des algorithmes. Enfin, l'effervescence et la forte médiatisation que suscite cette innovation peuvent laisser croire dans des cercles pas assez aguerris à l'analyse des données que ces dernières vont parler intelligemment d'elles-mêmes, hors de toute expertise et de toute idée ou hypothèse de départ. Sur tous ces sujets, les différents éclairages apportés dans ce numéro nous confirment qu'on ne peut réduire le débat sur les mégadonnées à une querelle entre les Anciens et les Modernes, mais qu'il faut explorer les différentes voies de régulation pour que cette innovation soit acceptée par tous.

Depuis l'identification, à la fin des années 90, des enjeux techniques et économiques de l'augmentation exponentielle du volume des données enregistrées, le phénomène BigData est devenu une réalité. Le sujet des «mégadonnées», selon la terminologie suggérée par les autorités françaises, est passé d'une réflexion stratégique et de prospective dans le milieu de la recherche (en particulier des physiciens qui ont les premiers dû faire face à des données massives), de l'économie et des États à un sujet de réflexion collective qui concerne de nombreux corps de métiers et plus globalement la société toute entière.

1. Ce séminaire a été organisé par le groupe Statistique et Enjeux Publics de la SFdS, en la personne de Marion Selz et Françoise Dupont. Il a reçu le soutien de l'Ecole Nationale de la Statistique et de l'Analyse Economique – ENSAE, et du Centre d'accès sécurisé aux données – CASD – qui font partie du Groupement des Ecoles nationales d'Economie et Statistique – GENES. Le compte-rendu du séminaire, ainsi que de nombreuses vidéos et présentations, sont disponibles à l'adresse : http://www.sfds.asso.fr/385-Les_enjeux_ethiques_du_Big_Data_opportunites_et_risques. La coordination de ce dossier sur les mégadonnées est le fruit du travail commun de Marion Selz et Françoise Dupont avec le comité de rédaction de la revue.

Reposant sur l'analyse de données volumineuses pour créer de la connaissance et de la valeur économique, le phénomène est devenu progressivement tangible à travers un foisonnement de recherches et d'applications concrètes qui commencent à toucher la vie de tous les jours. Au départ limitées aux grands acteurs du commerce et du numérique, les applications pratiques touchent peu à peu les secteurs de l'assurance, de la santé, du tourisme, des communications, des transports, … Celles dont on parle le plus relèvent essentiellement de la sphère privée de l'économie et sont plutôt orientées vers le marketing, la détection de fraude et la maintenance. Arnaud Laroche nous en propose dans ce dossier une revue détaillée. De plus, ces applications sont présentées comme n'étant qu'un avant-goût de celles qui se préparent avec la révolution numérique. L'apparition des objets connectés pourrait toucher notre vie quotidienne de manière inédite : selon certaines prévisions, nous aurions 10 objets connectés en moyenne en 2020.

Ce phénomène qui accorde aux données une place centrale conduit différentes cultures professionnelles à se rapprocher davantage pour maîtriser le potentiel des nouvelles données disponibles, les nouvelles infrastructures, les problèmes de sécurité, les enjeux de protection de l'anonymat, les risques de dérive de démarches algorithmiques pures. Les statisticiens, les mathématiciens et les informaticiens sont au cœur du mouvement mais les juristes sont aussi concernés pour les questions de propriété intellectuelle et de protection de l'anonymat. En réalité, presque toutes les disciplines sont concernées que ce soit dans la recherche ou au delà (physiciens, biologistes, médecins, informaticiens, mathématiciens, statisticiens, économistes et commerciaux, sociologues). D'autres corps de métiers seront touchés.

Les données sont «le nouvel or noir» : 90% des données récoltées depuis le début de l'humanité ont été générées au cours des deux dernières années. En 2011 on générait 5 exaoctets (5 milliards de gigaoctets) en deux jours ; en 2013 cette quantité était générée en 10 minutes. De quelles données parle-t- on ? Les mégadonnées sont d'abord des données émanant de l'utilisation d'internet au sens large et des communications : 247 milliards d'emails chaque jour, 133 millions de blogs, le trafic internet pourrait remplir environ 7 milliards de DVD. Plus généralement les données sont l'accumulation de traces laissées par des capteurs de toute sortes qui proviennent d'une activité via le web (consultations de sites, moteurs de recherches, discussions et commentaires sur les réseaux sociaux, espaces de stockage en ligne, outils collaboratifs en ligne …..). Elles peuvent également provenir des systèmes de gestion interne d'entreprises de secteurs aussi différents que la banque, les télécommunications, l'énergie, la logistique ou le transport, des capteurs mesurant température et pression dans des process de fabrication, etc. A toutes ces données qui sont déjà dans le paysage viendront s'ajouter les données issues des objets connectés (80 milliards de produits connectés prévus pour 2020).

Ces données qui sont caractérisées en premier lieu par leur volume sont également de natures très variées : données numériques, texte, photos, son, vidéos ; enfin elles peuvent être un mélange de ces différents formats (si l'on pense par exemple aux données produites par un examen médical). Avec les progrès de la recherche sur les algorithmes d'analyse et sur les infrastructures qui stockent et véhiculent les données, tous ces formats sont devenus exploitables dans des délais performants voire instantanément. Au delà du discours maintenant largement diffusé autour des « 3 V » (volume, vitesse, variété) on ajoute parfois d'autres V : en particulier V pour véracité qui indique que toute donnée transporte des informations qui sont fondamentalement conditionnées par le contexte de leur recueil et par toutes les conventions qui y ont présidé. Elles ne peuvent, comme nous le rappelle Antoinette Rouvroy dans ce dossier, être utilisées comme si elles représentaient la totalité d'une vérité appréhendée hors de tout contexte.

Ces nouveaux traitements présentent également la caractéristique d'être plus exploratoires. L'effet de mode est tel que certains vont jusqu'à annoncer un nouveau paradigme où les traitements permettraient, un peu comme par magie, sans réflexion préalable et sans expertise, de faire parler les données de façon pertinente en rendant caduques les démarches statistiques classiques. Arnaud Laroche nous rappelle que l'analyse des données a été présentée à ses débuts comme en rupture totale avec ce qui préexistait : il n'en était rien, et c'est la même chose aujourd'hui. Il y a une certaine continuité entre les méthodes statistiques classiques et les méthodes nouvelles adaptées aux Big Data.

A qui appartiennent les mégadonnées? Certaines portent quasi exclusivement sur des systèmes physiques gérés par des entreprises sans lien avec des comportements humains, mais 70% de ces données sont générées par des individus et sont porteuses d'informations personnelles. En raison de l'architecture informatique adoptée par internet, 80% de ces données sont stockées de façon centralisée par les entreprises. Benjamin Nguyen nous indique dans son article que l'anonymisation à 100% des données personnelles n'est pas possible, et que pour limiter les risques, une décentralisation de leur stockage est une bonne piste de protection.

Le stockage et l'utilisation de données qui sont en grande partie, mais pas uniquement, des données personnelles permettant d'identifier leur émetteur et contenant des informations sensibles soulève des questions juridiques et éthiques très délicates. Philippe Aigrain pense qu'une partie de la solution à ces problèmes repose sur le développement de modèles économiques et d'architectures alternatives à ceux qui sont actuellement proposés : il s'en explique dans l'interview qu'il nous a donnée.

Ces données représentent également un enjeu de propriété stratégique sur le plan économique pour les entreprises qui les détiennent (en particulier les « GAFA » : Google, Apple, Facebook, Amazon). La révélation en juin 2013 du dispositif de surveillance Prism de la NSA a créé un électrochoc, une partie importante des données étant stockées par ces entreprises soumises à la législation américaine et en particulier au Patriot Act. Les États européens et les entreprises européennes se mobilisent sur les enjeux de souveraineté et de protection des données.

Le potentiel de recherche et de développement économique (8% du PIB européen en 2020 selon le cabinet « Boston consulting group ») et en particulier de création d'emploi, crée un grand enthousiasme des milieux économiques autour du phénomène BigData. Nous sommes au début de ce phénomène et donc dans l'enthousiasme qui provoque une véritable bulle médiatique. Les retombées économiques sont déjà palpables : en 2014, le chiffre d'affaires dégagé grâce à ces technologies est estimé à 2.9 milliards de dollars (2,3 Mds en 2013), soit environ 2,2 milliards d'euros, pour la zone Europe de l'Ouest. En 2018, ce marché pourrait représenter 6,9 milliards de dollars d'investissements (estimation International Data Corporation - IDC). Les espoirs fondés sur ces nouvelles techniques ne relèvent pas uniquement de la sphère marchande. Des avancées sont également attendues dans la recherche, en particulier du côté médical, mais aussi en génétique, en astronomie… Les pouvoirs publics ont ainsi placé ce thème au cœur de leur réflexion stratégique en en faisant en 2013 une des sept priorités de la France pour 2030 et en lançant un plan stratégique dont François Bourdoncle a assuré le copilotage avec Paul Hermelin (Capgemini). Dans sa contribution, François Bourdoncle nous met en garde contre tout attentisme dans cette véritable révolution technologique. Il souligne le risque de perte de pouvoir économique liée à une culture législative européenne plus protectrice que la loi américaine.

Les questions juridiques et éthiques que soulève cette révolution technologique font l'objet de réflexions, mais également de pressions autour des initiatives législatives. Les autorités européennes de protection des données réunies au sein du groupe dit « de l'article 29 »[2] portent la réflexion sur la protection des données et le respect de la vie privée. Le 8 décembre dernier elles réaffirmaient les valeurs communes de l'Europe et proposaient des actions concrètes pour élaborer un cadre éthique européen. Elles affirment que la protection des données personnelles est un droit fondamental et que celles-ci ne peuvent être traitées comme un pur objet de commerce. Le règlement européen en discussion depuis déjà plusieurs années pourrait être finalisé en 2015. Sophie Vulliet-Tavernier (CNIL) aborde les réflexions en cours à la CNIL. Elle précise dans sa contribution le contexte de tension entre la protection des droits des individus et la possibilité d'innover dans lequel la CNIL opère en France.

Côté Français, dans le contexte du développement de l'économie numérique, le Conseil d'État a diffusé en septembre dernier une étude sur la protection des droits fondamentaux dans la société numérique. Cette analyse réalisée avec l'aide de nombreux professionnels venus d'horizon variés (économistes, juristes, sociologues,...) témoigne de la diversité et de la complexité des questions à résoudre : propriété intellectuelle, droit à la protection des données pour les individus, droit à l'oubli. Elle formule cinquante propositions pour faire évoluer le droit dans le cadre de la préparation de la loi sur le numérique qui sera débattue en 2015. Parmi celles ci, on y trouve le renforcement de la CNIL(comme autorité de protection des données européennes), la définition d'un droit des algorithmes prédictifs, des obligations des plateformes envers leurs utilisateurs en vertu du principe de loyauté, la possibilité de définir une action collective destinée à faire cesser les violations de la législation sur les données personnelles.

Les tensions entre les nécessités économiques et les exigences de protection des droits des individus sont très fortes. Quel est le chemin qui réconcilie les bénéfices de l'innovation et la préservation des libertés individuelles? Depuis peu en France, le débat commence à atteindre le grand public à travers différents articles qui s'interrogent sur la portée du phénomène pour le citoyen. Avec l'apparition des objets connectés, la préparation de la loi sur le numérique en 2015 pourrait être l'occasion d'une réelle appropriation des enjeux et des risques de ces innovations par l'ensemble de la société, avec un débat ouvert et large sur les modes de régulation de ces nouveaux systèmes. Des approches différentes sont proposées par les contributeurs à ce numéro. Le marché de la publicité digitale que présente Nicolas Grislain préfigure sans doute les évolutions d'autres secteurs. Il propose une vision des risques d'exploitation abusive des données, mais présente aussi les forces venant limiter ce risque. Alain Godinot soumet un regard de citoyen, et des pistes de régulation qui reposent sur l'éducation des plus jeunes, des dispositions juridiques et des codes éthiques professionnels et une initiative citoyenne. Philippe Aigrain, on l'a vu, défend l'idée qu'il faut surtout travailler sur des architectures informatiques alternatives. Antoinette Rouvroy, enfin, suggère de mettre en place un contrôle public des algorithmes et des systèmes de décisions et d'introduire dans la gestion de l'innovation des protections de ce qui fait la singularité des individus à travers des codes d'éthique.

Ce dossier le prouve abondamment : le débat ne se situe pas entre des « passéistes », qui refuseraient l'innovation, et des « modernistes » qui l'adopteraient sans réflexion. Toutes les contributions ici rassemblées témoignent d'une conscience vive des potentialités que recèlent les BigData pour le progrès de la connaissance et des applications. Toutes évoquent aussi les risques qu'un développement incontrôlé pourrait faire courir aux individus ou aux sociétés. La recherche de solutions pour parer ces risques est une préoccupation largement partagée.

2. Le groupe de travail européen dit « Groupe Article 29 » est composé de représentants des autorités nationales chargées de la protection des données, du Contrôleur européen de la protection des données et de la Commission européenne. Son organisation et ses missions sont définies par les articles 29 et 30 de la directive 95/46/CE, dont il tire sa dénomination.

La « vague de fond » du BigData

Entretien avec Arnaud LAROCHE

Co-fondateur et dirigeant de Bluestone, société de conseil en Datascience

Ce qui caractérise l'ère du « BigData », ce n'est pas seulement la taille des fichiers de données, ou la vitesse à laquelle on doit les traiter : c'est surtout la place qu'on leur attribue dans les processus industriels et de services. Désormais, c'est « la donnée » qui pilote des applications et des décisions. Les cas d'usage se multiplient, du micro-ciblage de publicités au pilotage de réseaux d'énergie ou de transport. Des secteurs comme les télécommunications, l'assurance, la banque sont directement concernés ; et l'information économique et sociale publique sera elle aussi touchée. Ce vaste mouvement donne à réfléchir : quelle est la place des analyses statistiques traditionnelles par rapport aux algorithmes « data-driven » ? Comment donner une maîtrise au citoyen-consommateur sur ses propres informations ? Comment concevoir et mieux maîtriser les implications sociales de la généralisation des algorithmes ?

Statistique & Société : Pouvez vous présenter votre société « Bluestone » ?

Arnaud Laroche : C'est une société que nous avons créée en 1998 avec quelques amis issus de l'ENSAE[1] pour proposer un service d'exploitation de la donnée aux entreprises désireuses de mieux comprendre leur environnement, pour prévoir et agir. Avec le temps, et les changements technologiques, nous avons évolué dans notre orientation sur la façon de traiter les données, nous sommes passés d'une approche statistique « classique » fondée sur les modèles à une approche qui fait place aux démarches nouvelles comme le « machine learning » où l'on laisse davantage parler les données. Mais il s'agit toujours d'utiliser les données pour répondre aux grands enjeux de l'entreprise. Aujourd'hui Bluestone compte 120 « ,data-scientists ».

S&S : Quelle définition donnez-vous du « BigData » ?

AL : Le BigData, c'est une vague de fond qui résulte de quatre évolutions, ou révolutions. La première, c'est la digitalisation de notre monde, élément déclencheur : nous sommes désormais entourés de capteurs, nous laissons partout des traces informatiques. Ensuite, il y a la révolution technologique : le remplacement de gros « supercalculateurs » par une myriade de petites machines travaillant en parallèle, dont le nombre peut être augmenté ou réduit en fonction des besoins. Cette révolution technologique divise les investissements requis pour entrer dans ce domaine et les met à portée de petites sociétés innovantes comme la nôtre. En troisième lieu, vient le progrès de la science des algorithmes capables d'opérer sur de très

1. Ecole Nationale de la Statistique et de l'Administration Économique.

gros volumes de données de façon plus exploratoire, comme ceux qui relèvent du « machine learning », etc. Et enfin, last but not least, depuis quelques années, et surtout depuis un an ou deux, les dirigeants d'entreprises ont changé d'attitude vis-à-vis des données. C'est une sorte de révolution culturelle : autrefois, les données de l'entreprise étaient vues comme des sous-produits des activités de gestion, analysées par des équipes de « data-mining » dont l'influence dans l'entreprise était réduite ; aujourd'hui, on construit des applications, des services, des processus qui sont conduits par les données. Le « buzz médiatique » est à la fois cause et conséquence de cette sensibilisation du « management ».

S&S : A partir de quelle taille des données est-on dans le domaine du « BigData » ? Peut-on esquisser un ordre de grandeur ?

AL : Je ne m'y essaierais pas. Bien avant qu'on parle de « BigData », on traitait dans certains domaines (astrophysique, génomique) de grandes bases de données. Mais c'était avec des architectures informatiques centralisées. L'émergence d'outils et de technologies différentes, notamment « Open source » autour de l'écosystème « Hadoop », permet de faire plus de choses à moindre coût, tant du point de vue des quantités de données que du point de vue du temps : on peut désormais réagir en continu à des données évolutives. Et sur cette nouvelle base technique s'est développé un changement « culturel » de la relation aux données : on met les données à la racine des efforts de l'entreprise pour traiter des enjeux économiques, sociétaux, etc. C'est cet ensemble technologique et culturel qui caractérise le « BigData ».

S&S : Avec cette définition, qui fait vraiment du BigData en France ? Il y a beaucoup d'entreprises qui en sont là ?

AL : Aujourd'hui, on met cette estampille partout, y compris sur des applications traditionnelles. Mais il y a des secteurs où déjà de vrais projets « BigData » au sens où je viens de le définir sont en place : télécommunications, assurances, un peu dans les banques. Dans beaucoup d'autres secteurs, on passe actuellement d'une phase d'observation à une phase d'industrialisation, et la demande de compétences sur les technologies et les outils BigData est en pleine explosion.

S&S : Pouvez-vous donner des exemples d'applications ?

AL : Les usages les plus connus relèvent de la personnalisation de la « relation-client ». Il s'agit d'utiliser les données internes de l'entreprise et des données web pour mieux gérer la relation client. On peut citer comme exemple le microciblage en temps réel du client potentiel réalisé en France par Critéo[2] pour offrir aux annonceurs une publicité sur internet à la performance personnalisée déclinable à l'échelle mondiale. Dans la même veine Netflix, récemment introduit en France a fondé son modèle économique sur des algorithmes qui recommandent à ses clients des films suceptibles de les intéresser. Les films ainsi proposés à l'abonné le sont sur la base d'algorithmes et de critères calculés sur des grandes masses. Le choix ainsi adapté au client semble pousser à la découverte, mais il présente également par construction le risque d'enfermer dans une mono culture. Les banques qui ont une relation client plutôt basée sur l'offre produit travaillent sur une plus grande personnalisation de l'offre en s'appuyant sur des données clients internes (transactions) croisées avec des données externes sur les moments de vie (recherches des clients sur le web via les cookies, parcours de recherches sur le web).
Ce sont les applications les plus médiatisées aujourd'hui. Mais il y en a beaucoup d'autres, moins visibles par le grand public.
On assiste depuis un an en France à une montée en puissance sur la « maintenance prédictive », en particulier dans l'industrie. En effet, d'une part l'utilisation intensive de capteurs se généralise dans les process de fabrication pour mieux contrôler les paramètres techniques (température

2. http://www.criteo.com/fr/what-we-do/technology/

pression…) et donc le pilotage, d'autre part les données sont stockées et analysées plus facilement et plus rapidement. Ces évolutions permettent de mettre en place des systèmes de détection des anomalies des systèmes basés sur les signaux faibles, qui ne peuvent pas être perçus dans des contrôles qualité de type échantillonnage. Ils permettent de détecter les problèmes plus en amont et de gagner en efficacité et en rapidité par rapport aux systèmes antérieurs basés sur les seuls experts des métiers concernés. On trouve de tels systèmes en France pour les forages de Total. General Electric[3, 4] propose une offre packagée aux entreprises industrielles. Les capteurs embarqués dans les avions permettent d'identifier les dysfonctionnements plus en amont pour mieux programmer les opérations de maintenance. L'afflux de données permet de mieux comprendre les enchainements temporels et de cerner plus précisément le lien entre les réparations effectuées et la résolution des dysfonctionnements constatés. Bien sûr, cette analyse ne peut se faire sans l'apport de l'expertise propre au métier. On pourrait trouver d'autres exemples de telles innovations dans la gestion des réseaux de transport, d'eau, d'électricité.

Les techniques d'analyses sont également mobilisées pour optimiser les chaines logistiques « supply chain management »: dimensionnement des entrepôts, optimisation des tournées de livraison, des capacités d'un parc de transport. Citons comme exemple l'initiative récente de Chronopost visant à diminuer ses délais de livraison[5, 6].
Les récurrences dans la configuration des données peuvent également être utilisées pour faire émerger des suspicions de fraude et cibler les contrôles (douanes, carte bleue, assurance chômage).
Les données comme celles de la téléphonie mobile, ou celles qui viennent de la gestion des réseaux de transport peuvent être mobilisées pour réaliser des analyses fines des flux afin de dimensionner les infrastructures, d'optimiser les transports. Elles apportent une amélioration par rapport aux enquêtes auprès des usagers qui sont coûteuses et ne peuvent être réalisées que parcimonieusement. Enfin, les compteurs électriques intelligents devraient permettre à terme de faire des diagnostics et des recommandations à distance.

S&S : Les bonnes conclusions, les bonnes décisions vont elles découler naturellement de la seule analyse des données ?

AL : On voit renaître aujourd'hui la vieille controverse entre les analyses « conduites par les données » (« data-driven ») et les analyses reposant sur des modèles. La statistique traditionnelle adopte la démarche hypothético-déductive, qui utilise des modèles, teste des hypothèses, cherche à comprendre ; le data mining cherche des corrélations sans hypothèses préalables, et cherche à prévoir. Vieille controverse : elle était déjà vive en France lors du renouveau de l'analyse factorielle vers 1970. Aujourd'hui certains de mes confrères disent : « avec le BigData, plus besoin d'être intelligent », « du moment que ça marche, c'est bon ». Cela me semble totalement à l'opposé de la réalité. Certes, il y a des contextes dans lesquels l'efficacité à court terme prime, sans qu'on ait besoin de savoir « pourquoi ça marche » : pour faire en temps réel les meilleures propositions commerciales, pour détecter le plus vite possible les pannes, on peut concevoir des algorithmes « data-driven ». Mais en même temps, s'engager dans une telle voie demande un surcroît d'intelligence, car il faut apprendre à contrôler ces algorithmes pour en maitriser la pérennité. Fonctionnent-ils convenablement ? Fonctionnent-ils de façon stable ou dérivent-ils dans la durée ? Pour le savoir, il faut les soumettre à un véritable « monitoring » dans lequel on doit s'interroger sur l'interprétation des phénomènes, sur la valeur structurelle des modèles, et faire appel à des techniques d'expériences contrôlées comme les sondages. Pour moi, les deux écoles devraient se répondre plutôt que de s'opposer.

3. http://www.ge-ip.com/ii/industrial-internet
4. http://www.ge-ip.com/products/rtoi/c564
5. http://www.decideo.fr/Chronopost-peaufine-ses-delais-de-livraison-avec-le-Data-Science-Studio-de-Dataiku_a7501.html
6. https://evenement.inter.laposte.fr/labpostal/images/conference/data_daitaku.pdf

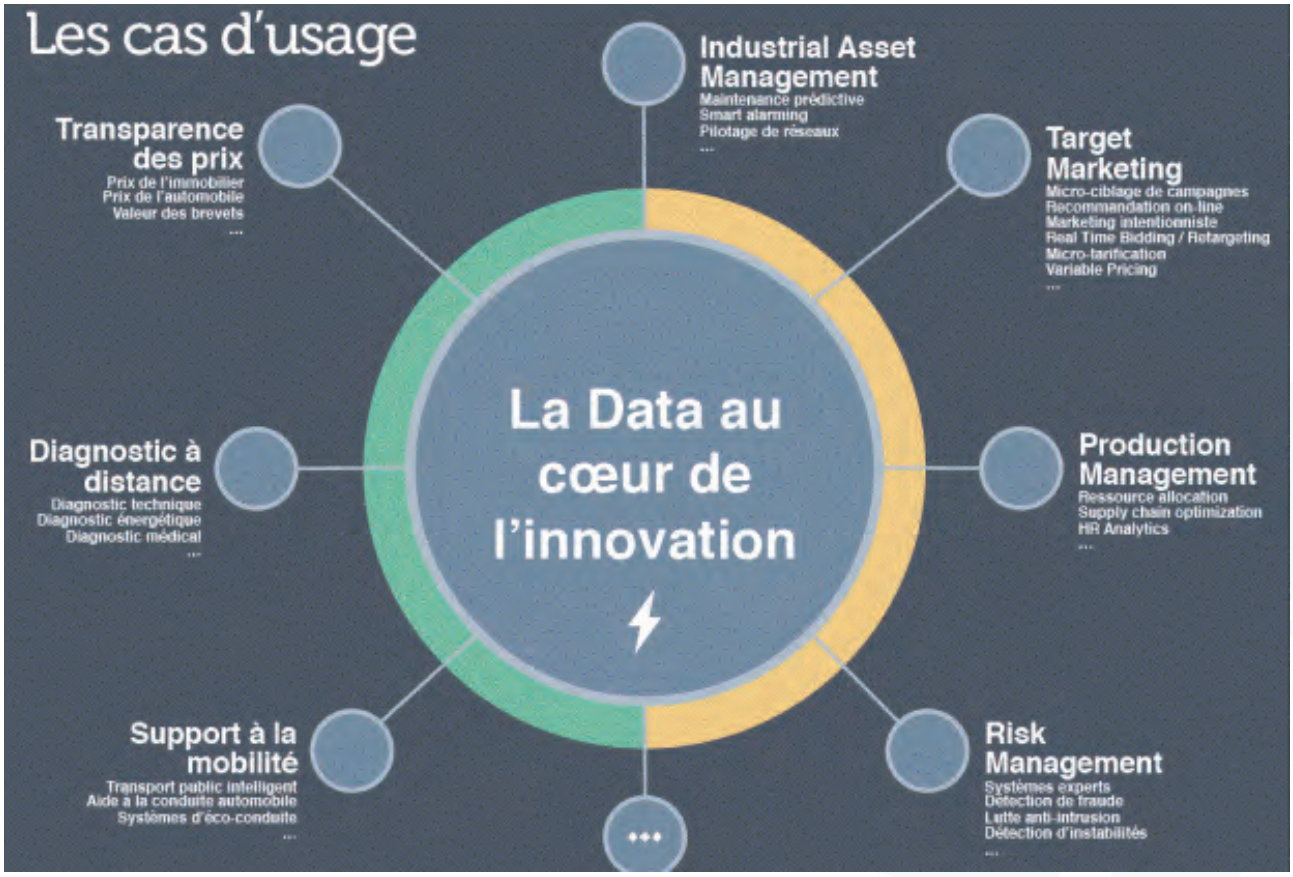


Figure 1 : Sept cas d'usage des « BigData »

S&S : Faut-il pour caractériser le BigData faire des distinctions entre les différents types de données ? Les données structurées, les « traces » qu'on laisse sur Internet, les données textuelles, etc. ?

AL : Il y a bien entendu de grandes différences du point de vue des techniques de traitement des données. Par exemple, lorsqu'on a affaire à des données issues du recueil de formulaires administratifs, les matrices « observations-variables » sont bien remplies, il y a en général peu de valeurs manquantes ; alors que, lorsqu'on utilise les traces laissées sur Internet par les consommateurs pour concevoir un système de recommandation de produits, la matrice « individus x produits » est très « creuse », et cela appelle des techniques de traitement particulières. En matière de traitement des textes, c'est pareil : le « text mining » existe depuis longtemps, mais le traitement des textes spontanés, récupérés sur des forums par exemple, pose des difficultés nouvelles. Il y a un foisonnement de recherches là-dessus pour mettre au point des algorithmes adaptés. Cela dit, ces différences ne me semblent pas être au cœur de la caractérisation du phénomène « BigData ».

S&S : : La statistique publique est-elle menacée par l'émergence du « BigData » ?

AL : Qu'est-ce qu'on attend de la statistique publique ? Qu'elle produise des chiffres sûrs, selon des méthodologies éprouvées, en respectant des principes clairs. Personne ne va s'amuser à lui faire de l'ombre sur ce terrain. La contrepartie est un certain manque d'agilité. Si des initiatives issues du BigData peuvent lui porter tort, c'est dans un domaine bien particulier, celui de la création d'indicateurs économiques avancés à partir de données captées « dans la vraie vie » à

partir de données qui n'ont pas fait l'objet d'un plan de recueil préalable. Un exemple, dans le domaine de l'immobilier, des prix des logements. La statistique publique s'appuie sur les bases des notaires et pour diverses raisons ne peut pas avoir moins de deux mois de retard par rapport à l'évènement. Un indice comme celui de « Meilleurs agents », fondé sur la captation de offres d'agences immobilières partenaires, paraît beaucoup plus tôt, et les 2/3 des articles de la presse spécialisée le citent. En ce cas précis, ce n'est pas du BigData ; mais cela pourrait le devenir. Selon moi, la statistique publique aurait tort de balayer d'un revers de la main ce genre de démarche en disant simplement « ce n'est pas propre » : elle devrait plutôt chercher à innover en traitant ce type de problème – utiliser des données tout venant pour calculer des indicateurs avancés fiables – avec le regard des statisticiens publics. L'exemple de « Google Flu » qui a prédit à tort une épidémie de grippe à New-York il y a quelques années, avec des conséquences fâcheuses pour l'action publique, montre qu'il y a là un réel besoin.

S&S : Venons-en maintenant aux risques des BigData pour les droits des individus. Comment les caractérisez-vous ?

AL : Tout d'abord, il faut savoir mettre en regard les nouveaux services offerts et les dangers réellement encourus. Le dévoilement de données personnelles à des tiers est beaucoup plus fréquent qu'il y a vingt ans : on délivre de l'information sur soi à beaucoup de gens, sans savoir toujours qui ils sont, où ils sont, et sans maîtriser ce qu'ils peuvent en faire et avec qui ils peuvent la partager. On le fait généralement en échange de services qui ne sont accessibles que si on a dit « oui » ! Et l'on désire obtenir ces services. Aussi, il serait vain de s'opposer à une telle déferlante. Je suis sceptique sur la capacité de résister, et je trouve qu'ériger la protection des données personnelles en un principe sacro-saint est une démarche vaine. Mais on peut s'efforcer de rendre de la maîtrise au citoyen-consommateur.

S&S : Comment ?

AL : J'ai un point de vue libéral, qui repose sur l'idéal d'un contrat informé entre l'individu qui veut avoir accès à de plus grands services et les fournisseurs qui ont besoin des données des individus pour développer ces services. Un tel contrat suppose qu'on donne à chaque individu un moyen simple de savoir quelle information sur lui-même il livre, à qui, et pour quoi faire. En particulier, il doit pouvoir contrôler les utilisations en cascade de ses données : à qui seront-elles transférées, ou vendues, et pour quels usages. Actuellement on est perdu : personne ne sait ce qu'il a lui-même autorisé. Il y a un manque de conscience de l'information qu'on laisse, et d'éducation sur les enjeux que cela comporte. Techniquement, un meilleur contrôle est possible : se développent actuellement des outils de « VRM[7] » permettant aux clients d'avoir accès à l'information détenue par les entreprises sur eux-mêmes, et leur donnant une certaine maîtrise sur ces contenus, par un renversement de la logique de « CRM[8] » dans laquelle les fournisseurs de services « managent » leurs clients.

S&S : N'est-il pas impossible de définir à l'avance tous les usages possibles des données ?

AL : On pourrait définir des catégories d'usages, que l'individu pourrait autoriser ou non, selon qu'il souhaiterait ou non accéder à des services plus étendus. Dès à présent on peut acheter de la donnée « Twitter », et il existe des modalités prévues pour cela dans Facebook. Il y a un modèle économique qui se construit autour de la réutilisation des données.

7. « Vendor relationship management » voir http://data-tuesday.com/2013/10/22/decouvrez-les-presentations-de-la-data-tuesday-vrm-8-octobre-2013/
8. « Customer relationship management »

S&S : L'émergence du BigData est-elle porteuse d'autres risques, cette fois au niveau de la société toute entière ?

AL : Se pose la question d'une éventuelle « sur-mathématisation du monde », c'est-à-dire du nombre de plus en plus grand des décisions prises par des machines. Qu'il s'agisse de finance, de décisions concernant des personnes, des interactions sociales, l'invasion des algorithmes n'est pas un mythe : une société vient même de faire entrer un robot dans son conseil d'administration ! Stephen Hawking, cosmologiste, alerte sur les dangers de l'intelligence artificielle et sur une possible perte de contrôle de l'homme sur la machine.

S&S : Vraiment ?

AL : Je pense à l'exemple du trading à haute fréquence : à un certain degré, les effets d'ensemble deviennent incontrôlables. Les modèles de scoring peuvent écarter des pans entiers de la population. La publicité sur Internet devient un marché financier, dans lequel les produits visibles sont déterminés par des algorithmes conçus pour montrer à chacun ce qui est censé lui convenir : cela entraîne un « normage de la société » où tout le monde est informé exclusivement selon sa place à l'intérieur d'une segmentation a priori. Encore une fois c'est une lame de fond, un mouvement qui se fait de toute façon ; mais un mouvement qu'il est souhaitable de contrôler par des démarches qui, elles, ne peuvent pas relever de l'algorithmique. Il faut y réfléchir, et mêler à cette réflexion des gens qui ne soient ni dans l'adhésion complète, par exemple du fait de leur implication professionnelle, ni dans l'opposition systématique qui conduit à ne penser qu'en termes de réglementation. De toutes façons, le BigData change notre monde : il faut y faire face hors des lobbies et des logiques doctrinaires.

Le BigData et la publicité en temps réel

Nicolas GRISLAIN
Co-fondateur de la société AlephD[1]

Lorsque l'utilisateur d'un site Internet affiche une nouvelle page, sur laquelle sont réservés des espaces publicitaires, en quelques millisecondes des algorithmes déterminent quelle publicité sera affichée, et quel prix sera payé pour cela par l'annonceur. Depuis quelques années déjà, il existe des algorithmes qui choisissent pour l'annonceur le contenu précis à afficher, en fonction des caractéristiques de l'utilisateur. Désormais les espaces publicitaires sont attribués à l'issue d'une enchère, mettant en concurrence les annonceurs. D'autres algorithmes déterminent, soit pour l'annonceur soit pour l'éditeur du site, le montant à enchérir. Pour construire une règle de décision optimale, constamment évolutive, il leur faut mémoriser et analyser d'énormes quantités de données issues d'enchères précédentes. Les techniques du BigData sont indispensables. Les données traitées sont pour partie des données personnelles, mais elles ne sont utilisées que par des machines, dans le cadre d'une problématique précise qui ne permet pas d'avoir une idée globale des individus.

Sur fond d'automatisation du marketing digital, la société AlephD s'est développée en proposant aux éditeurs (sites web, applications mobiles) des outils permettant d'augmenter leurs revenus en optimisant la mise en vente de leurs espaces publicitaires. Les stratégies d'optimisation reposent sur l'exploitation de grands volumes de données : *BigData*. Ces approches puissantes permettent d'adapter son action à chaque utilisateur. En tant qu'acteur d'un secteur fortement consommateur de données personnelles, AlephD a un point de vue original sur le traitement par le marché de ces données.

Émergence des enchères publicitaires en temps réel

Le développement d'internet et la nécessité de financer la création de contenus ont, dès le milieu des années 90, contribué à la forte croissance du marché de la publicité en ligne et notamment à l'utilisation de bannières publicitaires (*display ads*).
Les bannières publicitaires, initialement intégrées de manière statique dans les pages web, l'ont rapidement été par l'intermédiaire d'*ad-servers*, permettant de programmer la diffusion de campagnes différentes selon l'utilisateur, l'heure de la journée ou tout autre critère. Les ad-servers permettent d'adapter la vente des bannières à chaque « impression », c'est-à-dire à

1. La société AlephD offre des services de traitement de données aux éditeurs de sites Internet

chaque visite d'un espace publicitaire par un utilisateur à un instant donné.
Néanmoins, la mise en relation entre le site web ou éditeur, c'est-à-dire l'entité vendant les impressions, et les annonceurs, qui achètent les espaces, reste en grande partie manuelle. En raison notamment du coût relativement élevé de l'allocation manuelle des *impressions*, une part importante de ces *impressions* reste invendue. Dans un premier temps, se sont mis en place des montages complexes de cascades entre intermédiaires, les uns passant la main aux autres s'ils ne peuvent livrer de bannière.
En 2008 pour automatiser la négociation des ventes d'*impressions* et éviter les cascades de passation de relais entre intermédiaires qui ralentissent le processus de vente, l'*ad-exchange* (bourse d'échange d'impressions publicitaires) est créée.

Fonctionnement d'un ad-exchange

Sur un *ad-exchange*, chaque impression publicitaire est mise en vente aux enchères pendant le chargement du contenu d'une page web, c'est-à-dire en quelques millisecondes : on parle de *Real-Time Bidding* (RTB). Ce processus se déroule en 3 temps :

- Un utilisateur charge la page d'un site site.com. La balise (*tag*) HTML définissant l'espace publicitaire de la page chargée émet une requête (*ad-call*) à l'*ad-exchange*.
- L'*ad-exchange* reconnaît éventuellement l'utilisateur et l'identifie par un *user id* ; il envoie un appel à enchérir (*bid-request*) à tous les acheteurs potentiels. Cette requête contient l'identifiant du placement, l'identifiant de l'utilisateur ainsi que d'autres informations associées au placement, à l'utilisateur, à l'enchère ou au protocole (adresse IP, user-agent, *referer*).
- Les acheteurs (*bidders*) répondent en fournissant leur évaluation de l'impression (*bid*). L'*ad-exchange* attribue l'espace publicitaire au plus offrant. Ce dernier paye généralement le plus haut bid en dessous du sien (second *bid*) et délivre le contenu de la publicité à l'utilisateur.

Ce processus se déroule en près de 50 ms et aboutit à l'affichage de la bannière du gagnant sur *site.com*. Les contraintes de temps réel font que les stratégies des acheteurs sont mises en œuvre par des algorithmes qui évaluent l'opportunité d'afficher une bannière à chaque impression.
L'enchère simultanée au second prix (enchère de Vickrey) est le mécanisme d'attribution dominant du marché ; il est utilisé sur la quasi-totalité des plates-formes. Dans ce cadre, le prix de réserve[2] que fixe le vendeur a un impact sur le prix de vente. Il appartient au vendeur de fixer ce *prix de réserve*, ainsi que d'autres paramètres de l'enchère, comme le niveau d'information diffusé aux acheteurs.

Analyse micro-économique de la relation acheteur vendeur sur les plates-formes d'enchères

De fait, face aux algorithmes parfois sophistiqués des acheteurs, les éditeurs réalisaient jusqu'à récemment le paramétrage de leurs enchères de manière statique. Dans ce contexte, l'avènement des *ad-exchanges* et du RTB a conduit les acheteurs à développer des stratégies d'enchère élaborées les positionnant comme *faiseurs de prix*. Au contraire, les vendeurs (éditeurs) qui n'ont pas développé les capacités technologiques pour analyser chaque enchère individuellement ni pour agir sur celles-ci se retrouvent dans la situation inconfortable de *preneurs de prix*.

2. C'est-à-dire le prix minimum demandé par l'éditeur du site

AlephD et le *big data* au service des éditeurs

Sur la base de ce constat, la société AlephD (www.alephd.com), créée fin 2012, a construit des outils d'analyse et de prise de décision en temps réel pour permettre aux éditeurs d'optimiser leurs revenus publicitaires enchère par enchère, c'est-à-dire des milliers de fois par seconde en moins de 10 ms.
Pour réaliser cette tâche, AlephD enregistre des rapports d'enchère des milliers de fois par seconde. Ces rapports contiennent a minima un identifiant d'utilisateur chargeant la publicité, un identifiant de placement publicitaire ainsi que des données de prix : premier prix et prix payé. Ces informations sont d'une part traitées en flux (*online processing*), et d'autre part stockées pour des traitements en gros (*batch processing*).
Elles représentent des volumes importants (plusieurs téraoctets de données chaque mois) ; elles arrivent avec un débit élevé (*high velocity*) et sous des formes relativement variées. Les données analysées par AlephD correspondent donc assez bien à la définition donnée par *Gartner* du big data comme coïncidence de 3Vs : *volume, velocity, variety*. En particulier, le volume de données et leur débit ne permettent pas de traiter ces données sur une seule unité de traitement, même en considérant des ordinateurs très haut de gamme.

Traitement en gros des données

Pour construire une règle de décision optimale, AlephD estime un modèle d'apprentissage automatique (*machine learning*) dont l'espace des paramètres est de très grande dimension afin de prendre en compte les relations complexes entre variables. Ce modèle de décision donne un paramétrage optimal de chaque enchère sur la base d'agrégats historiques (ensemble de grandeurs caractéristiques d'une entité) construits à la volée.
Compte tenu du volume des données, ce calcul est réparti sur différents serveurs qui réalisent chacun une part de l'estimation. Afin de gérer la complexité liée à la distribution du calcul et à l'agrégation des résultats donnés par chaque serveur, le paradigme de calcul *map-reduce* est utilisé pour exprimer le processus d'estimation.
L'approche *map-reduce* est une manière de formaliser une opération sur un grand nombre d'observations qui consiste à appliquer à chaque observation une transformation (*map*) puis à agréger les résultats des étapes *map* par une fonction d'agrégation (*reduce*). L'intérêt de cette approche est qu'un calcul exprimé dans ce formalisme est assuré de pouvoir être distribué sur un grand nombre de serveurs et donc de passer à l'échelle. Cette garantie permet à AlephD de pouvoir traiter un volume de données 10 fois plus important en multipliant le nombre de serveurs effectuant l'estimation par 10.

Traitement en ligne

Des milliers de fois par seconde, AlephD reçoit un compte-rendu d'enchère qu'elle utilise pour dresser des agrégats historiques (ensemble de grandeurs caractéristiques de l'historique d'une entité) pour chaque utilisateur.
Ce traitement est réalisé de manière distribuée par un ensemble de serveurs traitant chacun une partie du flux de données et recombinant les résultats de manière adéquate.
Ce traitement en flux permet de tenir compte en temps réel de la dernière information disponible et de pouvoir optimiser les revenus de l'éditeur de manière très réactive.

Nécessité du *BigData*

Historiquement, le réflexe du statisticien face à un jeu de données de grande taille est d'extraire un échantillon de données, c'est-à-dire de ne conserver qu'une fraction de ses observations sélectionnées aléatoirement. C'est ce que font les instituts statistiques pour calculer des

agrégats macroéconomiques ou ce que font les instituts de sondage. Cette approche est valide tant que l'on ne cherche qu'à calculer un nombre limité d'agrégats.
La problématique d'AlephD est au contraire de pouvoir quantifier un nombre très important de paramètres. En effet, l'introduction du RTB a permis aux acheteurs de pouvoir fixer un prix et acheter chaque *impression* individuellement, c'est-à-dire pour chaque utilisateur sur chaque placement. Une stratégie de vente efficace pour l'éditeur ne peut se concevoir qu'à l'échelle de l'impression et il devient nécessaire de modéliser le profil en terme d'historique d'enchère de chaque utilisateur pris individuellement.
Ce passage d'un monde où l'on calcule quelques agrégats statistiques à un monde où l'on cherche à dresser un portrait individuel de chaque utilisateur nécessite d'abandonner l'échantillonnage et de traiter les données exhaustivement.
En outre, la valeur de certaines connaissances peut décroître très rapidement dans le temps. Dans ce cas le traitement en temps réel des données peut devenir nécessaire. C'est le cas de certaines grandeurs traitées par AlephD.
Cette nécessité de prendre des décisions personnalisées pour un grand nombre d'entités sur la base de données à péremption rapide est la raison essentielle d'une approche de type *BigData*.

Traitement des données personnelles par le marché

En tant qu'acteur d'un secteur fortement consommateur de données individuelles, AlephD dispose d'un observatoire privilégié sur le traitement par le marché de ces données.
Il est clair que la numérisation croissante des activités humaines et la possibilité donnée par les outils du *BigData* mettent à mal l'anonymat en ligne et fragilisent les tentatives de protection de la vie privée. Plusieurs points sont cependant à noter :

- Dans de nombreux cas, l'exploitation commerciale de données à l'échelle individuelle se fait dans le cadre d'une problématique bien précise qui ne permet pas d'avoir une idée globale des individus ni de retrouver leur identité. Certaines entreprises fournissent, par exemple, des informations sur la probabilité qu'un utilisateur soit un robot. De nombreux robots sont conçus pour augmenter artificiellement le nombre des visites ou des clics. Identifier ces robots nécessite de modéliser ce qu'est un comportement probable d'utilisateur humain et de qualifier chaque utilisateur individuellement. Dans cet exemple, l'entreprise n'est jamais amenée à constituer un profil complet d'un individu : elle ne menace pas la vie privée des utilisateurs.
- Dans d'autre cas les statistiques constituées à l'échelle des individus ne sont utilisées que par des machines. C'est le cas d'entreprises telles que Criteo où des modèles mathématiques sont conçus pour évaluer au mieux la probabilité qu'un utilisateur clique sur une bannière. Dans ce cas toute la chaîne est automatisée et l'information non pertinente pour réaliser l'objectif fixé est naturellement délaissée. Là encore, même s'il est très facile pour Criteo de prédire votre propension à cliquer sur telle ou telle publicité sachant que vous avez cliqué sur telle page d'un site d'e-commerce, il lui est très difficile de reconstruire le profil détaillé d'un individu en particulier.
- À côté de cela, l'utilisateur prend conscience de la valeur de ses données personnelles, et il apprend à gérer la diffusion de ces données. Plus le temps passe et plus il peut décider d'échanger un accès aux informations le concernant contre un contenu gratuit qu'il juge de qualité suffisante ; il peut également décider de payer un service contre l'assurance de la protection de ses informations personnelles. Certains acteurs mettent en avant cet aspect comme un argument commercial et implémentent de réels dispositifs de protection des données (on peut citer le chiffrement des données *cloud* par Apple).

En tant que précurseur de l'exploitation en masse des données personnelles le marché de la publicité digitale préfigure sans doute les évolutions d'autres secteurs. Il illustre les risques d'exploitation abusive des données, mais aussi les forces venant limiter ce risque.

Ne manquons pas la révolution industrielle du BigData !

François BOURDONCLE
Co-fondateur de la société Exalead[1], co-chef de file[2] du plan BigData français

L'irruption du BigData n'est pas seulement une révolution technologique. C'est surtout une modification profonde des rapports économiques entre les entreprises dans de très nombreux secteurs, et il n'est pas exagéré de parler de révolution industrielle. La lutte pour les marchés et les profits se joue aujourd'hui avec des armes nouvelles : la connaissance de caractéristiques des consommateurs ou des clients, et la capacité d'exploiter ces caractéristiques pour s'imposer aux acteurs traditionnels. Dans cette véritable guerre, les grandes entreprises mondiales nées à la fin des années 1990 et au début des années 2000 ont pris des positions fortes. Mais beaucoup d'innovations sont encore à venir, et les pouvoirs publics veulent encourager les initiatives des entreprises françaises. L'indispensable protection des individus et de la société contre les risques que comportent ces innovations doit être recherchée dans des modalités nouvelles d'application de la loi, qui peuvent constituer des avantages comparatifs au plan international.

Trois cycles d'innovation technologique

Nous connaissons actuellement le début de la troisième vague d'innovation due aux technologies de l'information. Dans les années 1980, on a vécu l'informatisation des entreprises et de leurs processus : c'était l'ère des sociétés comme SAP, ORACLE ou Microsoft, et des usages internes de l'informatique dans les entreprises. A partir de 1995 et au début des années 2000, les technologies de l'information ont pénétré le grand public : c'est l'époque de la naissance de Google, d'Amazon, de Facebook ; c'est la généralisation du téléphone mobile et des réseaux sociaux, tous secteurs fonctionnant sur le modèle économique de la publicité. Les technologies informatiques ont progressé en parallèle : le « transactionnel » et la mise à jour des données ont fait place aux moteurs de recherche et à un nouvel impératif technique : optimiser l'accès aux données et leur utilisation par des applications dont la facilité d'usage était devenue un critère de qualité déterminant. A l'issue de cette deuxième vague d'innovation, les géants qui en sont issus ont désormais une avance colossale pour obtenir et croiser les données, de façon à proposer des algorithmes prédictifs sur lesquels sont fondés des services, mais également pour la capacité à traiter des volumes de données, et ce en temps réel, comme cela n'a jamais

1. Filiale du groupe Dassault Systèmes pour les moteurs de recherche et les BigData
2. Avec Paul Hermelin, président directeur-général de CapGemini

été le cas dans l'histoire. Aujourd'hui, ces énormes sociétés ont besoin de relais de croissance : elles s'attaquent au monde physique, et c'est la troisième vague d'innovation. Prenons l'exemple de la cartographie. Avec Google Maps, Google a un projet de cartographie virtuelle du monde physique réel dans un sens très large, entrant à l'intérieur des bâtiments, des centres commerciaux, etc. Et là où Apple avait mis au travail 200 ingénieurs pendant deux ans pour concevoir un système qui s'est révélé finalement à côté de la plaque, Google emploie 5 000 cols bleus en Inde pour vérifier la numérisation des quantités d'information sur les équipements, les infrastructures, les panneaux de signalisation, etc. Il ne s'agit plus de « tertiaire », mais d'une véritable industrie connectée au monde réel.

Au-delà de la technique, une révolution industrielle…

Le BigData change le paysage concurrentiel dans de nombreux secteurs qui se croient protégés. Les nouveaux acteurs imposent leur rythme d'innovation à des entreprises industrielles traditionnelles qui pouvaient auparavant conserver leur profitabilité sans changer profondément leurs pratiques.
Car c'est en termes économiques qu'il faut analyser ce qui se passe, plus précisément en termes commerciaux. La bataille du commerce est vieille comme le monde : elle consiste à prendre les clients des concurrents. Elle se joue aujourd'hui avec des armes nouvelles, qui sont la connaissance des caractéristiques des consommateurs, et la capacité d'exploiter ces caractéristiques comportementales pour emporter les marchés. Ces armes, ce sont celles du BigData, que possèdent les grandes entreprises nées du deuxième cycle d'innovation technologique.
Et il faut bien comprendre que l'attaque va se porter sur les marges des entreprises traditionnelles. En coupant ces entreprises de l'information sur leur clientèle, les grands acteurs du BigData se mettent en capacité de les réduire à des rôles de sous-traitants, rémunérés uniquement pour leur technicité, c'est-à-dire mal rémunérés ! Le contexte est souvent un mouvement de « servicisation » des industries manufacturières traditionnelles, et de « réinternalisation » de beaucoup d'industries de services, comme on peut le voir sur des exemples du passé récent ou du futur proche.

De nombreux exemples existants…

Premier exemple : la distribution d'organes de presse par une plate-forme comme « l'Apple Store ». Non seulement cette plateforme demande aux entreprises de presse une marge élevée, de l'ordre de 30% ; mais le contrat qu'elle leur propose ne prévoit aucune rétrocession d'information sur leurs lecteurs. Les entreprises en question perdent la relation client, et ne sont plus rémunérées que pour leur technicité « d'écriture d'articles » : le pouvoir de décision passe ailleurs. On pourrait citer aussi ce que « iTunes » a fait à l'industrie du disque.
Qui maîtrise la relation-client dans le cas de la téléphonie mobile ? De plus en plus, ce sont les fabricants de terminaux intelligents, qui s'adressent directement à la clientèle qui dispose d'un haut revenu, laissant aux opérateurs traditionnels le soin de gérer les services, et aussi, il est vrai, de rivaliser commercialement, mais seulement en direction des petits utilisateurs qui ne constituent pas des cibles très intéressantes d'un point de vue commercial.
Troisième exemple, le commerce de proximité : Amazon s'impose comme nouvel intermédiaire entre les consommateurs et les producteurs ou les autres formes de distribution, de la même façon qu'en leur temps les Darty et autres s'étaient imposés comme intermédiaires. En utilisant ses capacités de recommandation par connaissance des comportements d'achat, Amazon prend une part importante de la vente directe, et ensuite peut s'associer à d'autres circuits de distribution, en leur imposant son propre service-clients… et ses marges ! Le bras de fer récent entre Amazon et le groupe Hachette en est un exemple.

Le cas du tourisme est également significatif : sur les 9 milliards d'euros de chiffres d'affaires procurés par les nuitées touristiques en France, 20% passe désormais par les plateformes de réservation en ligne, dont trois, Expedia, Bookings et Travel, représentent 90% du total. Ces plateformes prélèvent des marges de 23% en moyenne, alors que les intermédiaires traditionnels, les agences de voyage, prélevaient 10%. Elles ont un pouvoir de rétorsion considérable contre les hôteliers : elles peuvent dé-référencer des hôtels qui ne sont remplis que via ces centrales de réservation. Les accords commerciaux que ces plateformes passent avec les hôteliers ou les chaînes d'hôtels comportent parfois des clauses léonines : par exemple, toute ristourne consentie par un hôtelier à un client doit être signalée à la plateforme de façon à pouvoir être généralisée à tous les clients de celle-ci. Progressivement les hôteliers sont transformés en « gestionnaires des murs », coupés qu'ils sont de leur clientèle.
Dans tous ces exemples, le rapport avec les BigData est patent : c'est la capacité de maîtriser les grandes masses d'information sur la clientèle, et de les analyser pour fournir des services innovants, qui fournit à une société le pouvoir de s'introduire, puis de s'imposer sur un marché. Et ce n'est pas fini !

Et d'autres qui vont venir

On peut proposer quelques exercices d'anticipation : les choses ne se passeront peut-être pas précisément toutes comme cela, mais dans tous ces secteurs des prémisses sont déjà observables.
Voulez-vous payer moins cher votre assurance automobile ? Si Google faisait cette proposition à des internautes, en échange d'un accès à des données détaillées sur leur manière de conduire, données recueillies sans effort par des capteurs automatiques, que répondraient-ils ? Peut-être certains auraient-ils à cœur de protéger leur vie privée ; mais si cet opérateur était à même de proposer des tarifs concurrentiels aux bons conducteurs, combien résisteraient ? Du coup, les assureurs traditionnels seraient déséquilibrés, ne conservant que les risques les plus élevés. Il est probable que Google trouverait alors un sous-traitant pour gérer les sinistres, récupérant par là-même une information supplémentaire pour faire une analyse encore plus fine des risques, et maîtriser encore plus le marché. Peut-être les choses ne se passeront-elles pas exactement ainsi : mais les assureurs se préoccupent de cette possibilité, et ils ont raison.
Les secteurs de production industrielle « lourde » ne sont pas à l'abri, parce que certains biens, qui étaient des biens « de propriété » deviennent des biens « d'usage », valorisés par les services qui sont offerts avec eux ; et qui dit usage dit usage connecté, données, et optimisation.
On a déjà observé ce phénomène dans l'industrie aéronautique, qui est en pleine mutation vers cette « servicisation ». Cela a commencé par le cas des hélicoptères militaires : les armées veulent recourir à la location, acheter des heures de vol au lieu d'acheter des appareils. Du coup la maintenance incombe au constructeur des appareils : celui-ci ne dégagera de marge que s'il internalise cette maintenance et s'il sait l'optimiser.
Pour les moteurs d'avion, objets compliqués à produire s'il en est (il n'y a que trois fabricants dans le monde), l'idée s'est imposée que ce qui compte, ce n'est pas tant le prix d'achat que le « coût total de possession » pendant toute la durée d'usage. La maintenance de ces moteurs fait désormais l'objet d'une forfait à l'heure de vol, en lieu et place d'une maintenance opérée par les compagnies aériennes elles-mêmes. Du coup, progressivement la valeur s'est déplacée vers la technicité de la mise en œuvre opérationnelle des moteurs, la gestion des temps d'utilisation, des pannes, etc., toutes choses nécessitant l'analyse en temps réel de quantités colossales des données (1/2 téraoctet de données produite par moteur et par heure de vol).
Un processus similaire affecte le secteur automobile. Les constructeurs automobiles allemands, constatant la disparition progressive du « milieu de gamme » envisagent d'utiliser ce segment pour « serviciser », miser sur la location. Mais plutôt que de livrer des flottes entières à des entreprises de services, avec des marges faibles, ils semblent envisager de rendre eux-mêmes

le service, pour garder le bénéfice d'une connaissance des comportements des conducteurs. Au passage, cela leur permettrait de préciser les risques d'accidents et de les incorporer : les assureurs risquent d'être concurrencés de ce côté-là aussi.
L'industrie du logiciel n'échappe pas à cette tendance : aujourd'hui ce qui est vendu par SAS c'est un service, alors que la valeur marginale du software tend vers zéro. Mais les marges ne pourront se maintenir qu'en liant au logiciel des données exclusives permettant un service lui aussi exclusif.
Même le secteur administratif risque d'être touché. Je prendrai un exemple dans le domaine de la protection sociale en matière de santé. Les pré-diabétiques peuvent maintenant être munis de capteurs permettant d'alerter sans délai sur toute remontée de leur taux de glucose dans le sang. Il devient possible de proposer une protection préventive aux personnes acceptant d'entrer dans le jeu de la captation et de la transmission de données, alors que le système de protection sociale repose sur l'aspect curatif, sur la prise en charge des soins ex-post.

Les opportunités et les risques

On ne peut pas arrêter cette révolution industrielle dont le nom de code est « BigData» en créant des lignes Maginot règlementaires. Tout au plus peut-on essayer de gagner du temps pour s'adapter. Il s'agit d'économie et d'innovation : des emplois sont détruits, d'autres sont créés. Plutôt que de vouloir arrêter le processus, mieux vaut faire en sorte que les emplois recréés le soient dans notre pays. Le BigData, ce n'est pas principalement des technologies ou des outils, c'est d'abord cette puissante révolution dont les impacts sont énormes. C'est pour définir les moyens d'en tirer le meilleur parti en France que Paul Hermelin et moi-même avons reçu mission en 2013 de la part du ministre de l'économie pour jeter les bases d'un « plan BigData » français.
Et les risques ? Ils sont réels : risques d'atteintes à la vie privée des personnes, risques sociaux également. On pourrait en ajouter d'autres : risque de destruction d'emplois, de captation des richesses, etc. Il faut traiter tous ces risques sérieusement si l'on veut que les nouveaux usages puissent se développer.
Pour cela, on ne peut pas se reposer exclusivement sur la législation existante. La loi « Informatique et Libertés » de 1978 met en avant le respect de la finalité initiale des traitements : or, par définition, le BigData c'est la réutilisation de données au-delà de la finalité initiale pour laquelle elles ont été collectées. S'en tenir à la lettre de la loi reviendrait à interdire toutes les applications qui ont été évoquées. Il s'agit plutôt d'essayer de réfléchir à une réglementation qui fait référence à la finalité de « l'usage actuel » plutôt que de la finalité de la « collecte initiale ».
Nous devons rechercher une approche équilibrée, qui crée les conditions de la confiance des utilisateurs sans empêcher les innovations indispensables pour prendre place dans ce monde nouveau. Cette approche peut passer par la labellisation de processus industriels complets, avec les garanties convenables sur l'utilisation des données tout au long du processus, les anonymisations nécessaires, etc. On pourrait la mettre en œuvre par un système de « rescrit », analogue à ce qui se pratique en matière fiscale[3], par lequel s'exercerait le contrôle d'organismes comme la Commission Informatiques et Libertés. Tout ceci mérite d'être discuté, et ne nécessite pas forcément une modification de la loi à court terme.
Un système de normes souples, qui inspire une confiance justifiée sans paralyser l'innovation, serait de nature à soutenir les entreprises françaises qui se lancent dans des projets du BigData, et même à leur procurer un atout sur la scène internationale.

3. Dans le système du rescrit fiscal, une entreprise ou un particulier interroge l'administration fiscale sur la manière d'appliquer la règlementation fiscale dans le cadre de son activité, en lui fournissant toutes les informations nécessaires : la réponse de l'administration fiscale donne une sécurité juridique au contribuable, pour autant qu'il ait donné une description sincère et complète de son projet.

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

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

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

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

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

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

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

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

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

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

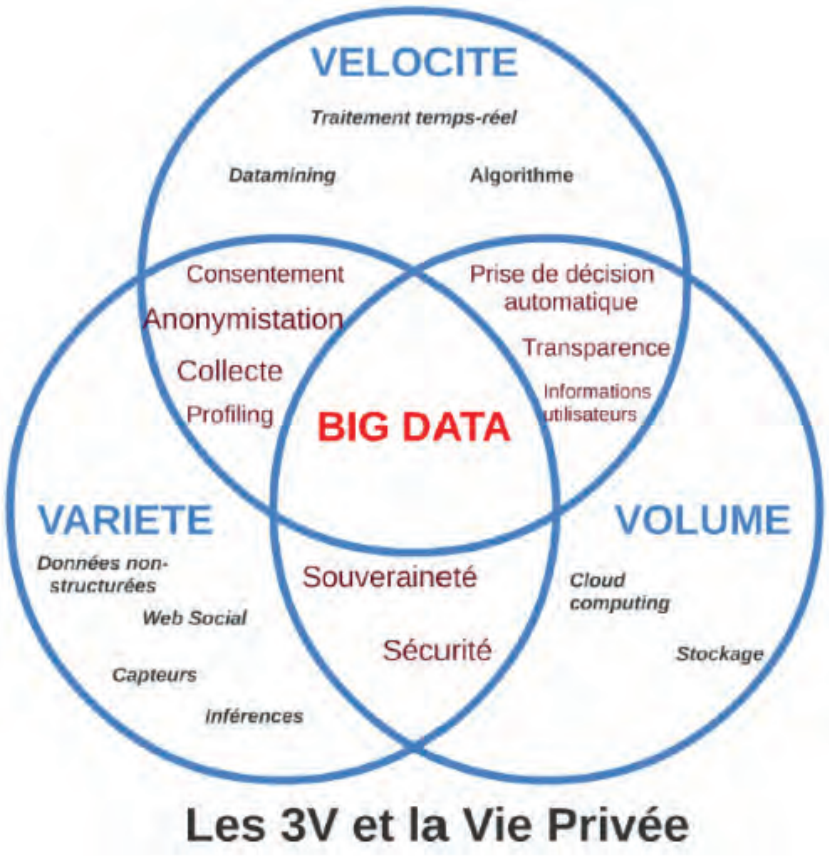


Figure 1 : Les mots-clés du BigData (source CNIL)

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

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

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

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

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

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

Finalité et pertinence : des principes remis en question ?

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

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

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

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

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

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

La « gouvernementalité algorithmique »

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

Loyauté de la collecte et respect des droits des personnes

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

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

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

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

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

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

données du web social sont-elle utilisables librement ?

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

Quelles mesures particulières de sécurité ?

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

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

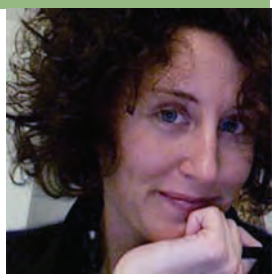
Enfin, en termes de stockage des données, les traitements BigData reposeront généralement sur le *cloud computing*, seul à même de fournir espace, souplesse et vitesse nécessaires. En matière de protection des données personnelles, les enjeux du cloud computing, bien connus, sont à la fois juridiques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

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

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

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

BigData : de nouveaux outils à combiner aux savoirs établis et à encadrer par la délibération publique



Entretien avec Antoinette ROUVROY

Chercheuse en Philosophie du Droit au Fonds National de la Recherche Scientifique (Belgique), rattachée au Centre de recherche Droit, Information, Société, Université de Namur

A la différence de la statistique classique qui repose sur des conventions et sur des hypothèses, les algorithmes qui fouillent les mégadonnées en font surgir sans médiation apparente des corrélations qui suggèrent en « temps réel » (sur le mode de l'alerte, de la recommandation, de l'aide à la décision), des actions immédiatement exécutables. Ces algorithmes sont la source de multiples profilages, qui délimitent les opportunités de chacun, à la place de normes communes. Porteuse de services étendus et personnalisés, aussi bien que de découvertes scientifiques, la révolution numérique est aussi un risque pour tout ce qui ne se laisse pas réduire à la rationalité économique, et notamment pour la justice sociale et pour la délibération collective. Les individus entrent presque tous dans le jeu, en permettant l'exploitation de leurs « traces », parce que les architectures de choix ne leur laissent voir que les côtés positifs de leur consentement. Plutôt que vers un renforcement de leur pouvoir sur leurs données personnelles, c'est vers un contrôle public des systèmes de décision susceptibles d'affecter les personnes qu'il faudrait se diriger, à travers la loi ou grâce à des codes de déontologie professionnelle

Statistique & Société : Votre réflexion envisage les BigData à la fois comme une nouvelle source de savoir et comme une nouvelle source de pouvoir – la fameuse « gouvernementalité algorithmique ». Nous allons aborder ces deux aspects successivement, si vous le voulez bien. D'abord, en quoi l'exploitation des mégadonnées est-elle pour vous différente de la mise en œuvre traditionnelle des outils de la statistique ?

Antoinette Rouvroy : Pour moi, *BigData* signifie surtout le franchissement d'un seuil à partir duquel nous serions contraints (par la quantité, la complexité, la rapidité de prolifération des données) d'adopter une rationalité purement statistique, inductive, se bornant à repérer des patterns, c'est-à-dire des *motifs* formés par les corrélations observées entre des données numériques, indépendamment de toute explication causale. La répétition de ces « motifs » au sein de grandes quantités de données leur confèrerait une valeur prédictive. Ainsi voit-on apparaître, grâce à la *visualisation* algorithmique des relations subtiles entre les données, un tout nouveau type de « savoir », exploitable dans une multitude de domaines. L' « intelligence » des algorithmes consiste en leur capacité à traiter statistiquement ces quantités massives,

complexes (textes, images, sons, localisations, trajectoires,...), relativement peu structurées, de données dans un temps record, pour en faire surgir non pas des relations causales explicatives mais des corrélations statistiquement signifiantes entre des éléments *a priori* sans rapport.

S&S : C'est donc très différent de la démarche classique adoptée par la statistique publique ?

AR : Lorsqu'on compare les pratiques statistiques nourries par les BigData aux statistiques classiques, on constate une série de glissements, tant du point de vue des finalités que du point de vue des pratiques et techniques impliquées.

L'une des finalités classiques des statistiques est *l'objectivation*, la confirmation d'hypothèses posées *a priori*, alors que les nouvelles pratiques statistiques impliquées dans le traitement des BigData visent précisément à dispenser d'avoir à poser des hypothèses *a priori*, et de *découvrir* celles-ci directement dans les mégadonnées par la grâce d'algorithmes capables d'y détecter des corrélations statistiquement signifiantes. Une autre des finalités classiques des statistiques est la *quantification* : elle consiste à rendre commensurables ou comparables entre eux, en les exprimant sous une forme chiffrée, des éléments de réalité hétérogènes. Ainsi, le *benchmarking des chercheurs* permet-il de comparer entre elles les performances de chercheurs spécialisés dans des domaines de recherche très différents. De telles opérations de *benchmarking* présupposent une série de conventions d'équivalence – toujours imparfaites, controversées -, établies à l'issue de discussions parfois longues entre représentants des disciplines concernées. En aval, elles permettent, ou obligent, suivant l'évaluation positive ou négative que l'on peut faire de ces systèmes, à utiliser, pour évaluer les mérites des uns et des autres, un langage commun, qui est celui des chiffres. Les pratiques statistiques nourries par les BigData s'écartent également de cette finalité d'organisation de la commensurabilité : le *datamining* ne vise pas à établir de commensurabilité qui permette la discussion, mais, prenant au sérieux l'incommensurabilité irréductible des situations, dispense d'avoir à discuter en faisant surgir des mégadonnées elles-mêmes, automatiquement et si possible en temps réel, les patterns, profils ou catégories les plus opérationnels en fonction des finalités (de gestion des ressources humaines, de sécurité, de marketing, etc.). Avec les BigData, nous ne sommes plus dans le monde un peu lent et conflictuel des conventions de quantification ouvertes sur l'espace public : le datamining s'inscrit dans un système d'immanence totale, dans lequel la collecte des données n'est plus orientée ni limitée par aucun système de catégories conventionnelles antécédentes ; on a supprimé la couche d'intermédiation que constituait la catégorie statistique et instauré, du même coup, un régime d'indistinction entre la réalité et sa représentation.

Enfin, dans la statistique classique, l'idée que l'on ne prend pas en compte la totalité des données disponibles est très importante. On *sélectionne* les données sur lesquelles on veut travailler, alors que les ambitions dominantes des promoteurs d'applications nourries par les BigData reposent sur l'idéal d'une coextension de la base statistique au réel (numérisé) lui-même. Dans le contexte des pratiques statistiques classiques la sélection consiste aussi à ne pas tenir compte des points trop éloignés de la « moyenne » ou de la « normale », à les considérer comme sources d'erreurs et de perturbations, à les considérer comme du « bruit » à exclure de la base statistique. L'ambition des applications nourries par les BigData est d'éviter cette sélectivité et ainsi, véritablement, d'*épuiser tout le possible* en tenant compte des cas les plus singuliers, les plus éloignés de la « moyenne » ou de la « normale ». A la différence des objets statistiques plus classiques à propos desquels on peut toujours argumenter qu'ils ne sont pas *représentatifs* de la réalité, qu'ils accordent trop ou pas assez d'importance aux situations singulières, les modélisations algorithmiques nourries par les *BigData* ont l'air d'absorber tout ce qui n'était pas pris en compte par les statistiques classiques : les points trop éloignés de la moyenne (qui pouvaient faire dire que les statistiques, ça ne vaut que pour les grands nombres, pas pour les cas individuels), ce qui n'entrait pas « dans les cases », c'est-à-dire dans les catégories statistiques établies par convention... Dans le monde des BigData, on peut prendre en compte

tous les points, y compris les plus atypiques, et on peut justement chercher à analyser les comportements les plus rares : on s'émancipe de tout rapport à la moyenne et à la « normale ». Dès lors, la personnalisation automatique des interactions administratives, sécuritaires, commerciales etc. à l'échelle industrielle n'est plus un oxymore.

Les statisticiens devraient se tenir fermement aux règles de leur discipline

Souligner ces différences, ce n'est pas nier l'intérêt pour les statisticiens d'investir dans ces nouvelles données et ces nouvelles méthodes, bien au contraire. Il me paraît tout à fait évident que dans de nombreux domaines de la connaissance (génétique, épidémiologie, astronomie, climatologie, etc.), les BigData ont un extraordinaire potentiel de reconfiguration de la perception, avec tout ce que cette révolution de l'accès au "réel" peut ouvrir comme perspectives inédites ne fût-ce que parce qu'elle nous débarrasse partiellement du joug perceptuel imposé aux sens ordinaires (qui nous rend inévitablement "prisonniers" d'un point de vue toujours trop partial et d'une perspective toujours trop partielle).
Mais les dispositifs nourris par les BigData sont un mode particulier de production de « ce qui compte comme réel ».

Il s'agit donc de ne pas se laisser leurrer par les promesses « d'objectivité par la totalité et par l'automaticité » que certains profèrent, et d'être conscients de ce qu'ils risqueraient de perdre s'ils prenaient ces promesses pour argent comptant.

J'insiste particulièrement sur le caractère de convention sociale attaché aux définitions statistiques : selon Alain Desrosières, là résident à la fois la fragilité et la force, des statistiques. Les statisticiens devraient se tenir fermement aux règles de leur discipline, aux modes de véridiction qui leur sont propres et qui n'ont pas à céder devant les classifications produites par les algorithmes de *datamining*. Si, en raison de leur origine conventionnelle, les objets statistiques servant de références dans les débats publics sont toujours suspects de répercuter et de « naturaliser » les biais, préjugés et normativités sociales dominantes, l'absence de convention présidant au profilage algorithmique ne garantit qu'une objectivité de façade qui naturalise de manière cette fois absolument invisible, les normativités sociales rendues indiscutables.

S&S : Comment la nouvelle rationalité de l'exploitation des BigData permet elle de mettre en place de nouveaux moyens de gouverner les personnes ?

AR : En exploitant les *profils* induits par les corrélations, on peut détecter, sans avoir à les rencontrer ni à les interroger personnellement, ni même à les identifier précisément, les risques et opportunités dont sont porteuses les personnes. Les algorithmes produisent des catégorisations impersonnelles, évolutives en continu, en fonction des attitudes, des trajectoires, etc. Etre profilé consiste à être appréhendé dans un réseau de formes « percées » qui ne peuvent jamais nous contenir totalement, mais qui tracent, en pointillés mobiles, nos trajectoires à venir. Être « profilé » de telle ou telle manière affecte les opportunités qui nous sont disponibles, et, ainsi, l'espace de possibilités qui nous définit : le gouvernement algorithmique ne s'intéresse pas tant à ce que nous avons fait et faisons, qu'à ce que nous aurions pu faire ou pourrions faire dans l'avenir, c'est à notre dimension « potentielle », « virtuelle » donc, qu'il s'adresse. La spécificité, ou la radicale nouveauté de la gouvernementalité algorithmique tient notamment en ceci qu'elle affecte les individus en neutralisant (en privant d'effets disruptifs, sans pour autant les supprimer) leurs dimensions inactuelles (la dimension de la spontanéité, de la potentialité), sans pour autant les assujettir à aucune « norme » - à la différence de la discipline - ni mobiliser leurs capacités d'entendement et de volonté - à la différence de la loi.

L'exemple le plus connu est celui du marketing : si l'on en croit Eric Schmidt, directeur chez Google, bientôt la technologie deviendra tellement efficace qu'il deviendra très difficile pour les personnes de voir ou consommer quelque chose qui n'aurait pas été prévu pour elles. En sens inverse, aucune norme sociale – nous avons vu déjà que le monde des BigData s'était émancipé de tout rapport à la « normale » ou à la « moyenne » - n'impose ni ne suggère de limites à l'exploitation des possibilités de profit : une personne ayant des addictions, « droguée au chocolat » par exemple, se verra inviter à acheter indéfiniment le produit sans aucune autre limite que sa propre satiété. Il risque de devenir de plus en plus difficile de résister à la « manipulation digitale » : d'une part, le « temps réel » est un temps dans lequel les humains n'ont pas la possibilité de prendre du recul relativement à leurs propres pulsions (d'achat, par exemple), un temps dans lequel on fonctionne sur le mode de l'alerte et du réflexe, plutôt que sur le mode de la réflexivité ; d'autre part, chaque consommateur se retrouve seul face à la sollicitation, parce celle-ci est adaptée à ce qu'il a de singulier, excluant de ce fait toute possibilité de résister ensemble. Il en résulte que nous n'avons plus même à former ni à formuler par nous mêmes nos désirs : ceux-ci nous précèdent sous une forme adaptée à la fois à l'offre et à ce que chacun de nous a de plus singulier, de plus éloigné des grands nombres.

Autre exemple : un programme de financement de l'Union Européenne suscite des innovations techniques favorisant le maintien des personnes âgées à leur domicile. Avec les meilleures intentions du monde, les visionnaires du futur qui répondent à cet appel d'offres proposent d'équiper d'un très grand nombre de capteurs les appartements des personnes âgées, pour permettre d'intervenir face à toute forme d'évènement ou de comportement inattendu (une chute, une immobilité prolongée, des déplacements nocturnes, un défaut de prise d'un médicament,...). A aucun moment ces innovateurs, préoccupés de réduire toute incertitude, ne pensent qu'une certaine intimité peut être pour ces personnes indispensable, fût-ce au prix de leur sécurité. Il ne s'agit pas d'un simple pas supplémentaire dans le progrès technique. Ces appareils ne sont pas faits pour faire quelque chose à notre place : ce sont des appareils qui vont nous faire faire des choses, en vertu d'une notion de besoin détecté en temps réel à partir des enregistrements qui seront faits, et donc selon une normativité immanente qui fait l'économie de la volonté des personnes elles-mêmes, et de toute délibération autour de la nature de ces besoins.

S&S : On pourrait arriver à des situations extrêmes avec les objets connectés, comme dans le cas de ce bracelet mis au point par une société américaine, que l'on porterait en permanence pour diminuer le coût de son assurance automobile, si on accepte que soient ainsi mesurés les temps d'exercice physique et de sommeil...

AR : C'est un très bon exemple. En matière d'assurance, le profilage par des algorithmes peut permettre de former des groupes de plus en plus restreints, évolutifs, de manière à ajuster en permanence les primes demandées aux risques de sinistres : la logique actuarielle est alors poussée à l'extrême, au détriment des principes de mutualisation qui fondent l'assurance, et qui supposent qu'on accepte l'existence d'une part irréductible d'incertitude. Toute incertitude est vue comme un résidu à neutraliser par un raffinement de l'algorithme. Les catégories produites par le *datamining* ne sont pas nécessairement « justes » ni « équitables ». Elles le seraient si, par exemple, les notions de justice actuarielle (en fonction de laquelle toute distinction de traitement économiquement rationnelle serait actuariellement juste, chacun ayant à contribuer au fonctionnement de l'assurance en payant des primes ajustées à « son » risque, c'est-à-dire à la probabilité qu'il bénéficie, plus tard, de la compensation d'un dommage qui se serait réalisé) et de justice sociale se recouvraient parfaitement, ce qui n'est bien évidemment pas le cas. Une distinction de traitement qui exclurait par exemple systématiquement les personnes victimes de violences conjugales du bénéfice de l'assurance vie, quels que soient le sexe, l'origine sociale de ces personnes, sur base d'une attribution de profil de risque établi par une méthode inductive de *datamining*, pourrait bien être algorithmiquement et économiquement « rationnelle »,

actuariellement justifiée, et socialement injuste. On perçoit bien, en l'occurrence, le danger associé au déploiement d'un « régime de vérité » numérique impartial et opérationnel mais qui dispenserait de toute discussion politique, de toute décision collective, et de toute contestation relative aux critères de besoin, de mérite, de dangerosité, de capacités qui président aux catégorisations bureaucratique et/ou sécuritaire des individus et comportements. Notons en passant que l'individualisation « parfaite » des risques et opportunités signifierait tout aussi bien la fin de la raison d'être des assurances, dont le rôle premier n'est certainement pas d'individualiser la charge des risques mais au contraire, de constituer des « contrats sociaux » restreints entre des personnes, les assurés, qui, soumis à des risques comparables, s'engagent à prendre en charge collectivement les coups du sort qui s'abattraient sur certains d'entre eux. Le *datamining* permettrait le passage d'une société actuarielle à une société post-actuarielle.

L'alliance des visionnaires bienveillants et des commerçants intéressés

S&S : C'est donc ce que vous appelez « gouverner au moyen d'algorithmes », en visant spécifiquement les algorithmes qui à partir d'analyses de données comportementales ou autres proposent des actions qui n'ont plus qu'à être exécutées. Mais qui est là-derrière ? Qui promeut ce genre d'algorithmes ?

AR : Je n'ai pas une théorie du complot, rassurez-vous ! Je constate une curieuse convergence, parmi les « promoteurs » d'une gouvernementalité algorithmique, entre des groupes d'opinions qui a priori semblent très éloignés, mais qui ne le sont pas tant que cela en réalité. D'une part, un courant « anarchiste pro-cybernétique » ou crypto-anarchiste voit dans les applications en réseaux la promesse d'une société sans État ni institutions – on rejoint l'idée de l'im-médiation, d'une sortie de la logique de la représentation. On pourrait penser, naïvement, que ce rêve d'immanence corresponde aux idéaux de la pensée critique héritée des années 1960-1970 (Gilles Deleuze, Félix Guattari, Michel Foucault,...) ; mais alors que cette pensée critique était véritablement une pensée du « virtuel », une pensée de l'ouverture au non numérisable, au non maîtrisable, ce à quoi nous assistons aujourd'hui avec cette modélisation du social à même le social numérisé, c'est à une clôture du numérique sur lui-même, à une neutralisation du possible. D'autre part, la gouvernementalité algorithmique s'inscrit parfaitement dans la continuité de la gouvernementalité néolibérale dont elle n'est en somme qu'un « perfectionnement ». Appuyés sur une forme d'« idéologie technique », ces gens conçoivent l'idéal d'une société qui se gouvernerait toute seule, en temps réel, l'a priori étant que les individus, une fois émancipés du joug des institutions, se trouveraient dans une situation de parfaite égalité en termes de moyens et de bien-être. La possibilité de modéliser le social sans intermédiation, en dehors des conventions imposées par des autorités de toutes natures, rejoint leurs rêves. Ils essaient de faire du « design utile », avec une réelle bienveillance, mais sans prendre garde qu'ils ont un certain point de vue qui n'est pas forcément celui des utilisateurs. Ils auraient intérêt à s'entourer d'anthropologues...ou de philosophes, qui pourraient leur rappeler les limites de l'idéologie technique, qui renvoie dans l'arrière-fond invisible les constructions techniques, et les visions du monde qui les sous-tendent, et qui propose à la place une « interprétation globale du réel censée valoir par elle-même, comme si le réel lui-même parlait... »[1]. D'autre part, les promoteurs habituels d'une gouvernance mondiale de type néolibérale se retrouvent très bien dans le projet d'une gouvernementalité algorithmique destituant l'Etat et l'espace de délibération publique. Très prosaïquement, les grandes sociétés multinationales ont bien vu qu'avec les systèmes fondés sur des algorithmes, toute distinction de traitement des individus qui serait économiquement justifiée devient automatiquement légitime : l'émancipation de toute norme libère de toute contrainte, de tout scrupule, les objectifs de maximisation des profits. Ainsi des enthousiasmes très libertariens s'articulent à des intérêts qui font peu de cas de la liberté !

1. Pierre Macherey, « Idéologie : le mot, l'idée, la chose. Langue, discours, idéologie, sujet, sens : de Thomas Herbert à Michel Pêcheux », 17/01/2007, http://stl.recherche.univ lille3.fr/seminaires/philosophie/macherey/macherey20062007/macherey17012007.html

S&S : Alors, selon vous, il faut rejeter tout en bloc ?

AR : Ce n'est certainement pas ce que je dis. Je vois de nombreux domaines scientifiques comme l'astronomie, l'épidémiologie, la climatologie, la génétique, etc. dans lesquels les BigData vont permettre de faire des découvertes inattendues, et très intéressantes, surtout si on sait les combiner aux savoirs établis et aux théories existantes. Et bien sûr, de multiples applications utiles sont à la portée des BigData en aval de ces découvertes.
En revanche, quand il s'agit d'interventions dans la vie des personnes et dans la vie sociale, j'essaie d'attirer l'attention sur les risques de cette nouvelle forme de « gouvernementalité » pour la justice sociale et pour la délibération publique, les deux étant bien entendu liés.
Le danger d'injustice sociale est démultiplié si les procédures mises en place excluent de fait les espaces de délibération publique. Les conclusions se présentent comme des « vérités numériques » impartiales et opérationnelles, qui dispenseraient de toute discussion politique, de toute décision collective, et de toute contestation relative aux critères de besoin, de mérite, de dangerosité, de capacités qui président aux catégorisations bureaucratique et/ou sécuritaire des individus et comportements.

S&S : En multipliant leurs « traces » numériques, et en permettant l'utilisation de ces données, les individus entrent massivement dans ce jeu. Pourquoi ne sont-ils pas plus méfiants ?

AR : En ce qui concerne les données à caractère personnel (dont on sait qu'elles ne sont pas absolument nécessaires aux opérations de profilage des personnes, qui peuvent très bien être réalisées en n'utilisant que des données anonymes, des métadonnées, etc.), le succès des règles de conservation des données par défaut ou, pour le dire autrement, le manque de succès des options permettant de déroger à cette règle de conservation des données tient, si l'on en croit Cass R. Sunstein, se fondant sur l'économie comportementale, à la combinaison de trois facteurs principaux : 1) Le premier facteur est l'inertie des comportements dès lors qu'effacer « ses traces » demande un effort dont on ne sait au juste s'il vaut vraiment la peine, étant donné que chacune des données émanant de nos activités en ligne nous paraît à nous-mêmes, *a priori* (indépendamment des opérations de recoupement, de croisement, de modélisation auxquelles elles pourraient contribuer), de peu d'importance. La règle par défaut, quand bien même nous avons la possibilité d'y déroger très facilement « en quelques clics » prévaudra toujours lorsque l'enjeu ponctuel, actuel, n'apparaît pas significatif aux yeux de l'internaute. 2) Le second facteur favorisant la règle de conservation par défaut consiste en ceci que, dans une situation d'incertitude quant à la marche à suivre, l'utilisateur moyen aura tendance à considérer que la règle par défaut, puisqu'elle a été pensée par d'autres que lui, réputés plus experts et puisqu'elle est probablement suivie par la plupart des autres personnes, est sans doute la meilleure option pour lui aussi. 3) Enfin, le troisième facteur consiste dans le fait que les individus soient généralement plus sensibles au *risque de perdre* un avantage dont ils ont ou croient avoir la jouissance en se maintenant dans la situation dans laquelle ils se trouvent qu'à *l'opportunité de gagner* quelque chose en changeant. C'est une variante du phénomène d'inertie mais à travers laquelle les concepteurs, les « designers », les « marketeurs » peuvent avoir une prise sur les individus : ils peuvent réduire la probabilité que les utilisateurs s'écartent de la règle par défaut dans l'ajustement des « règles de confidentialité » en évoquant tout ce qu'ils ont à perdre dans la mesure où la rétention des « traces numériques » est ce qui permet de leur offrir un service plus personnalisé, plus adapté à leurs besoins en temps réel en fonction du lieu où ils se trouvent, ou de leurs goûts, un service plus rapide et efficace, et que l'effacement leur fera perdre tous ces avantages suffira généralement à éviter que l'utilisateur ne s'écarte de la règle par défaut.

Superviser les architectures de choix

S&S : Que suggérez-vous ?

AR : Cette question des « architectures de choix » - qui affectent le consentement des personnes à la conservation de « leurs » données, mais aussi leurs propensions à « faire confiance » aux recommandations d'achat automatisées qui leur sont envoyées, ou à se fier aux « rankings » de Google pour évaluer la pertinence et la valeur des contenus informationnels fournis par le moteur de recherche - est très importante : ce sont elles qui conditionnent en partie la capacité des personnes à intervenir de façon réfléchie et non pas de façon réflexe. Une supervision éthique et juridique des architectures de choix spécifiques aux plateformes est à mettre en place, qui soit fondée sur une typologie fine des acteurs et de leurs intérêts, suivant que ces intérêts sont plus ou moins alignés sur les intérêts des « utilisateurs » : un hôpital ne devrait pas être traité de la même façon qu'un commerçant. Il s'agirait de ménager la possibilité de processus de véridiction et de justification, donc de mises à l'épreuve des productions des catégorisations émanant des BigData en tenant compte de la nature des intérêts en jeu. Suivant une typologie fine des acteurs et surtout de leurs intérêts, il est possible de distinguer les situations dans lesquelles les acteurs – ceux qui exploitent les données d'une part, et ceux que l'on appelle parfois un peu abusivement les utilisateurs (consommateurs, citoyens,... qui sont aussi, en partie, les producteurs des données) – ont des intérêts alignés et les situations où c'est l'inverse. Exemples : l'hôpital et les patients ont, en principe, même si ce n'est pas toujours complètement le cas, des intérêts alignés, c'est-à-dire convergents vers la guérison, bonne pour le patient, pour la réputation de l'hôpital, etc. alors qu'une compagnie d'aviation et les voyageurs peuvent avoir des intérêts désalignés, la compagnie souhaitant faire payer les voyageurs le plus cher possible tout en restant concurrentiels, et les voyageurs ayant, eux, intérêt à payer le moins cher possible. Le « profilage » peut jouer soit dans l'intérêt des deux parties, soit dans l'intérêt de l'une d'entre elles seulement, au détriment de l'autre. Ainsi la classification des patients dans certains « profils » thérapeutiques est dans l'intérêt tant de l'hôpital que du patient, alors que le profilage des voyageurs dans un certain « profil » en fonction de leur prédisposition à vouloir payer un certain prix pour un certain voyage (la « personnalisation » ou le profilage algorithmique permettent d'ajuster les prix en fonction de la disposition à payer [willingness to pay] de chaque client ; au plus le client aura « besoin » de voyager à telle date rapprochée, au plus cher sera son billet), est le plus souvent défavorable au voyageur (la somme qu'il aura à débourser pour voyager étant adaptée à l'élasticité de sa disposition à acheter un billet en fonction d'une évolution des prix). En fonction, donc, des applications, et des intérêts en jeu, il importe, afin de protéger la « partie faible » (celle qui subit, généralement, une asymétrie d'information, n'étant pas au courant des logiques de traitement de données qui président aux décisions qui l'affectent et ne se sachant pas « profilée »), de veiller à assurer des possibilités de contester les productions algorithmiques (« catégorisations » ou « profilages »).

En ce qui concerne les algorithmes eux-mêmes, je ne crois absolument pas à la possibilité de les rendre « transparents », comme certains le proposent aujourd'hui. A la technicité des processus algorithmiques et aux perspectives de dispositifs auto-apprenants les rendant difficilement intelligibles pour le commun des mortels s'ajoutent les obstacles juridiques imposés par le secret industriel ou le secret-défense. Mais je crois au recours à la loi, et aux codes de déontologie. Le projet de nouveau règlement européen sur la protection des données contient un article, déjà présent dans la directive de 1995 et dans la loi française de 1978, qui stipule que :

> *«Toute personne a le droit de ne pas être soumise à une mesure produisant des effets juridiques à son égard ou l'affectant de manière significative prise sur le seul fondement d'un traitement automatisé destiné à évaluer certains aspects*

personnels propres à cette personne physique ou à analyser ou prévoir en particulier le rendement professionnel de celle-ci, sa situation économique, sa localisation, son état de santé, ses préférences personnelles, sa fiabilité ou son comportement (…) »

Au-delà de la loi, je crois possible de persuader les acteurs du domaine de l'intérêt qu'ils ont à se doter de codes professionnels pour encadrer leurs pratiques. Ils ont accès à des possibilités de manipulation de l'intellect des gens ! Certains des « grands principes » formant l'ossature des régimes juridiques de protection des données à caractère personnel pourraient rester très inspirants moyennant une série d'adaptations (dans le contexte des BigData, nous n'avons plus affaire que minimalement à des données à caractère personnel, les données ne servant plus tant à identifier qu'à catégoriser, etc.). Je pense notamment au principe de loyauté de la collecte des données qui pourrait contribuer à discipliner certaines pratiques de « smarter marketing » qui, au lieu de s'appuyer sur les capacités d'entendement et de volonté des consommateurs, capitalisent au contraire sur les faiblesses de leur volonté et de leur entendement, détectées en temps réel, de manière à les faire passer à l'acte d'achat sans qu'ils aient eu l'occasion de même comprendre leurs propres motivations, ou encore à des pratiques de marketing qui n'ont pas pour but de vendre quoi que ce soit mais seulement de recueillir davantage d'informations sur les consommateurs de manière à mieux les profiler. Je parle ici de marketing et cela paraît trivial, mais on peut imaginer que ces mêmes techniques soient utilisées à des fins d'individualisation de la communication électorale, par exemple, ou à des fins de conditionnement des enfants à la surconsommation dès le plus jeune âge. L'enjeu, c'est notre intégrité mentale.

S&S : Si l'on vous suit bien, ce n'est pas la protection des données à caractère personnel qui est le point fondamental ?

AR : Nous n'avons appris à nous méfier que des traitements automatisés de données à caractère personnel : or celles-ci n'interviennent que marginalement dans les phénomènes qui nous intéressent ici. Une sorte de fétichisation de la donnée personnelle – renforcée par le droit positif actuel – nous fait passer à côté de ce qui fait aujourd'hui problème. Les nouvelles formes de pouvoir qui s'exercent sur les individus passent beaucoup moins par les traitements de données à caractère personnel et l'identification des individus que par des catégorisations impersonnelles, évolutives en continu, des opportunités et des risques, c'est-à-dire des formes de vie (attitudes, trajectoires,…). Un profil, ce n'est en réalité personne – personne n'y correspond totalement, et aucun profil ne vise qu'une seule personne, identifiée ou identifiable. Ce n'est pas le risque d'identification qui est le plus dangereux, c'est le risque de catégorisation, sans outil de critique des catégories et de récalcitrance par rapport à elles. Et donc, ce n'est pas « plus de privé » qu'il nous faut, c'est au contraire « plus d'espace public ».

S&S : Finalement, est-ce qu'on n'est pas un peu dans une « Querelle des Anciens et des Modernes » ? Que devrions-nous dire aux jeunes générations, de statisticiens, ou simplement de citoyens ?

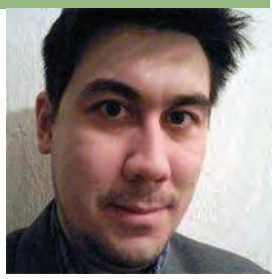
AR : Je l'ai déjà dit : je ne suis pas une adversaire des BigData, j'en perçois très bien les avantages en termes d'avancement de la connaissance scientifique et en termes de nouveaux services utiles. Mais j'en perçois les dangers, d'autant plus marqués que nous vivons actuellement dans une sorte de « bulle spéculative » à propos des BigData. J'aimerais faire partager cette vision équilibrée.

Aux statisticiens je dirais : engagez-vous dans ce mouvement, mais conservez vos principes professionnels et la lucidité de vos prédécesseurs sur l'insertion sociale de leur discipline et sur son rôle dans la constitution d'un espace public de délibération.

Pour le reste, j'ai bon espoir, et je me réjouis du débat qui s'instaure en France notamment à propos des possibilités et des risques des BigData. En ces temps de crise de la représentation et

d'assèchement de l'espace public, ces débats sont une occasion pour repenser collectivement des questions essentielles. Quelle partie de nos fonctions cognitives voulons-nous déléguer à des algorithmes ? Dans quel rythme temporel voulons-nous vivre ? Quelle place voulons-nous faire à la mémoire des tentatives, de l'ineffectué (seules les choses effectuées laissent des traces numériques, rejetant dans l'oubli les projets trop tôt abandonnés, les utopies non tentées par nos prédécesseurs, mais qui sont autant de sources d'inspiration pour l'avenir) ? Et surtout, comment ménager encore l'espace de l'événement, de l'imprévisibilité toujours renouvelée, et donc de la liberté ?

Techniques d'anonymisation



Benjamin NGUYEN

Insa[1] Centre Val de Loire et Inria[2] Paris-Rocquencourt

L'opposition entre une donnée qui permet d'identifier une personne et une donnée anonyme n'est pas une opposition absolue. C'est pourquoi il existe plusieurs méthodes d'anonymisation, plus ou moins efficaces. On utilise souvent aujourd'hui la « k-anonymisation », la « l-diversité », ou la « confidentialité différentielle », trois techniques dont les principes sont donnés dans cet article. Les différentes techniques sont à juger à la fois sur la sécurité qu'elles procurent, et sur ce qu'elles laissent subsister comme analyses possibles.

Il existe légalement deux types de données : des données à caractère personnel, et des données anonymes. Les données sont à caractère personnel « *dès lors qu'elles concernent des personnes physiques identifiées directement ou indirectement* » pour citer la CNIL. Au contraire, toute donnée qu'il est impossible d'associer avec une personne physique sera dite « *anonyme*. » Il est intéressant de constater que la loi française définit une impossibilité forte, puisqu'elle précise que « *pour déterminer si une personne est identifiable, il convient de considérer* ***l'ensemble des moyens*** *en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.* » Plus mesuré, le projet de règlement européen prévoit que « *pour déterminer si une personne est identifiable, il convient de considérer* ***l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre****, soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne. Il n'y a pas lieu d'appliquer les principes de protection aux données qui ont été rendues suffisamment anonymes pour que la personne concernée ne soit plus identifiable.* » En d'autre termes, le projet de règlement reconnaît une quantification et gradation dans la méthode d'anonymisation.

Cette définition sous entend qu'il existe plusieurs méthodes d'anonymisation, plus ou moins *efficaces* (au sens d'une protection plus ou moins « forte »). Pourquoi n'utiliserait-on pas toujours « la meilleure », ou dit autrement, est ce qu'il y a coût à payer pour avoir une anonymisation forte ? On pourrait imaginer qu'une partie du coût serait un coût temporel, i.e. qu'il serait très long de calculer une « bonne » anonymisation. On pourrait aussi penser que la force de l'anonymisation est inversement proportionnelle à la quantité (ou précision) des données publiées i.e. si on publie des informations au niveau d'un département on est à peu près 4 fois moins anonyme que si on les publie au niveau d'une région. Bien que tous ces facteurs entrent en jeu, le facteur déterminant est le *type de modèle d'anonymisation* utilisé. Il faut en effet prendre garde, car celui-ci peut restreindre l'exploitation future des données anonymes à certains types de calculs. Pour que le lecteur puisse facilement se représenter les données manipulées, nous considérons

1. Institut national des sciences appliquées
2. Institut national de recherche en informatique et en automatique

une base de données constituée d'un ensemble d'enregistrements (appelés n-uplets) ayant chacun une structure identique, c'est-à-dire les mêmes champs, par exemple : numéro de sécurité sociale, nom, adresse, date de naissance, salaire, etc. (voir Figure 1). On repère dans un n-uplet des données dites *sensibles* comme une pathologie médicale, le salaire voire l'adresse.

Numéro de sécurité sociale (*Identifiant*)	**Age**	**Code postal**	**Sexe**	**Pathologie** (*Donnée sensible*)
2023475123123	75	75005	F	Cancer
2067875123123	40	75012	F	Grippe
1101175123123	12	78000	M	Grippe

Figure 1. Une base de données personnelles

Dans cet article, nous allons décrire cinq types de modèles d'anonymisation, qui cherchent à cacher ou briser le lien existant entre une personne du monde réel, et ses données sensibles : la pseudonymisation, le *k*-anonymat, la *l*-diversité, la *t*-proximité et la *differential privacy* (confidentialité différentielle, au sens du calcul différentiel). Nous illustrerons leur utilisation possible et leur degré de protection.

La pseudonymisation

La pseudonymisation consiste à supprimer les champs *directement* identifiants des enregistrements, et à rajouter à chaque enregistrement un nouveau champ, appelé *pseudonyme*, dont la caractéristique est qu'il doit rendre impossible tout lien entre cette nouvelle valeur et la personne réelle. Pour créer ce pseudonyme, on utilise souvent une *fonction de hachage* que l'on va appliquer à l'un des champs identifiants (par exemple le numéro de sécurité sociale), qui est un type de fonction particulier qui rend impossible (ou tout du moins extrêmement difficile) le fait de déduire la valeur initiale. On voit ainsi que deux entités possédant des informations sur une même personne, identifiée par son numéro de sécurité sociale, pourraient partager ces données de manière anonyme en *hachant* cet identifiant. Il est également possible d'utiliser tout simplement une fonction aléatoire pour générer un identifiant unique pour chaque personne, mais nous verrons plus bas que cela ne résout pas tous les problèmes.

Le gros avantage de la pseudonymisation est qu'il n'y a aucune limite sur le traitement subséquent des données. Tant que l'on traite des champs qui ne sont pas directement identifiants, on pourra exécuter exactement les mêmes calculs qu'avec une base de données non-anonyme. Ainsi, on montre dans la Figure 2 un exemple de calcul de la moyenne d'âge pour une pathologie donnée. L'utilisation de données pseudonymisées ne nuit pas à ce calcul.

Pseudonyme

Donnée sensible

ID	Age	CP	Sexe	Pathologie
1	75	75005	F	Cancer
2	40	75012	F	Grippe
3	12	78000	M	Grippe

Pathologie	MOY(age)
Cancer	75
Grippe	26

Figure 2. Pseudonymisation et exemple de calcul

Toutefois, la pseudonymisation n'est pas reconnue comme un moyen d'anonymisation, car elle ne donne pas un niveau de protection suffisamment élevé : la combinaison d'autres champs peut permettre de retrouver l'individu concerné. Sweeney l'a mis en évidence aux Etats-Unis en 2001 en croisant deux bases de données, une base de données médicale pseudonymisée et une liste électorale avec des données nominatives. Le croisement a été effectué non pas sur des champs directement identifiants, mais sur un triplet de valeurs : code postal, date de naissance et sexe, qui est unique pour environ 80% de la population des Etats-Unis[3] ! Elle a ainsi pu relier des données médicales à des individus (en l'occurrence le gouverneur de l'Etat).

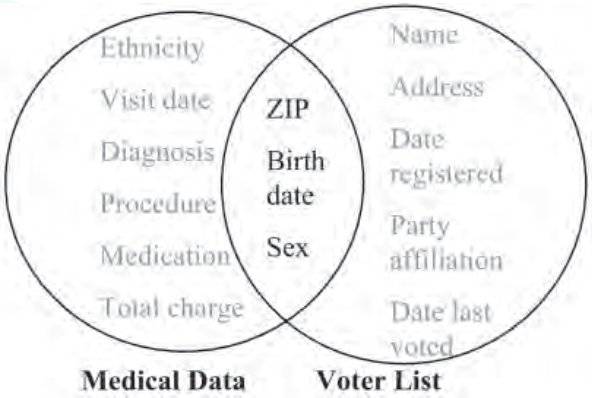


Figure 3. Un exemple de recoupement d'une base anonyme (source Sweeney 2002)

3. Autrement dit : une personne qui appartient à ce groupe de 80% de la population est seule à posséder son triplet code postal - date de naissance – sexe. Dans le complément de 20%, les personnes partagent leurs triplets avec une ou plusieurs autres personnes.

Le *k*-anonymat

Afin de se protéger contre ce type d'attaque, appelée *record linkage*[4], Sweeney a proposé la technique de *k*-anonymat. Celle-ci va flouter la possibilité de lier un *n*-uplet anonyme à un *n*-uplet non anonyme de la manière suivante : 1) déterminer les ensembles d'attributs (appelés *quasi*-identifiants) qui peuvent être utilisés pour croiser les données anonymes avec des données identifiantes ; puis 2) réduire le niveau de détail des données de telle sorte qu'il y a au moins *k* *n*-uplets différents qui ont la même valeur de *quasi*-identifiant, une fois celui-ci généralisé (on dit alors que les individus font partie de la même classe *d'équivalence*). « Généraliser » signifie en fait « enlever un degré de précision » à certains champs. Ainsi, il est impossible d'être sûr à plus d'une chance sur *k* qu'on a bien lié un individu donné avec son *n*-uplet anonyme. L'avantage du *k*-anonymat est que l'analyse des données continue de fournir des résultats exacts, à ceci près qu'on ne peut pas dissocier les individus d'un groupe. Dans la Figure 4, nous montrons un exemple de généralisation des champs activité et âge d'une base de données médicales sur des étudiants et enseignants d'une université. Les étudiants sont identifiés par leur niveau d'étude (L3, M1, etc.), qui se généralise en « étudiant », et les enseignants par leur position académique (doctorant, maître de conférences, etc.), qui se généralise en « enseignant ». Nous traçons dans cette Figure l'origine de chaque *n*-uplet flouté.

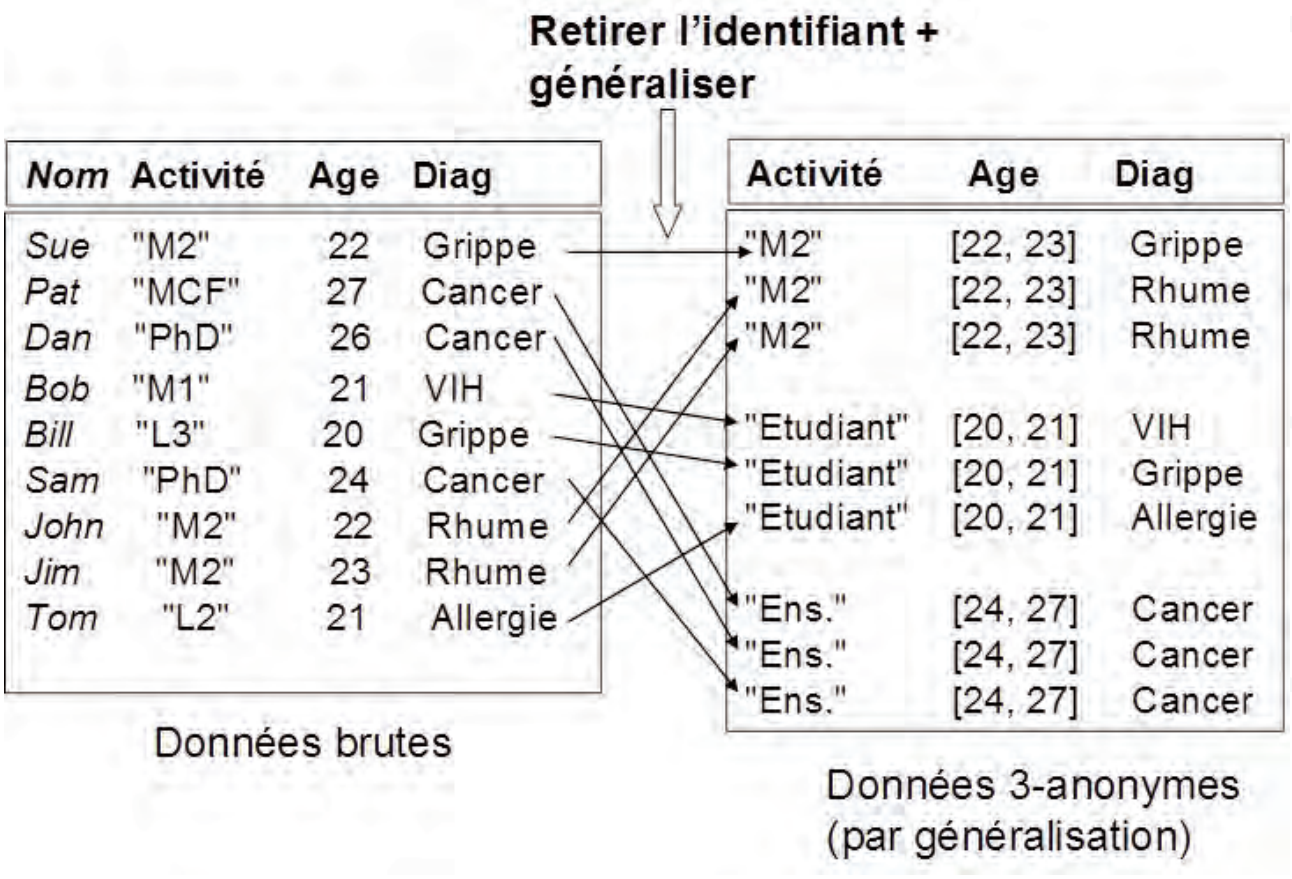


Figure 4. Anonymisation d'une table sur des données universitaires

Toutefois, une certaine quantité d'information sera déjà dévoilée, en particulier de l'information négative : si on connaît le quasi-identifiant d'une personne, on pourra exclure tout un ensemble de valeurs, ou bien savoir qu'elle a de plus grandes chances d'avoir une certaine valeur sensible.

4. Liaison entre enregistrements

Certains cas peuvent aussi apparaître : si tous les individus d'une classe d'équivalence possèdent les mêmes valeurs sur un champ intéressant l'attaquant, alors celui-ci sera capable d'identifier cette valeur. Par exemple, en considérant les données de la Figure 4, on peut déduire qu'un enseignant ayant un âge entre 24 et 27 ans a forcément le cancer. Si on sait que Sam est un doctorant de 24 ans, alors on peut en déduire qu'il a le cancer.
Enfin, un problème technique important subsiste pour réaliser le k-anonymat : être capable de déterminer les généralisations à effectuer pour produire les quasi-identifiants, ce qui peut être fait soit par un expert humain qui connaît le domaine, ou bien par un calcul informatique, souvent très coûteux pour une base de données réelle.

La *l*-diversité

Comme on l'a vu à la Figure 4, il est possible de déduire des informations dans certains cas pathologiques, sans faire le moindre croisement, par exemple si tous les individus d'une classe possèdent la même valeur sensible. Le modèle de la l-diversité répond à ce problème, en rajoutant une contrainte supplémentaire sur les classes d'équivalence : non seulement au moins *k* *n*-uplets doivent apparaître dans une classe d'équivalence, mais en plus le champ sensible associé à la classe d'équivalence doit prendre au moins l valeurs distinctes[5]. Dans l'exemple de la Figure 5, on voit que pour constituer de telles classes on doit parfois regrouper ensemble des étudiants et des enseignants. Leur activité est alors désignée de façon encore plus générale (« université »). Notons qu'on peut également lister les valeurs possibles, par exemple avoir une modalité « Étudiant ou Doctorant » (Etu/PhD).

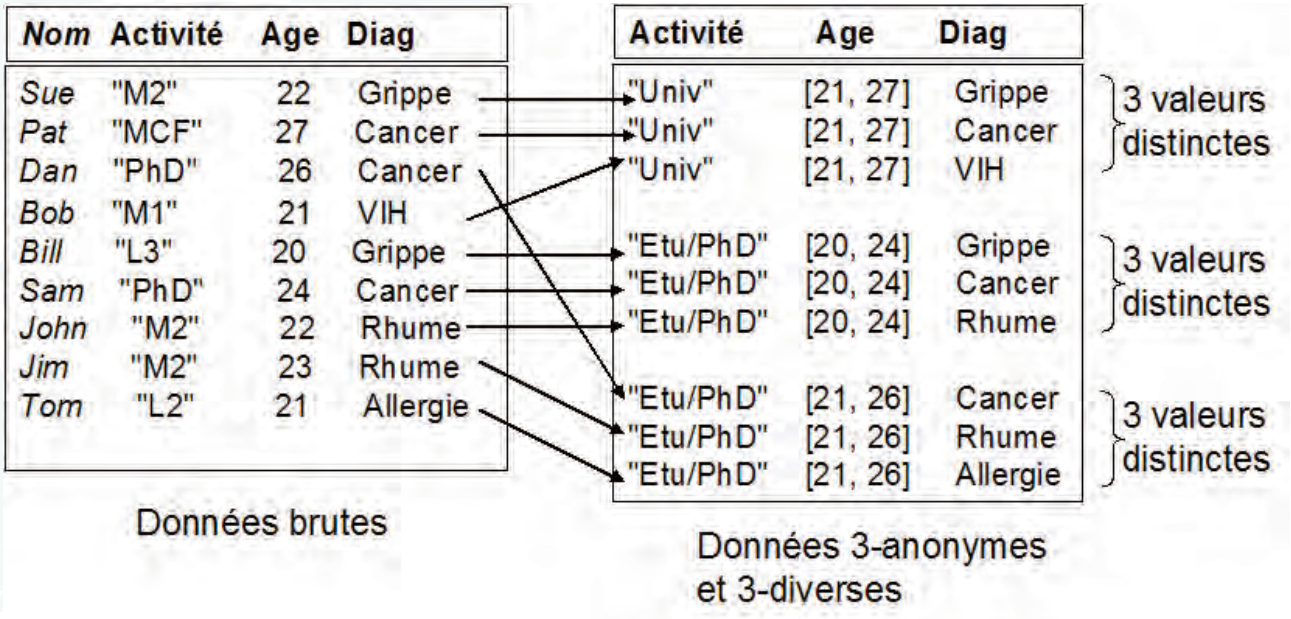


Figure 5. Données *l*-diverses

Cependant, en menant une attaque par croisement du même type que celle de Sweeney, il reste possible de déduire des informations. On voit par exemple dans la Figure 5 qu'on peut déduire qu'un étudiant de 20 ans aura une probabilité 0.33 (soit 1/*k*) d'avoir la grippe, 0.33 d'avoir le cancer et 0.33 d'avoir un rhume… et surtout aucune chance d'avoir une autre pathologie. Si on sait que Bill est la seule personne de la base dans ce cas de figure, alors on peut déduire des informations sensibles à son sujet.

5. On peut généraliser à plusieurs champs sensibles.

La *t*-proximité

Pour essayer de réduire encore l'information qui peut être observée directement, on introduit le modèle de la *t*-proximité, toujours à partir d'un regroupement de données en classes d'équivalences selon le processus du *k*-anonymat. Ce nouveau modèle est basé sur une connaissance globale de la distribution des données sensibles, c'est-à-dire en ce cas les pathologies, pour essayer de faire coller au mieux les valeurs sensibles d'une classe d'équivalence à cette distribution, et ainsi éviter le problème de déduction d'informations soulevé par la *l*-diversité. Le facteur *t* que nous ne détaillons pas ici, indique dans quelle mesure on se démarque de la distribution globale.

Age	Sexe	Département	Pathologie	Nombre d'individus
<45	M	75	Grippe	400
<45	M	75	Rhume	800
>45	M	75	Grippe	500
>45	M	75	Rhume	1000
<35	F	75	Grippe	300
<35	F	75	Rhume	600
>35	F	75	Grippe	600
>35	F	75	Rhume	1200
...				

Figure 6. *t*-proximité

La t-proximité souffre de plusieurs problèmes, le plus important étant sans doute son utilité ! En effet, il parait évident d'exploiter des données *k*-anonymes ou même *l*-diverses pour découvrir des corrélations entre des données appartenant au quasi-identifiant et des données sensibles. Toutefois, le but même de la t-proximité est de réduire au maximum ces corrélations, puisque toutes les données sensibles de chaque classe d'équivalence vont se ressembler ! Ainsi, comme on le voit dans la Figure 6, la *t*-proximité permet surtout de répondre à la question suivante : *comment partitionner mes données de telle sorte que toutes les partitions se ressemblent en termes de distribution ?* Par exemple, si on imagine une base de données nationale sur des pathologies, comment regrouper les départements, classes d'âge et sexes, de telle sorte qu'on ait la même distribution des pathologies dans chaque sous-groupe. On peut s'interroger du jeu de données qui résulte de cette opération lorsqu'on souhaite précisément réaliser une analyse qui fait ressortir les facteurs qui différencient les individus.

La confidentialité différentielle (Différential Privacy)

Nous concluons ce survol des techniques d'anonymisation par la *confidentialité différentielle*, une méthode très en vogue dans les milieux de la recherche en informatique depuis quelques années, car contrairement aux méthodes précédentes, elle est la seule à donner des garanties

formelles, c'est-à-dire des preuves mathématiques, sur la possibilité de borner les informations qu'on peut apprendre sur les individus. Cette méthode introduit un échantillonnage des données vraies (avec une probabilité α), et une génération de données fictives avec une probabilité $\beta >> \alpha$ (mais ces données doivent naturellement rester réalistes...). Les garanties formelles sont cruciales, et permettent de quantifier le risque de ré-identification des *n*-uplets, d'où l'engouement pour cette méthode. En effet, en observant le jeu de données anonymes, l'information qu'on peut obtenir sur le fait qu'un n-uplet soit vrai ou faux est doublement bornée : on n'est jamais sûr qu' un *n*-uplet soit vrai avec une probabilité supérieure à α, ni qu'il soit faux avec une probabilité inférieure à β.

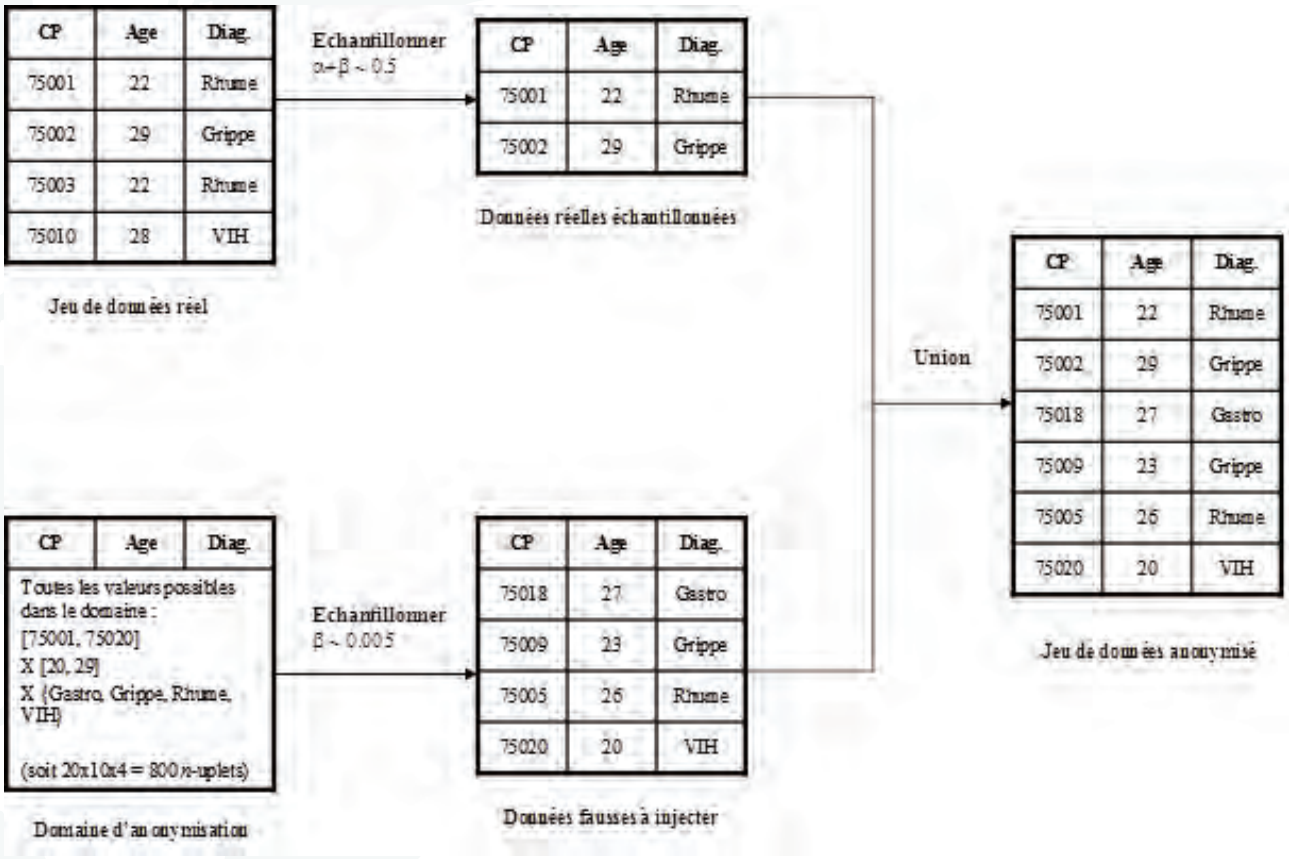


Figure 7. Confidentialité Différentielle

La confidentialité différentielle oblige à calculer un estimateur d'un agrégat que l'on souhaite connaître. Prenons l'exemple du calcul du nombre moyen de malades de la grippe par département, et supposons pour simplifier que les données fictives sont générées de manière équiprobable[6]. On peut estimer le nombre total de malades de la grippe par la fonction suivante, dont l'objectif est de soustraire le bruit (connu) introduit :

$$Nb_{Rhume\ estimé} = \frac{(Nb_{Rhume\ anonyme} - \beta \times Nb_{Rhume\ domaine})}{\alpha} = (2 - 200 \times 0.005)/0.5 = 2$$

Le taux d'erreur peut également être estimé. Cependant, seules certaines fonctions d'agrégation peuvent être calculées avec une erreur bornée : moyenne, nombre total, etc. En revanche, on voit bien que calculer la valeur maximale d'une donnée numérique ne fait pas sens.

6. Il faudrait en réalité baser cette génération sur la distribution réelle des maladies.

Outre cette restriction, le problème principal de la mise en œuvre de la confidentialité différentielle réside dans la vraisemblance des données fictives. Ainsi, cette technique s'applique surtout lorsqu'on cherche à protéger des données de géolocalisation, où il est facile de générer des données fausses « plausibles », et où les fonctions qu'on peut calculer avec cette technique d'anonymisation restent utiles (en particulier la densité et la distance). En revanche, comme on le voit sur l'exemple, il parait plus difficile d'exploiter cette méthode d'anonymat sur des données médicales.

Conclusion

Le problème de l'anonymisation des données, en vue d'assurer leur innocuité tout en permettant une analyse poussée, reste un problème ouvert. Même s'il existe des solutions pour garantir une certaine protection des données d'un individu (confidentialité différentielle), celles-ci sont difficiles à mettre en œuvre dans tous les domaines. Aussi, et même si elles ne permettent pas de réduire totalement le risque, on pourra de manière pratique recourir à des techniques de *k*-anonymat et /-diversité. En revanche, contrairement à ce que peut laisser penser son nom, la pseudonymisation n'est pas une technique d'anonymisation à proprement parler, et ne doit pas être utilisée en tant que telle.
Enfin, notons que de nombreux chercheurs, dont les équipes d'Alex Pentland aux Etats-Unis, ou Philippe Pucheral en France, préconisent d'utiliser la décentralisation du traitement des données, pour lutter entre autres contre le problème du croisement de données à l'insu des personnes concernées. Décentraliser les données signifie que chaque personne concernée par les données gère elle même leur stockage dans un serveur personnel de données (par exemple dans un espace personnel sur le *cloud*, ou dans du matériel sécurisé à son domicile) et autorise ou refuse les traitements.
En effet, puisque l'anonymisation parfaite n'existe pas, il est fondamental que l'utilisateur soit bien informé sur le modèle utilisé, ses performances et ses risques, et qu'il puisse adhérer, en connaissance de cause, aux traitements effectués.

Références

L. Sweeney : " k-anonymity: a model for protecting privacy" *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10(5), 2002.

A. Machanavajjhala , D. Kifer , J. Gehrke , et M. Venkatasubramaniam : "*L-diversity : Privacy beyond k-anonymity*" ACM Transactions on Knowledge Discovery from Data, 1(1):2007.

N. Li, T. Li, S. Venkatasubramanian : " t-closeness: Privacy beyond k-anonymity and l-diversity", *International Conference on Data Engineering*, 2007.

C. Dwork : "Differential Privacy", *International Colloquium on Automata, Languages and Programming*, 2006.

A. Pentland : Social Physics: how good ideas spread : the lessons from a new science - *Penguin Press*, 2014.

T. Allard, N. Anciaux, L. Bouganim, Y. Guo, L. Le Folgoc, B. Nguyen, P. Pucheral, Ij. Ray, Ik. Ray, S. Yin : Secure Personal Data Servers: a Vision Paper, dans *Very Large Data Base*s, 3(1): 25-35, 2010.

Quels droits, et quel pouvoir pour les individus ?

Entretien avec Philippe AIGRAIN

Informaticien et essayiste, co-fondateur de « La Quadrature du Net »

Dans le droit actuel, les droits des individus face aux traitements informatiques reposent sur deux principes de base : la finalité et le consentement. Chaque traitement a une finalité claire ; chaque individu, informé de cette finalité, donne son consentement pour participer au traitement. Ces principes établis dans les années 1970 pour des recueils de données structurées sont mis en difficulté aujourd'hui où le champ des données s'est considérablement étendu, puisqu'il couvre les traces laissées par les internautes ou les opinions qu'ils expriment. Et l'anonymisation n'est pas la panacée. Suffira-t-il pour autant de définir de nouveaux droits ? On a également besoin d'une attention plus forte portée aux architectures informatiques et aux modèles commerciaux, qui conditionnent la réalité de l'exercice d'un droit comme celui du consentement.

Statistique & Société : Quel est le projet de « La Quadrature du Net » ?

Philippe Aigrain : La Quadrature du Net a été créée en 2008 par cinq individus pour défendre les droits fondamentaux et les libertés dans l'espace numérique. Je le souligne : pour nous, il ne s'agit pas de droits ou de libertés spécifiques à l'espace numérique, il s'agit de la mise en œuvre dans cet espace des droits généraux, tels qu'ils sont énoncés notamment dans la Déclaration universelle des droits de l'Homme. Nous sommes maintenant une association loi de 1901, qui intervient soit en réaction à des projets de textes juridiques, soit spontanément, pour proposer des chartes de droits adaptées aux enjeux du numérique, des politiques publiques porteuses de développement humain dans ce domaine, et des conditions de développement des technologies favorisant leur usage social. Nous sommes financés à plus de 50% par les dons individuels, le reste provenant du soutien de fondations.

S&S : Dans quels domaines intervenez-vous ?

PA : Dans deux grands domaines. D'une part, celui des droits fondamentaux : et c'est ce qui nous concerne essentiellement dans cet entretien. D'autre part, la Quadrature du Net intervient sur tout ce qui concerne lles enjeux culturels et sociaux dans le monde numérique.

S&S : Alors vous êtes un groupe de « geeks » ?

PA : Pas du tout. Certes nous avons des compétences en informatique, et nous défendons depuis l'origine le modèle des logiciels libres ; mais nous intervenons surtout sur le terrain juridique, national ou international, à l'aide de membres qui ont des compétences reconnues

en droit et en sciences politiques. Nous sommes consultés par les pouvoirs publics : par exemple, je fais partie des experts membres de la commission de l'Assemblée Nationale qui se penche là-dessus[1]. Quant il y a une saisine du Conseil Constitutionnel sur un sujet de notre compétence, nous nous efforçons d'éclairer son jugement.

S&S : Avez-vous des équivalents à l'étranger ?

PA : Dans le monde anglo-saxon, nos équivalents sont des ONG professionnalisées, comme « Open Rights » ou encore la « Foundation for information policy research » (FIPR) au Royaume-Uni. En Allemagne, je peux citer le « Chaos Computer Club » qui est la plus ancienne association dans le domaine, ou « Digitale Gesellschaft ». L'influence anglo-saxonne est sensible au niveau international : ainsi, au niveau européen, l' « EDRI » (« European digital rights »), établie à Bruxelles, est très liée au monde anglo-saxon Au niveau mondial, l'association « Article 19 »[2] créée en 1987 se consacre à la défense de la liberté d'expression.

S&S : Venons-en donc aux droits individuels dans l'espace numérique. Par la loi « Informatique et liberté » de 1978[3], il est prescrit que « les données sont collectées pour des finalités déterminées, explicites et légitimes » (article 6) et que « un traitement de données à caractère personnel doit avoir reçu le consentement de la personne concernée » (article 7, sous réserve des exceptions énumérées dans cet article). Est-ce que ce cadre légal est toujours pertinent ?

PA : Si - ce qui n'est pas mon cas - on considère les données en tant que « richesse », circulant soit librement selon les modèles d'OpenData, soit par vente comme c'est le cas aux Etats-Unis des données de santé, la finalité ne peut pas être définie au moment de la collecte, car ce sont les futurs exploitants qui vont décider des exploitations possibles. Du coup, la doctrine qui est à l'origine de la CNIL, selon laquelle tout va bien à condition que chaque individu donne son consentement pour une finalité connue, ne fonctionne plus. Pour ma part, je remets en cause cette vision des données : pour moi, l'acceptation des modèles commerciaux fondés essentiellement sur la publicité demande à être critiquée, nous y reviendrons [4]. Mais dans ce contexte, il est clair que le système « finalité-consentement » est affaibli. Par contre, sa pertinence n'a pas disparu dans d'autres contextes : ainsi d'une étude sur le rapport alimentation santé dans laquelle les enquêtés, correctement informés des finalités de l'étude, ont apporté volontairement des informations hautement confidentielles les concernant.

S&S : Alors, est-ce qu'il faut défendre pied à pied cette référence à la finalité ? Ou faut-il imaginer quelque chose de nouveau ?

PA : Il faut protéger les contextes dans lequel le concept de finalité a encore un sens. Mais les finalités sont rarement aussi simples. Souvent, la collecte d'information a plusieurs finalités : et on constate souvent que « les mauvaises chassent les bonnes » : des finalités secondes chassent celles qui avaient été affichées en premier. Ainsi, lorsque dans un service public on s'intéresse aux données pour améliorer la qualité des services aux usagers et pour réduire les coûts, on finit par viser seulement ce dernier objectif. Autre exemple : les données collectées pour accroître la satisfaction du consommateur le sont aussi pour générer des recettes publicitaires, et cet objectif devient prédominant.

Il faut aussi se rendre compte que la notion de « données » s'est élargie à des contextes nouveaux, tout différents de ceux qu'on connaissait en 1978. On appelle maintenant « données »

1. Commission de réflexion sur les droits et libertés à l'âge numérique
2. Ainsi nommée d'après l'article 19 de la déclaration universelle des droits humains.
3. Cette loi a été modifiée à plusieurs reprises depuis, mais reste en vigueur à ce jour ; ses principes sont inchangés.
4. La publicité représente 95% des profits commerciaux de l'exploitation des données ; le reste va aux réductions de coûts par la maîtrise des risques en assurance, par exemple.

des choses très diverses : les « traces » laissées par les individus dans leur parcours d'usage de certains services ; les « informations » capturées par hasard ; ou encore les « communications » entre personnes, et les « expressions » d'opinions laissées sur des forums qui relèvent de tout autres mécanismes de droits… Pour ces catégories de données, le principe de finalité qui a été conçu pour des formulaires précis s'applique malaisément. Il risque même de conduire à protéger des choses qui ne devraient pas l'être, comme des expressions publiques, et à en protéger insuffisamment d'autres.

S&S : L'anonymisation est-elle une solution ?

PA : Le « miracle de l'anonymisation » viendrait nous sauver, et nous autoriser à mettre en circulation soit libre soit commerciale ces données, pourvu qu'elles aient été préalablement « anonymisées » ? Il y a de nombreux spécialistes qui travaillent là-dessus, qui cherchent à identifier les failles de l'anonymat : mais il est difficile de distinguer l'accessoire de l'essentiel. Le problème, c'est qu'une anonymisation parfaite enlève en général tout intérêt aux données ! Le chiffrement est en partie une alternative, mais peut poser lui-même ses problèmes de sécurité. Je ne crois pas que la solution soit : un droit assez contraignant pour les données personnelles, et une libre circulation des données qu'on dit anonymisées.

S&S : Alors, faut-il créer de nouveaux droits ?

PA : Attention : la vraie loi n'est pas toujours où l'on croit. On a tendance à déclarer des droits très forts, et à remettre à des décrets d'application le soin de préciser comment ces droits vont s'appliquer. A ce moment là, c'est le décret d'application, voire le cahier des charges de l'application informatique, qui définit la réalité du droit. Et c'est pourquoi les architectures informatiques et les protocoles d'accès sont cruciaux : le pouvoir des individus passe par là.

S&S : Pouvez-vous donner un exemple ?

PA : Oui, l'exemple du dossier médical personnel (DMP), sur lequel j'ai personnellement travaillé. Je citerai deux aspects. La loi prévoit que le malade aura le pouvoir de contrôler qui aura accès à ses données, de façon précise. Mais, comme c'est impraticable dans le contexte hospitalier, une clause supplémentaire a été introduite : le consentement donné à un membre d'une équipe de soins sera réputé avoir été donné à l'ensemble de cette équipe, y compris les personnels administratifs semble-t-il. Autre exemple : la loi de 2005 formule un principe essentiel : chacun dispose d'un droit d'accès à sa propre information de santé. Mais pour la mise en œuvre de ce droit, à l'ère numérique, en ce qui concerne le DMP, il est pour le moment prévu qu'il faudra envoyer une lettre aux autorités compétentes, pour finalement recevoir une liasse de papier !

S&S : Est-ce que certaines architectures informatiques sont plus propices que d'autres à l'exercice des droits des individus ?

PA : Certainement. Dans le cas du DMP, la décentralisation du stockage des données vers les individus eux-mêmes, avec bien entendu les procédures de sauvegarde nécessaires, donnerait un contrôle bien plus réel. On peut aussi envisager une « décentralisation partielle », selon un modèle intermédiaire qui a déjà été mis en œuvre à grande échelle dans certains pays. Mais il faut réfléchir de façon conjointe aux architectures et aux droits, tant ceux des individus que ceux de la collectivité. Autre exemple : prenez l'hébergement des médias – vos photos, vos films, vos enregistrements musicaux. Il n'y a en réalité aucun problème pour héberger cela soi-même, dans des espaces sous contrôle de l'usager : des dispositifs simples peuvent être développés pour cela. Il existe une société française, OVH, qui est un leader de l'hébergement de sites, mais on préfère citer comme succès des entreprises qui développent des technologies pour la publicité ciblée. L'hébergement de contenus de chacun s'est déplacé vers des sites centralisés

du type Youtube, ou aujourd'hui du cloud d'Apple, dans lesquelles l'usager ne maîtrise plus le devenir de ses objets, en échange de possibilités largement imaginaires dans la plupart des cas, comme de partager ses contenus avec des millions d'autres personnes.

S&S : Alors, qu'est-ce qui pourrait être amélioré pour garantir aux individus un pouvoir effectif sur leurs données ?

PA : La définition juridique des droits, et des sanctions, reste indispensable : si on ne l'a pas, une approche purement technologique et sociétale ne marchera pas, elle restera confinée dans de petits cercles de volontaires désireux d'appliquer des pratiques « vertueuses ». Mais je le répète : la question de l'effectivité des droits est pour moi soumise à celle des architectures informatiques et des modèles commerciaux. Je ne crois pas qu'on puisse progresser par une approche purement juridique. Mais on peut imaginer des architectures et des modèles économiques qui permettent un plus grand contrôle des individus. J'ai déjà parlé des architectures décentralisées : j'ajouterai que le contrôle peut être « logique » sans être nécessairement physique. Ainsi, les créateurs de pages Web chez les fournisseurs d'accès internet gardaient-ils à la fin des années 1990 un contrôle de fait sur leurs contenus, sans rien savoir de leur implémentation physique.

Les conditions du recueil du consentement à l'utilisation des données sont essentielles. Ainsi dans le cas de cookies, il faut souvent donner son consentement pour accéder à la suite du site : l'utilisateur n'a pas vraiment le choix ! De plus, le consentement doit être réversible, et les exigences de contextualisation doivent être respectées (de la même façon que la lumière doit rester allumée au cinéma pendant les publicités ou que les publicités rédactionnelles doivent se différencier clairement des articles dans la presse écrite).

S&S : Et les modèles commerciaux ? Comment jouent-ils sur les droits ?

PA : En ce qui concerne les modèles commerciaux, je considère que c'est la « tache aveugle » de l'action publique, alors que l'on constate ici clairement à quel point certains modèles en vigueur sont pernicieux tant sur le plan social que sur le plan macro-économique. L'industrie du livre, qui n'a jamais reposé sur le modèle de la publicité, continue à représenter la plus grande part de la valeur ajoutée des industries culturelles, alors même que le temps de lecture décroit au profit du temps consacré aux media audiovisuels. Pourquoi la publicité est-elle le revenu fondamental des services sur le web ? Parce qu'il est plus facile de prélever des sommes importantes auprès d'un petit nombre d'annonceurs que de prélever de petites sommes auprès d'un grand nombre d'utilisateurs finaux. Mais ce n'est pas sans conséquence sur l'émergence d'oligopoles. Dans tous les domaines, les modèles commerciaux sont gérés par de la réglementation, une politique de la concurrence : c'est d'ailleurs une politique de ce genre qui a permis le succès de l'Internet fixe en France. Tout le débat sur la « neutralité du Net » est en fait un débat sur les modèles commerciaux qu'on autorise. Le politique a renoncé à faire des choix, même lorsqu'il s'agit d'écarter ce qui est le plus nuisible, le plus inacceptable.

Encore un exemple, celui de la géo-localisation. Actuellement, les services offerts en cette matière reposent sur des services centralisés. Ces services ne permettent pas du tout aux gens de savoir où ils sont eux-mêmes ; mais ils permettent aux entreprises commerciales de savoir où les gens sont, leurs itinéraires, etc. de façon à leur proposer des restaurants, des moyens de transport, etc. C'est Google qui sait où vous êtes, pas vous ! On pourrait renverser la logique : faire en sorte que chacun puisse seul savoir où il se trouve, et décider de diffuser cette information à un prestataire de service s'il veut interroger celui-ci. Autre exemple encore, la RATP et les données issues du « Pass Navigo » : la CNIL a prouvé que les analyses de trafic nécessaires au fonctionnement de la RATP pouvaient se faire avec des données anonymes. Cependant, actuellement, le client doit payer 5€ de plus par mois s'il veut une carte anonyme

empêchant que ses données soient exploitables à d'autres fins. Pourquoi ? Parce que la RATP veut être un acteur du marché de la publicité.

S&S : Si l'essentiel se joue au niveau des architectures informatiques et des modèles commerciaux, que peut faire la puissance publique ?

PA : Les pouvoirs publics peuvent agir par des politiques d'innovation, en appuyant les initiatives qui donnent aux usagers un contrôle physique, ou au moins logique, sur leurs données. En ce qui concerne les modèles économiques et les usages, les politiques répugnent beaucoup à interférer ; mais l'évidence de l'évasion fiscale à grande échelle permet d'espérer que les États vont se mobiliser. Au moins, on pourrait ne pas agir dans l'autre sens : ne pas favoriser systématiquement les solutions centralisées ou seulement compatibles avec le modèle de la publicité. En matière de transmission de fichiers et de droits d'auteur, les pouvoirs publics ont fait une guerre sans merci au partage « peer to peer » en tant qu'architecture, pour favoriser le « streaming » ou le « download » centralisé qui se prêtent tout autant à des usages non autorisés et sont nuisibles à la diversité culturelle.

Il faut aussi reconnaître l'existence de « données d'intérêt public » sur lesquelles on ne peut pas raisonner en termes de propriété. A titre d'exemple, il est inimaginable que les accords de partenariat public-privé ne soient pas rendus publics ! Ces informations sont en effet des « biens communs », à rendre utilisables par tous pour autant qu'on ait assuré la sécurité des données individuelles.

S&S : Se pose aussi le problème de la réutilisation de données d'opérateurs privés pour établir des informations d'intérêt général. Par exemple, l'opérateur de téléphonie mobile Orange exploite ses données sur les communications (métadonnées, données de géo localisation) et permet à d'autres de les exploiter. Il s'agit d'un gisement statistique qui est loin d'être anecdotique. La statistique publique pourrait l'utiliser pour mieux connaître le tourisme ou les déplacements. Quel statut pour de telles données ?

PA : Je ne suis pas favorable à rémunérer les individus qui sont à l'origine de ces données, pour ne pas les engager dans des négociations totalement asymétriques avec des opérateurs infiniment plus puissants qu'eux. En revanche, leur consentement devrait être obtenu sur la base d'un système « d'opt-in » : il faut alors un acte positif de l'usager pour que ses données puissent être utilisées.

S&S : Qui dit consentement dit « biais de sélection » : ce n'est pas bon pour la statistique !

PA : Le consentement a toujours un coût ! C'est une question de volonté, de choix politique. Quand on a mis en place le consentement aux essais cliniques, il y a quelques années, cela a entraîné des difficultés sérieuses, des conséquences multiples : mais on est arrivé à y faire face. L'argument de la commodité ou de l'incommodité de faire autrement sert souvent à fermer le choix politique. Les statisticiens doivent trouver des solutions pour atteindre leurs objectifs en considérant le consentement comme un impératif supérieur.

Une nécessaire exigence éthique

Le point de vue d'un citoyen[1]

Alain GODINOT
Adhérent de la Société française de statistique (SFdS)

L'exploitation des mégadonnées offre de multiples opportunités dans la sphère de l'économie et dans celle de la connaissance. Elle présente aussi pas mal de risques, aussi bien pour les libertés individuelles qu'en termes de dévoiement des savoirs et de manipulation de l'opinion. Comment concilier ces risques et opportunités ? C'est une question à laquelle, vraisemblablement, seule une action continue et de longue haleine permettra de répondre en œuvrant de manière coordonnée au niveau planétaire dans quatre voies : l'éducation dès le plus jeune âge, l'adaptation du droit, l'approfondissement de la déontologie professionnelle et l'élaboration d'une charte universelle de comportement.

Des myriades d'informations de toute nature disponibles à tout instant en tout lieu, sous forme de textes, de photographies ou de documents vidéo ; des moyens techniques performants pour les exploiter à un coût décroissant ; un encadrement juridique encore trop enfermé dans les frontières nationales ; des citoyens pas toujours conscients ni qu'ils répandent à longueur de journée des traces informatiques, ni de l'utilisation susceptible d'en être faite ; des acteurs aux motivations diverses, des plus désintéressées aux plus manipulatrices : le décor est planté pour un monde de possibilités aussi inquiétantes que stimulantes. Comment ouvrir celles-ci tout en se gardant de celles-là ? La question dépasse le cadre des simples mesures de protection individuelle. C'est bien la société humaine tout entière qui est au seuil d'une autre manière de s'informer et de communiquer. Dans ce nouvel univers mal exploré, la protection des libertés individuelles et des libertés publiques doit être éclairée par l'éthique.

Dans son acception moderne, l'éthique s'inscrit délibérément dans le réel : elle vise à faire émerger les règles collectives de comportement individuel les mieux adaptées à un monde changeant de manière à limiter les abus et, in fine, à assurer partout et à tout instant le respect d'autrui (et de soi-même !). On est clairement dans la dimension collective : comment avoir une action responsable vis-à-vis de son environnement au sens large ? Comment se comporter dans la pratique ? Cette éthique générale, ainsi entendue ici comme l'ensemble des règles s'imposant à la conduite des acteurs, va bien sûr se décliner en règles pratiques adaptées à chaque profession et constituant la déontologie de celle-ci.

Chaque être humain est à la fois producteur d'informations le concernant (parfois consciemment, de plus en plus souvent à son insu) et utilisateur d'informations qu'il va chercher lui-même ou qu'il recueille par le canal de médias de toute nature, auxquels il accorde une confiance inégale

1. Point de vue largement inspiré par ma participation au séminaire organisé par la SFdS le 22 mai 2014 (Les enjeux éthiques du BigData – Opportunités et risques)

en fonction de leur caractère plus ou moins institutionnel et de ses propres engagements et croyances. Dans l'exercice de son métier, au service d'une entreprise ou d'une institution, il ne lui est pas toujours possible d'apprécier les usages lointains des travaux qui lui sont confiés. Ce même être humain est la cible ultime aussi bien des recherches scientifiques les plus utiles que des manipulations, subtiles ou grossières mais toujours dangereuses, de l'opinion. C'est donc surtout au niveau de l'individu et à celui des institutions dont il dépend dans sa vie de citoyen, de producteur ou de personne assistée que se jouera le sort des sociétés informatisées. Dès lors, les moyens imaginables pour protéger le corps social et chacun de ses membres doivent être coordonnés entre eux et soutenus par une opinion publique bien informée.

L'esprit critique au cœur de l'enseignement

L'éducation est le premier de ces moyens.. On sait aujourd'hui que ce sont les tout premiers temps de la vie qui déterminent la structuration physiologique, affective et mentale de l'être humain. Il est donc essentiel que, dans une société d'un haut degré de technicité désireuse de maintenir ses valeurs démocratiques, l'éducation des enfants vise à former des citoyens avertis de leur environnement d'information et de communication.

Les enfants deviendront des adultes. Certains exerceront des professions leur donnant des pouvoirs sur autrui : magistrats, avocats, médecins, journalistes, etc. Il est donc indispensable, par exemple, que tous aient une claire notion des limites des instruments dont ils se serviront. Ainsi, une identification reposant sur l'ADN s'exprime en termes de probabilités et non de certitude. Un fichier contient inévitablement des erreurs, et plus il est volumineux, plus il en contient. Ou encore, parmi des occurrences très nombreuses d'événements de toutes sortes, il en est d'infiniment peu probables et qui pourtant se produisent, sans qu'il faille pour autant les ériger en clés de compréhension du monde. Autre exemple encore, l'inéluctable montée du nombre de corrélations qu'on peut observer en croisant à l'aveugle des monceaux d'informations disponibles, sans que ces corrélations signent autant de causalités.

Bref, le développement de l'esprit critique doit être au cœur de la démarche éducative. Un esprit critique tourné, non vers le dénigrement, mais vers l'interrogation rationnelle et le contrôle des informations que l'on reçoit. Cela suppose de bonnes méthodes de travail et de réflexion, au service d'une appréhension scientifique du monde.

À quoi devraient s'ajouter, en matière de protection de soi-même, une bonne connaissance de l'état du droit relatif à l'utilisation informatique des données, éclairée par la pratique de comportements personnels responsables vis-à-vis de soi-même (ne pas délivrer en place publique des informations de l'ordre de l'intime) et vis-à-vis d'autrui (ne pas rediffuser les informations de cette nature imprudemment mises en ligne par d'autres personnes).

L'indispensable protection juridique

Le législateur a bien sûr un rôle central dans l'élaboration des règles de droit et on ne peut que se réjouir des efforts de l'Union européenne pour faire émerger, en matière de traitement des données, un espace juridique de quelque cinq cents millions de personnes. La loi est assurément un rempart nécessaire contre les abus éventuels dans l'utilisation des mégadonnées. Mais le droit est une construction qui a besoin de temps. Il ne s'accommode pas d'approximations : toute entité juridique doit être précisément définie . Au surplus, la portée du droit est loin d'être universelle : en général, son domaine géographique est celui d'un État ; rarement celui de plusieurs États ; exceptionnellement la planète entière.

Face à cela, la technique caracole et se joue des frontières. Imaginative, créative, elle offre sans

cesse de nouveaux champs d'action dans le traitement de l'information. Dans un monde où (heureusement !) tout ce qui n'est pas interdit est permis, la partie est-elle perdue d'avance pour le droit ? Elle l'est s'il s'agit de réglementer dans le détail les innombrables traitements opérés par tous les acteurs. Elle ne l'est pas si l'enjeu est d'édicter des principes de portée universelle et intemporelle, en synergie avec les démarches déontologiques des personnes morales et l'adhésion des personnes physiques à des chartes de comportement. C'est ainsi que certaines initiatives privées veulent convaincre les pouvoirs publics d'adopter une « Déclaration des droits de l'homme numérique » . Comme en écho à cette attente, les autorités européennes de protection des données viennent de produire, en novembre 2014, une déclaration commune allant dans le même sens.

Il reste que l'efficacité du droit passe par son interprétation et son application uniformes aux quatre coins du monde. Pourquoi cette exigence ? Par souci de sécurité. D'une part, pour ne pas ouvrir trop grand l'accès à des comportements délictueux jouant sur des écarts des législations ou des jurisprudences. D'autre part et surtout, afin de sécuriser sur le plan juridique l'action des entreprises et des chercheurs puisant dans les mégadonnées auxquelles ils ont accès. On est loin du compte, tant sont nombreuses et diverses les cultures nationales, dont les législations sont inévitablement l'expression. Si les données individuelles sont un objet marchand ici et un objet à protéger là, une convergence est-elle possible ? L'histoire des dernières décennies et l'émergence de textes de portée supranationale donnent cependant quelques raisons d'espérer que des règles planétaires existeront un jour. Moins par la vertu des peuples et de leurs dirigeants que par les nécessités du commerce international de l'information et la pression d'opinions plus conscientes des enjeux.

Comment agir au mieux dans le respect de principes partagés ? Comment promouvoir une approche éthique de l'utilisation des mégadonnées ? Dans ce domaine comme dans les autres, l'éthique et le juridique ne s'opposent pas (heureusement !) mais ne coïncident pas nécessairement. L'éthique définit des règles qui complètent les règles juridiques. La loi peut ne pas m'interdire une exploitation de données individuelles que je trouverais potentiellement dangereuse dans ses effets sur le corps social. La loi peut m'interdire un traitement informatique que j'entreprendrai néanmoins au nom d'une exigence éthique supérieure.

Beaucoup de professions - le plus souvent des professions règlementées - se sont donc dotées de codes de déontologie professionnelle qui viennent compléter l'appareil juridique en édictant des devoirs et des règles de comportement à l'adresse de leurs membres.

L'importance des bons comportements professionnels

On pourrait imaginer que les gouvernements édictent des règles générales d'éthique, régulièrement mises à jour, qui auraient valeur juridique et auxquelles les codes professionnels devraient se conformer. Au moins dans le domaine de l'utilisation des mégadonnées, cette façon de faire, couplée avec l'inlassable recherche d'une unification internationale des règles, serait sans doute utile. Mais, comme le diable est toujours dans les détails, seule la pratique professionnelle quotidienne peut donner corps aux meilleures intentions. Au sein des entreprises et des institutions, un levier puissant pour l'action est l'image que ces entreprises et institutions veulent avoir dans l'opinion. De ce point de vue, la mise en place d'un label par lequel toute personne morale traitant des mégadonnées afficherait qu'elle s'engage à respecter les règles d'éthique nationales (ou mieux, internationales) serait de nature à améliorer son image . A minima, on peut espérer que ces entreprises ou institutions promeuvent au sein de leur personnel un questionnement préalable à toute utilisation des mégadonnées, comportant par exemple les interrogations suivantes : Que cherche-t-on ? Pour faire quoi ? Quelles données utilisera-t-on ? Quel en est l'origine ? A quels tests de validité les résultats d'une étude seront-ils soumis ? Etc. Sur le terreau d'une éducation qui aura sensibilisé les individus à ces questions, la

déontologie professionnelle a d'autant plus de chances de s'ancrer dans les pratiques que les entreprises et les institutions sauront développer une formation interne et des contrôles - au sens anglo-saxon du terme - tournés vers sa constante amélioration.

Nul n'est censé ignorer la loi. Certes, mais la loi est bien éloignée du quotidien et ses textes d'application sont trop compliqués pour que chaque citoyen, même éduqué au mieux, sache d'emblée comment il doit se comporter en toute circonstance, a fortiori s'il est confronté aux questions inédites que pose l'usage des mégadonnées. Indispensable en ce qu'elle édicte des principes et des règles de comportement, la loi est heureusement prolongée par une culture de déontologie professionnelle. Mais celle-ci se décline, précisément, par profession, voire par institution. Au fil de cette adaptation aux exigences propres à certains métiers, l'individu ne risque-t-il pas de perdre de vue, dans sa pratique quotidienne, des principes plus généraux ? Et la déontologie professionnelle ou d'institution ne risque-t-elle pas de se dévoyer en rites bureaucratiques ou en défense corporatiste ?

L'initiative citoyenne en renfort

Plus profondément, une limite de la loi et des démarches déontologiques tient à ce que l'une et les autres partent d'en haut et s'imposent aux citoyens ou aux salariés, sans que ces derniers les aient nécessairement intériorisées ou, peut-être, comprises dans leurs finalités. Cela ne condamne évidemment pas la démarche descendante, mais il manque au dispositif de protection un mouvement venant en quelque sorte d'en bas, c'est-à-dire des citoyens prenant des initiatives pour affirmer de façon autonome leur attachement à des valeurs, à des principes, qu'ils entendent respecter dans leur vie professionnelle et dans leur vie tout court.

Une forme possible de ces initiatives serait l'élaboration progressive, notamment dans les milieux associatifs, de règles de comportement visant, d'une part, à se protéger soi-même et, d'autre part, à veiller au respect scrupuleux des droits d'autrui. L'adhésion personnelle à ces règles de comportement - de préférence regroupées en un document unique (appelons-le charte, à la fois par commodité et pour lui donner une certaine solennité) - serait de nature à conforter un fonctionnement social plus responsable, plus précautionneux, dans un monde menacé par la réduction de la personne à un avatar informatique . On trouve en France des démarches allant dans ce sens (voir encadré). Elles devraient s'accompagner d'une veille citoyenne sur les mésusages possibles des mégadonnées au regard des exigences éthiques reconnues afin de les dénoncer en utilisant sans retenue les moyens d'alerte offerts par notre monde de communication.

Certains penseurs estiment que, du fait des évolutions techniques entraînées par l'informatique, nous sommes en train de vivre une rupture anthropologique. Si cela est vrai, il convient plus que jamais de conserver la maîtrise de la technique. De ce point de vue, l'existence de règles éthiques et déontologiques encadrées par des législations de portée universelle et soumises au contrôle de citoyens éclairés et responsables est à mon sens une condition nécessaire pour un fonctionnement harmonieux de nos sociétés de l'information.

Charte « Éthique et BigData »

La charte « Ethique & BigData » - qui concerne les « activités générales non réglementées » est une « co-construction d'acteurs académiques et industriels pour faciliter la création, la diffusion et l'utilisation des grands volumes de données (BigData) et ainsi participer à leur valorisation ». On en trouve le texte à l'adresse : http://www.cil.cnrs.fr/CIL/IMG/pdf/CharteEthiqueBigDatav5.pdf

On peut y lire l'engagement suivant :

Par l'adhésion à la présente Charte, je m'engage dans mes activités relatives à l'accès, à l'extraction, à la réutilisation de données dans le cadre d'une activité faisant appel à des jeux de données, à respecter les principes suivants :

- *exercer mon activité dans le respect des principes éthiques, et ce, envers les individus auxquels ces données sont liées, les personnes et entités intervenant à la collecte, la transformation ou la diffusion de ces données ;*
- *garantir autant que possible la traçabilité des données et informer l'ensemble des acteurs qui peuvent avoir à connaître des informations de traçabilité ;*
- *respecter l'ensemble des droits attachés aux données, que ces droits soient liés à l'acquisition ou à la transformation des données ;*
- *respecter les législations afférentes à la diffusion de données, qu'elles soient générales ou spécifiques à la nature des données concernées.*

A cette fin, je remplis la présente Charte Ethique et BigData et m'engage sur les informations qu'elle contient.

Comment modéliser la réussite scolaire en tenant compte de plusieurs niveaux d'analyse ?

Julien DANHIER, Céline TENEY
Chercheurs[1]

L'analyse des résultats aux tests « PISA » a abouti aux conclusions que non seulement les élèves issus de milieux défavorisés réussissaient moins bien que les autres, mais qu'ils réussissaient d'autant moins bien qu'ils étaient scolarisés dans des établissements où l'origine socio-économique moyenne des élèves est moins favorisée. Pour établir ce genre de résultat, les chercheurs ont recours à des modèles statistiques dits « multiniveaux ».

Nous nous plaçons dans le domaine de l'éducation pour présenter un exemple simple d'analyse multiniveaux afin de mettre en lumière l´intérêt de cette méthode (voir Danhier et Martin 2014 pour une application plus développée). À cet effet, nous allons analyser les données issues de l'enquête PISA (« Program for International Student Assessment »). Celle-ci est un projet de recherche mené par l'OCDE qui vise à évaluer "dans quelle mesure les élèves qui approchent du terme de leur scolarité obligatoire possèdent certaines des connaissances et compétences essentielles pour participer pleinement à la vie de nos sociétés modernes" (OCDE 2014: 23). Plus précisément, nous nous limitons aux 1963 élèves de l'enseignement secondaire ordinaire général (excluant ainsi ceux de l'enseignement de qualification et de l'enseignement spécialisé), scolarisés dans 79 écoles de la Fédération Wallonie-Bruxelles. Ces données sont hiérarchiques puisqu'elles reflètent une réalité structurée où les élèves sont regroupés dans des écoles. Afin de tenir compte de cette hiérarchie, les élèves et les écoles ont été utilisés comme unités aux premier et second niveaux.

Pour illustrer ce type d'analyse (voir figure 1), nous présentons, successivement, des modèles multi-niveaux afin d'expliquer la dispersion des résultats en mathématiques aux épreuves PISA 2012 (dont l'échelle s'étend d'environ 235 à 810 points). Le modèle 1 nous permet d'observer l'effet de variables socio-démographiques sur la réussite scolaire. Parmi les variables disponibles, nous nous sommes limités au genre (représenté par une variable dichotomique) et à l'origine socio-économique. Cette dernière est un indice composite couvrant les niveaux d'éducation des parents, leurs situations professionnelles et diverses possessions du ménage. Il se mesure sur une échelle continue allant de -2,7 à 2,5. Dans le modèle 2, l'effet du retard scolaire accumulé par l'étudiant est ajouté. Finalement, il est possible de mesurer l'effet propre de la composition socio-économique de l'école (modèle 3). Celle-ci est l'effet spécifique du regroupement des élèves, qu'il soit dû à des interactions directes entre pairs (discussions, motivations, disputes ou tensions entre différents groupes), à des pratiques du corps professoral (ajustement du style pédagogique ou attentes différentes relatives au groupe d'élèves) et à la qualité de l'école (problèmes de management des ressources humaines ou financement) (van Ewijk & Sleegers

1. Julien Danhier : Groupe de recherche sur les Relations Ethniques, les Migrations et l'Égalité (GERME) Université libre de Bruxelles (ULB) jdanhier@ulb.ac.be
Céline Teney : Centre for Social Policy Research University of Bremen celine.teney@uni-bremen.de

2010). Le tableau 1 reprend les modèles qui seront commentés successivement ci-dessous. Notons toutefois que si ce mode de progression est conseillé par Hox (2010), il n'est pas le seul possible et d'autres progressions peuvent ouvrir à d'autres interprétations.

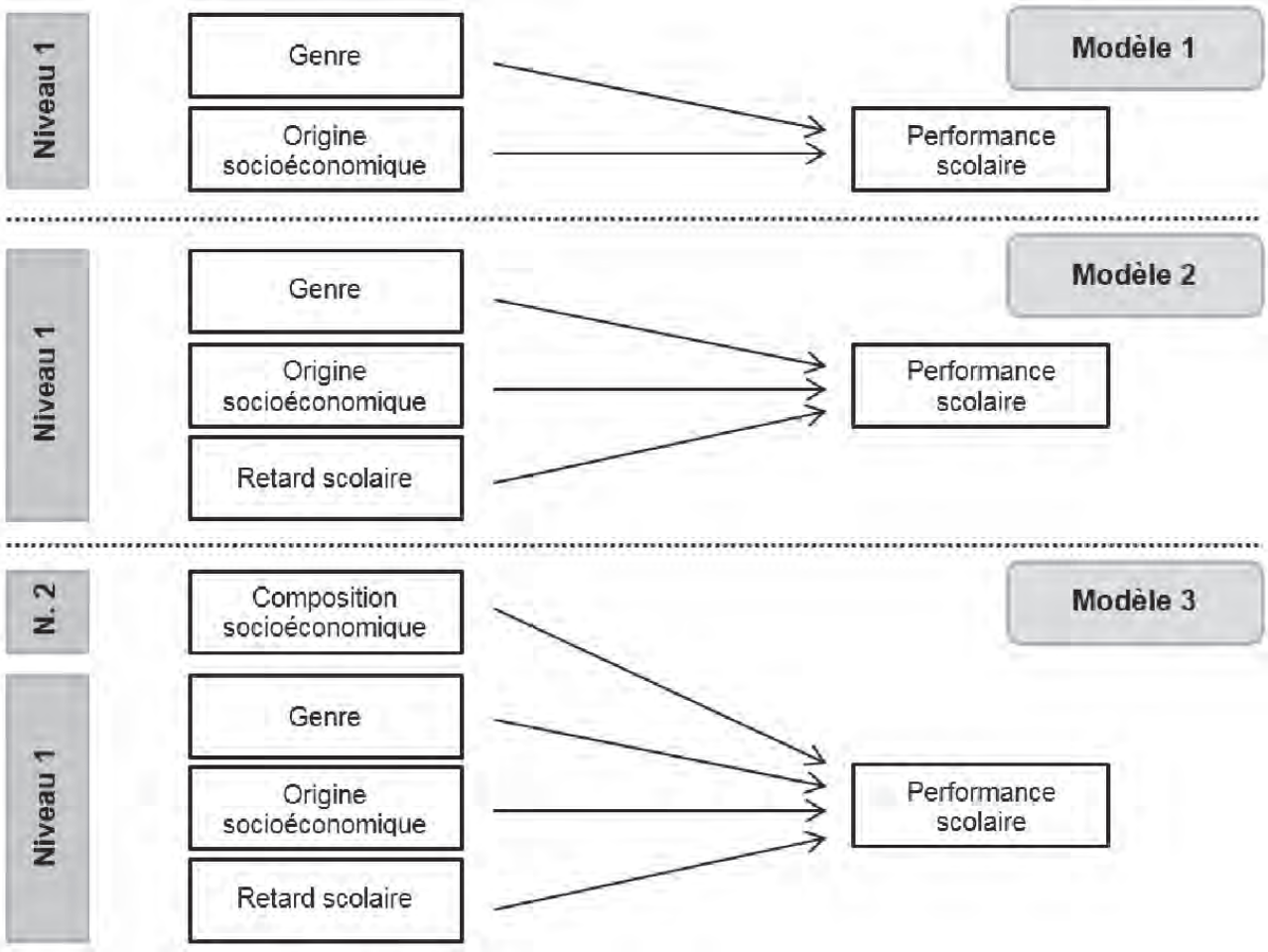


Figure 1 : Représentation des modèles successifs

Il est d'usage de commencer par produire un modèle dit « vide » dans lequel aucune variable explicative n'est spécifiée. La part dite « aléatoire » désigne les variances des « erreurs » à chaque niveau, à savoir la variance des écarts entre les résultats individuels et leurs moyennes dans chaque école (variance « élèves »: 5048) et celle des écarts entre ces dernières et la moyenne générale (variance « écoles »: 2597). Un tel modèle est utile pour deux raisons. Premièrement, il peut servir de base de comparaison pour apprécier la variance expliquée par les modèles suivants. Deuxièmement, il permet de voir comment la variance des scores obtenus par les élèves se répartit entre les niveaux. Ici, une variance « écoles » représente 34 % de la variance totale (qui correspond à la somme des variances « élèves » et « écoles »), ce qui confirme qu'il y a des agglomérats dans nos données. En d'autres termes, 34 % de la dispersion des résultats en mathématiques aux tests PISA est imputable à des différences entre écoles.

Une fois le modèle vide analysé, nous pouvons ajouter des variables explicatives, en commençant par les variables élèves, puis en ajoutant les variables écoles. Le tableau 1 reprend les valeurs prises par l'ordonnée à l'origine et les coefficients de régression (leur erreur standard apparaissant entre parenthèses) sous l'intitulé « part fixe ». Dans le modèle 1, l'ordonnée à l'origine correspond au score pour un individu présentant une valeur de 0 sur les échelles des variables explicatives. Tout comme dans une régression linéaire simple, les coefficients

associés aux variables explicatives représentent l'augmentation des scores d'un élève associée à une augmentation d'un point dans l'échelle de la variable considérée, toute chose étant égale par ailleurs. Un test de Wald permet de vérifier qu'un coefficient est significativement différent de 0. Ici donc, un garçon d'origine socio-économique moyenne obtiendra 525 points en mathématiques. S'il s'agissait d'une fille, elle aurait 19 points de moins. Enfin, si elle était d'origine plus défavorisée (d'un point sur cette échelle qui s'étend d'environ -2,7 à 2,5), elle aurait 21 points de moins. À titre de comparaison, l'OCDE a calculé qu'un écart de 41 points était, en moyenne, équivalent à une année de scolarisation.

Comme dans le cas de la régression linéaire simple, les variables peuvent être ajoutées par blocs successifs afin d'observer d'éventuels effets de médiation. Nous ajoutons donc le retard scolaire dans le modèle 2[2]. L'ordonnée à l'origine représente maintenant, le score d'un garçon d'origine socio-économique moyenne et à l'heure dans son parcours scolaire. L'effet propre du retard scolaire est énorme puisqu'une année de retard est associée à une baisse de 61 points aux tests PISA. Il est intéressant de noter que l'effet de l'origine socio-économique a fortement baissé. Ceci traduit un retard scolaire plus important chez les élèves d'origine défavorisée. On pourra ainsi dire que les scores moindres des élèves d'origine défavorisée traduisent d'une part leur origine, mais également leur présence plus importante parmi les élèves en retard.

L'ajout de variables au modèle réduit la variance des erreurs au niveau des élèves et des écoles, ce qui nous permet de calculer une variance expliquée à chaque niveau, à la manière du R2 de la régression linéaire (une mesure de l'adéquation entre le modèle et les données observées). L'introduction des variables socio-démographiques dans le modèle 1 s'associe à une réduction de la variance résiduelle attribuable aux "élèves" de l'ordre de 5,9% ((5048-4750)/5048). Lorsque l'effet du retard scolaire est pris en compte dans le modèle 2, cette réduction est de l'ordre de 39,0%. La diminution de la variance résiduelle attribuable aux "écoles" est de 25,8% (pour le modèle 1) et 76,6% (pour le modèle 2). Cela signifie que les variables mesurant des caractéristiques individuelles des élèves jouent à la fois au niveau des élèves, mais également au niveau des écoles. On pourra dès lors parler de l'effet du recrutement différentiel des écoles sur la dispersion des résultats des écoles. Il est possible de comparer plus finement les modèles entre eux et d'observer que 50 % de la variance entre écoles est uniquement expliquée par les caractéristiques scolaires des élèves tandis que 20 % de cette variance l'est par l'effet joint des caractéristiques scolaires et non scolaires.

Tableau 1 : Analyse multiniveaux

Paramètres	Modèle 0	Modèle 1	Modèle 2	Modèle 3
Part fixe				
Ordonnée à l'origine	512 (6,96) ***	525 (6,97) ***	569 (4,53) ***	575 (3,92) ***
Variables au niveau des élèves				
Genre (référence : homme)		-19,4 (4,19) ***	-26,1 (3,55) ***	-26,5 (3,54) ***
Origine socio-économique (-)		-20,5 (2,77) ***	-8,4 (2,23) ***	-6,19 (2,32) **
Retard scolaire			-61,4 (2,55) ***	-59,6 (2,47) ***
Variable au niveau des écoles				
Composition socio-économique				-49,4 (6,18) ***
Part aléatoire				
Variance « élèves »	5048 (387)	4750 (367)	3078 (238)	3070 (237)
Variance « écoles »	2597 (549)	1928 (393)	624 (131)	249 (73)
Ajustement du modèle				
R2 « élèves »	0,0	5,9	39,0	39,2
R2 « écoles »	0,0	25,8	76,6	90,4

Niveaux de significativité : non significatif (n.s.), 0,05=*, 0,01=**, 0,001=***

2. Afin de faciliter l'interprétation pour les non-statisticiens, seules les variables dont le 0 n'a pas de sens ont été centrées autour de la moyenne générale, à savoir, l'origine et la composition socio-économique.

Dans le modèle 3, nous ajoutons la composition socio-économique (parfois appelée tonalité) qui est une variable au niveau des écoles mesurant l´origine socio-économique moyenne des élèves fréquentant la même école. Ceci permet de tester si le regroupement d'élèves dans les écoles selon leur origine a un effet sur les résultats scolaires. Comme l'origine individuelle a été précédemment modélisée, il s'agit d'un effet supplémentaire non réductible aux origines socio-économiques individuelles. Nous observons qu'avec un coefficient de -49, la composition socio-économique a un effet significatif sur les résultats scolaires. En d'autres termes, un garçon d'origine moyenne, à l'heure dans son cursus et dans une école dont la composition socio-économique est dans la moyenne aura 575 points en mathématiques. Dans une école parmi les plus défavorisées (1 point sur l'échelle de la composition qui s'étend de -1,2 à 1) ce même garçon, avec la même origine sociale, aura 49 points de moins.
L'étape suivante dans la modélisation consiste à vérifier si l'effet des variables individuelles est identique dans toutes les écoles et à complexifier ainsi la part aléatoire du modèle. Il est possible, par exemple, que l'effet de l'origine socio-économique diffère d'une école à l'autre. Nous avons testé cette spécification dans un modèle non reporté ici et ce modèle apparaît comme moins pertinent pour nos données. Si l'effet de l'origine socio-économique avait été significativement différent d'une école à l'autre, nous aurions également pu tester si l'importance de cet effet dépendait de la composition de l'école (effet d'interaction entre deux niveaux) et s'il était par exemple plus important dans les écoles plus favorisées. Une telle hypothèse n'est pas confirmée ici.

L'analyse multiniveaux

L'analyse de régression multi-niveaux (ou hiérarchique) est une famille d'analyses développées pour tenir compte des agglomérats présents dans les données, soit parce qu'ils existent dans la réalité, soit parce qu'ils sont créés par le chercheur lors de sa collecte. Un exemple typique de données dites « hiérarchiques » est issu du monde de l'éducation où les élèves sont regroupés dans des classes, elles-mêmes regroupées dans des écoles. Sans être exhaustives, deux qualités de ce type d'analyse méritent d'être soulignées et en justifient l'usage.
Premièrement, lorsque ces agglomérats sont présents dans les données, les observations ne peuvent pas être considérées comme indépendantes. Pour reprendre l'exemple de l'éducation, des élèves fréquentant une même classe ou une même école évoluent dans un même contexte scolaire, parfois, avec les mêmes enseignants et ont donc tendance à avoir un profil scolaire et socio-démographique plus similaire que des élèves provenant d´écoles ou de classes différentes. Cette relation doit être statistiquement prise en compte dans l'analyse sous peine d'obtenir des résultats faussement significatifs et l'analyse multi-niveaux est une des méthodes qui permettent de le faire.

Deuxièmement, les caractéristiques, non seulement des individus, mais aussi des agglomérats peuvent avoir une influence. Dans le cas de l'éducation, les caractéristiques des élèves et des écoles peuvent jouer différemment et intervenir à différents niveaux. Ainsi, l´origine sociale d´un élève peut exercer une influence plus ou moins importante sur sa réussite scolaire selon que l´élève fréquente l'une ou l'autre école. De plus, si les caractéristiques individuelles peuvent influer sur la réussite, leur agrégation au niveau des écoles peut également exercer une influence significative. Dit autrement, l´origine sociale moyenne d´une école peut exercer un effet supplémentaire à l´origine sociale individuelle sur la réussite scolaire d´un élève. L'analyse multi-niveaux permet de modéliser des variables à chaque niveau et d'observer leur influence selon le niveau considéré.

Pour aller plus loin

L'analyse multi-niveaux permet des modélisations complexes et son usage ne se limite pas à l'exemple simple que nous venons de présenter. Nous invitons le lecteur à consulter les ouvrages de référence pour appréhender toute l'étendue des analyses possibles (voir notamment Hox, 2010 ; Snijders & Bosker, 2012).

L'analyse multiniveaux n'est toutefois pas la panacée. Elle n'est pas adaptée à toutes les analyses. Il s'agit d'une méthode complexe ayant certains prérequis. Son usage doit ainsi être raisonné et justifié. Nous relevons rapidement certains problèmes auxquels tout utilisateur sera confronté.
L'analyse exige, tout d'abord, un type particulier de données. Premièrement, l'échantillon doit être de taille suffisante, non seulement au niveau des élèves, mais encore au niveau des écoles. En dessous d'au moins 100 écoles, la prudence sera requise, car des simulations ont relevé des biais importants dans le cas d'échantillons restreints (Maas & Hox, 2005). Deuxièmement, il faut être attentif à la définition des niveaux et à l'absence éventuelle de certains. Ainsi, dans nos données, un niveau intermédiaire aurait dû être utilisé : celui des classes. Son absence a pour conséquence une redistribution de la variance entre les deux autres niveaux. Dans le cas d'un niveau supérieur manquant, toute la variance aurait été imputée au niveau le plus haut, à savoir celui de l'école (Opdenakker & Van Damme, 2000).

Au-delà des données, certains choix méthodologiques spécifiques doivent être faits et auront des conséquences sur les résultats obtenus. Le choix de la méthode de centrage est un de ceux-ci. Comme dans le cas de la régression linéaire simple, il est possible de centrer les valeurs d'une variable autour de leur moyenne afin d'en faciliter l'interprétation. Dans l'analyse multiniveaux il est également possible de centrer ces valeurs autour de la moyenne de l'école. Ce choix guidé par la question de recherche n'est pas marginal puisqu'il produira des résultats non équivalents (Enders & Tofighi, 2007).

Malgré sa complexité, la méthode s'est répandue et de nombreux programmes permettent aujourd'hui de l'utiliser de manière relativement intuitive. Parmi d'autres, nous pouvons mentionner des programmes spécialisés, comme MlwiN et HLM (dont les manuels sont aisément accessibles pour les débutants) ou des logiciels plus généraux comme SPSS, SAS, Stata, Mplus ou R.

Références

[1] Danhier, J., & Martin, É. (2014). Comparing Compositional Effects in Two Education Systems: The Case of the Belgian Communities. British Journal of Educational Studies, 62(2), 171 189.
[2] Enders, C. K., &Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychological methods, 12(2), 121 138.
[3] Hox, J. (2010). Multilevel analysis. Techniques and applications (2e éd.). New York : Routledge.
[4] Maas, C. J. M., & Hox, J. (2005). Sufficient Sample Sizes for Multilevel Modeling. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1(3), 86 92.
[5] OCDE, 2014. Résultats du PISA 2012: Savoirs et savoir-faire des élèves. Paris, OECD Publishing.
[6] Opdenakker, M.-C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: an illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. School Effectiveness and School Improvement, 11, 103–130.
[7] Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis. An introduction to basic and advanced multilevel modeling (2nd éd.). London : Sage.
[8] Van Ewijk, r. & Sleegers, P. (2010). Peer ethnicity and achievement: a meta-analysis into the compositional effect. School Effectiveness and School Improvement, 21(3), 237-265.

Mini-débat : jusqu'où va le libre choix des auteurs dans la présentation d'un graphique ?

Avec la participation d'Yves GUIARD

Chercheur LTCI - CNRS et Telecom-ParisTech

et de Thomas PIKETTY

Professeur à l'Ecole d'économie de Paris

Les graphiques statistiques doivent-ils obéir à des règles strictes, ou bien au contraire les auteurs d'une étude disposent-ils d'une certaine liberté pour présenter graphiquement leurs résultats ? A cette question trop générale, il ne peut pas y avoir de réponse unique : mais il est intéressant d'examiner des cas particuliers.

Yves Guiard, chercheur émérite au laboratoire « Traitement et Communication de l'Information » de Télécom-ParisTech, a proposé à la rédaction de Statistique et Société un article où il critique un graphique paru dans le livre « Pour une révolution fiscale » de Camille Landais, Thomas Piketty et Emmanuel Saez. Pour Yves Guiard, un graphique qui représente une variable en fonction de quantiles de la distribution d'une autre variable doit utiliser en abscisses des classes d'égale importance, faute de quoi le graphique risque de donner une idée fausse du phénomène étudié. L'article d'Yves Guiard est publié dans les pages qui suivent.

Nous avons communiqué cet article aux auteurs du livre en question. Thomas Piketty a répondu en leur nom par une réaction qui est reproduite ci-dessous page 79. Il y affirme la liberté des auteurs, dans ce cas comme dans d'autres, à choisir la représentation qui leur semble la plus opportune.

Le lecteur jugera !

La régressivité de l'impôt chez les très hauts revenus : des chiffres incisifs sous le scalpel émoussé de Landais, Piketty et Saez

Yves GUIARD[1]
Chercheur LTCI[2] - CNRS et Telecom-ParisTech

Le taux d'imposition varie selon le niveau de revenu. Le livre de Landais, Piketty et Saez paru en 2011, *Pour une révolution fiscale*, présente un graphique de cette variation utilisant en abscisse des quantiles de revenus. Le graphique des auteurs est malheureusement erroné, suggérant des conclusions fausses. Mais comme les auteurs fournissent leurs données de base, il est aisé d'en construire une représentation graphique exacte.

Landais, Piketty et Saez (LPS) ont publié en 2011 un petit livre abondamment commenté par la presse intitulé *Pour une révolution fiscale*. L'objet du livre, paru un an avant l'élection présidentielle de 2012, était de proposer une réforme du système d'imposition actuel dans lequel la progressivité de l'impôt a été durement mise à mal par l'accumulation, au cours des décennies, des abattements et autres niches fiscales, sans même parler de la dissimulation des ressources. L'ouvrage démarre tout naturellement sur un *état des lieux chiffré*. C'est exclusivement de cet état des lieux chiffré qu'il est question dans le présent article.

LPS ont créé le site http://www.revolution-fiscale.fr/ dans lequel ils offrent gratuitement au public le PDF intégral du livre. Autre particularité remarquable, le site donne tous les tableaux de chiffres qui ont servi à la construction des graphiques du livre. Le lecteur curieux peut donc s'amuser à refaire des figures à partir des chiffres. Or, me livrant moi-même à l'exercice, j'ai constaté avec surprise que LPS ont décrit le système fiscal actuel en s'appuyant sur une visualisation erronée de leurs propres données.

Il faut garder à l'esprit que si les revenus et l'imposition du plus grand nombre sont du domaine public, l'information devient de plus en plus confidentielle à mesure que l'on progresse vers les plus hauts revenus. Comment savoir à quels taux sont réellement imposées les plus riches si personne n'a une idée correcte du montant de leurs revenus ? C'est précisément des travaux de chercheurs comme Landais, Piketty et Saez, trois spécialistes des plus gros revenus français et états-uniens, que pourrait surgir un peu de lumière. D'où l'intérêt spécial de leurs estimations chiffrées. Si ces chercheurs-là, internationalement connus pour avoir mis au point des méthodes indirectes permettant de lever un peu le voile sur l'extrême richesse et sa fiscalité, ne nous disent pas la réalité objective, on se demande qui nous la dira.

Les auteurs ont livré au public une représentation graphique inexacte de leurs chiffres, produisant une image déformée de la progressivité fiscale et une vision franchement édulcorée de la régressivité fiscale. Pour prendre connaissance de leurs précieuses données numériques — dans ce qui suit, elles seront prises, délibérément, pour argent comptant — il va nous falloir

1. yves.guiard@ telecom-paristech.fr
2. Laboratoire Traitement et Communication de l'Information (UMR CNRS N° 5141).

refaire le graphique central du livre.

Une courbe curieuse

Comment le taux d'imposition varie-t-il en fonction du revenu dans la France de 2011? La réponse graphique proposée en page 50 du livre est reproduite dans la Figure 1.

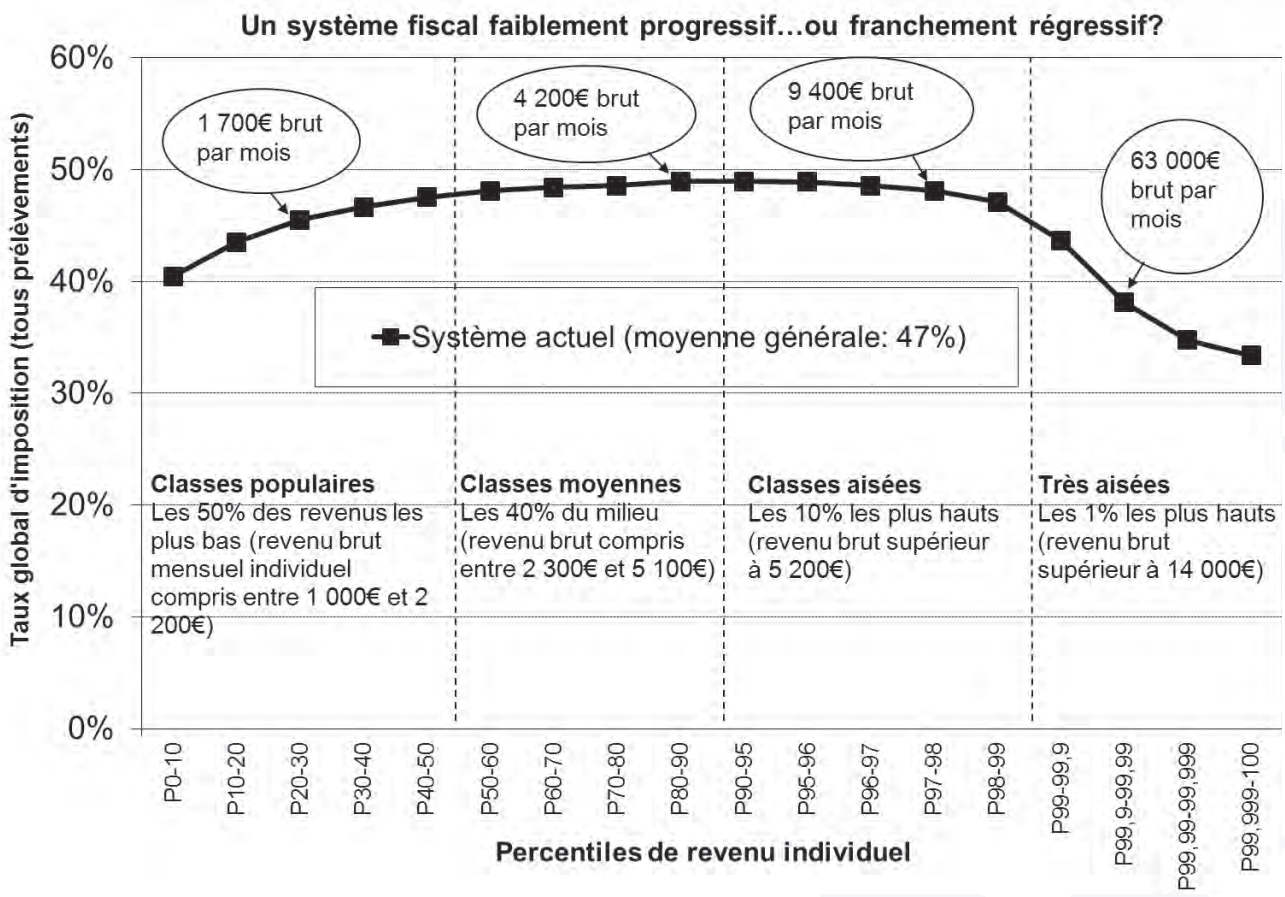


Figure 1. La figure originale de LPS. Légende originale : « *Le graphique montre le taux global d'imposition (incluant tous les prélèvements) par groupes de revenus au sein de la population des 18-65 ans travaillant à au moins 80 % du plein-temps. P0-10 désigne les percentiles 0 à 10, c'est-à-dire les 10 % des personnes avec les revenus les plus faibles, P10-20 les 10 % suivants, etc., P99,999-100 désigne les 0,001 % les plus riches. Les taux d'imposition croissant légèrement avec le revenu jusqu'au 95e percentile puis baissent avec le revenu pour les 5 % les plus riches. Note : Le taux moyen d'imposition des revenus primaires est ici de 47 % (et non de 45 %) car le graphique porte sur la population des 18-65 ans travaillant à au moins 80 % du plein-temps (et non sur la population adulte totale).* »

En débarrassant le graphique de tous ses commentaires et de la partition de l'axe horizontale en quatre classes arbitraires, on obtient la Figure 2 (le centile est expliqué dans l'Encadré 1 ci-dessous).

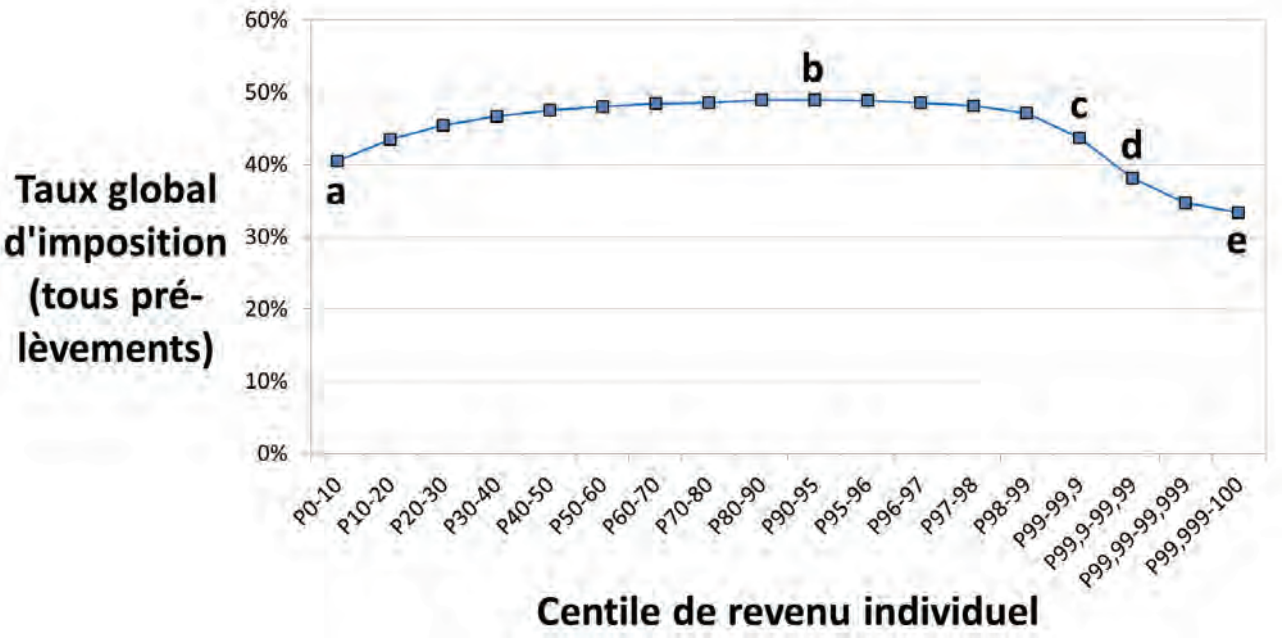


Figure 2. La courbe du livre réduite à son plus simple appareil

1. Le centile de revenu

Imaginons qu'on connaisse le revenu de chacun des 50 millions de contribuables français et qu'on les ait ordonnés du plus petit au plus gros, obtenant ainsi une série de revenus numérotés du N°1 au N° 50 millions. La notion de centile (*percentile* en anglais) est délicate. Le « centile de revenu » désigne, non pas un revenu mais un ensemble de personnes : par exemple, le premier centile de revenu c'est l'ensemble des 500 000 contribuables les plus pauvres, qui représentent 1% de tous les contribuables. Si toute l'information était disponible, on pourrait estimer le revenu de ce centile comme le revenu médian, c'est-à-dire celui du contribuable N° 250 000.

La distorsion qui affecte le graphique de LPS vient de ce que les classes de centiles sont de largeurs inégales. Sur l'axe horizontal de ce graphique la première classe étiquetée « P0-10 » correspond aux dix premiers centiles (10% des contribuables, soit cinq millions de personnes), tandis que la dernière classe, étiquetée « P99,999-P100 », représente une fraction de centile (0,001% des contribuables, soit 500 personnes). En fait sur l'axe horizontal de la figure de LPS la largeur de classe se contracte progressivement à partir de P90-95 selon une loi arbitraire que les auteurs n'expliquent pas.

On devine la source de la bévue : à mesure qu'on progresse vers les plus hauts revenus, le poids *économique* du centile augmente considérablement, selon une loi accélérée fort bien décrite par la fameuse courbe de Lorenz. Mais cela ne justifie évidemment pas qu'un écart horizontal de 0,001% (c'est celui qui sépare les deux derniers points) reçoive dans la représentation graphique le même espacement qu'un écart de 10% (par exemple, entre les deux premiers points). Il faut réaliser que le centile de revenu individuel représenté sur l'axe horizontal du graphique est une variable intrinsèquement *démographique* : il y est question de dénombrement d'individus. Par définition, les 1%, 5% ou 10% les plus riches, quel que soit leur poids économique, représentent exactement 1%, 5% ou 10% de la population. D'où la nécessité de respecter les abscisses de tous les points dans le graphique. Tout indique que LPS ont tout simplement amalgamé le critère économique et le critère démographique.

Le taux d'imposition commence par monter — l'impôt est d'abord *progressif* — et à partir d'un

certain niveau de revenu (l'abscisse du point **b**) il se met à redescendre — l'impôt devient *régressif*. Les auteurs n'ont pas jugé bon de commenter le détail de leur courbe. S'ils l'avaient fait, il leur aurait fallu expliquer deux bizarreries.

La première est que le maximum de la courbe (atteint au point **b**, où le taux d'imposition culmine à 49%) tombe à peu près au milieu de la figure, une conclusion que les auteurs assument pleinement en affirmant que « *le système est légèrement progressif jusqu'au niveau des « classes moyennes»* » (p. 48, guillemets dans l'original). Mais peut-on dire que le groupe de centiles étiqueté P90-95, qui apparaît effectivement au milieu de l'axe horizontal, représente les « classes moyennes » ? ce groupe ne devrait-il pas apparaître à proximité de la marque 100% ?

On s'interroge ensuite sur la bizarre convexité (courbure en U) adoptée par la courbe dans la région sensible des plus hauts revenus (du point **c** au point **e** de la Figure 2). Il n'est pas bien difficile d'interpréter la courbure concave (en U inversé) visible de **a** à **d** : à gauche du point **b**, où l'impôt est progressif, la concavité veut dire tout simplement que la progressivité de l'impôt s'essouffle avec l'augmentation du revenu. Que la courbure reste concave à droite du maximum signifie non moins simplement que la régressivité, une fois installée, s'aggrave. Ainsi, du point **a** au point **d**, la pente de la courbe se réduit progressivement avec le revenu, ce qui se comprend aisément (notamment si l'on songe à l'apparition graduelle puis à l'amplification de l'optimisation fiscale). Le problème c'est cette convexité finale, qui nous dit que dans la région des plus hauts revenus la *régressivité* fiscale se *résorbe avec l'enrichissement*. Qui est disposé à croire cela ? Le graphique nous dit également que tout en haut de l'échelle on serait imposé à un taux d'environ 33% — un résultat peu vraisemblable que les auteurs, curieusement, semblent prendre parfaitement au sérieux lors de leur discussion du bouclier fiscal (page 52). Nous allons voir que tous ces doutes se dissipent avec une représentation graphique adéquate des données.

Retour aux chiffres

On ne peut *voir* véritablement des chiffres qu'une fois qu'on les a convertis en graphique, mais la conversion numérique/graphique obéit à certaines règles. Face aux données numériques de LPS on a deux choix à sa disposition. Le premier consiste à construire une *courbe* uni-variée, où les points vont avoir une ordonnée mais pas d'abscisse. C'est ce qu'ont fait nos auteurs, à ceci près qu'ils n'ont pas prêté attention à la taille variable de leurs classes de revenu. À la différence de LPS, limitons-nous donc aux neuf premières lignes du tableau de LPS (voir annexe), celles où l'intervalle de classes est de 10% — de P0-10 à P80-90 inclus — en laissant donc de côté les neuf lignes suivantes où l'intervalle décroît progressivement. Nous obtenons la courbe de la Figure 3.

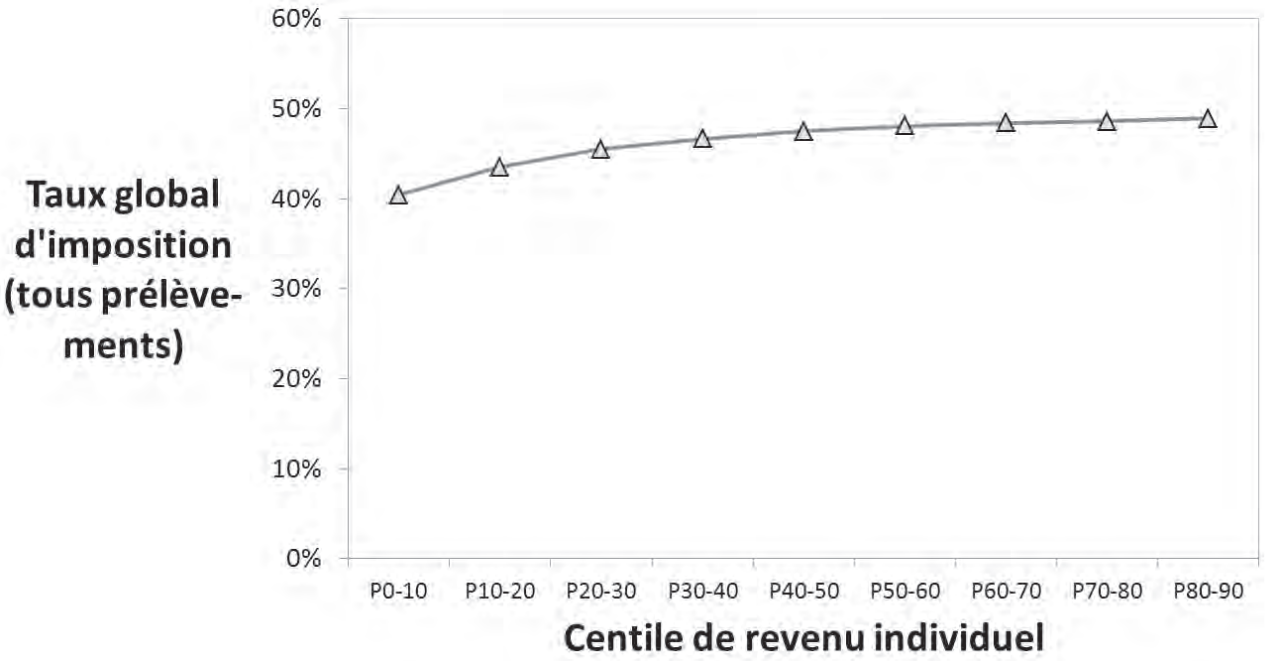


Figure 3. Courbe uni-variée

Bien qu'il ne rende compte que de la moitié des lignes du tableau de LPS, le graphique de la Figure 3 nous délivre une information presque complète, donnant le taux d'imposition pour 90% des contribuables. On constate que la pente montante, tant bien que mal, résiste jusqu'au bout. Au vu de cette courbe, qui traduit les chiffres de LPS sans aucune distorsion graphique, il paraît bien difficile de dire que le passage de la progressivité à la régressivité de l'impôt se produit au niveau des « classes moyennes ». L'impôt est progressif pour au moins 85% des contribuables les moins riches — ou, si l'on préfère, l'impôt est régressif pour au plus 15% des contribuables les plus riches.

L'autre option graphique correcte consiste à tracer une courbe bi-variée, que MS Excel appelle le « nuage de points » (voir annexe). Cette technique, à laquelle on s'étonne que LPS n'aient pas eu recours, va nous permettre de représenter fidèlement la totalité de leurs données et de voir véritablement la relation entre le taux d'imposition et le niveau de revenu. On obtient la Figure 4.

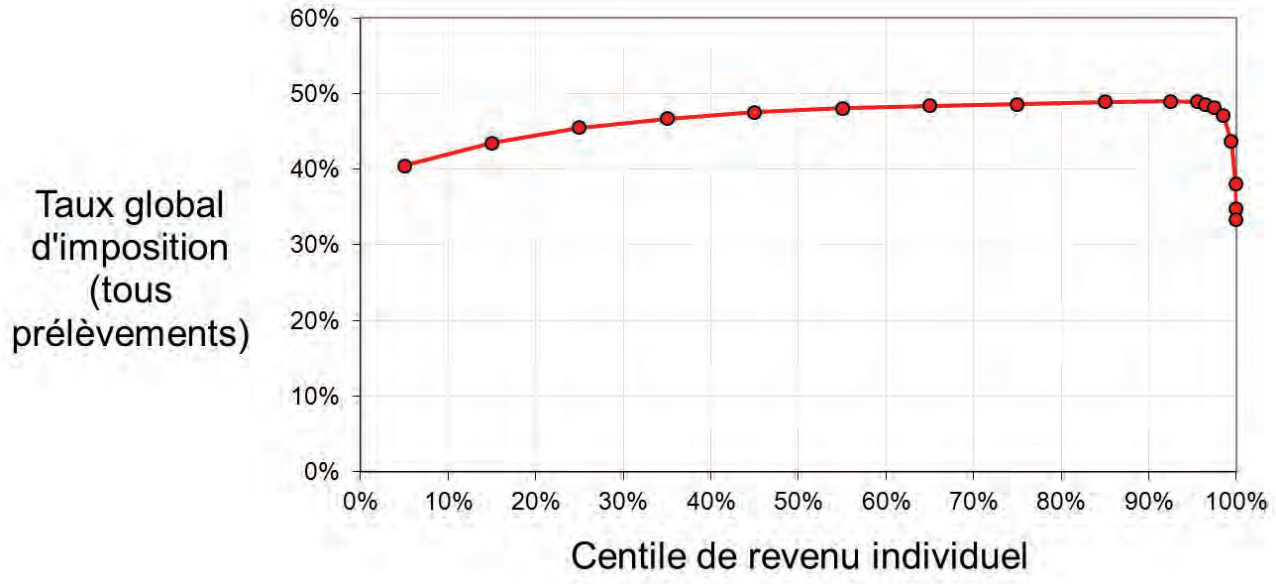


Figure 4. Courbe bi-variée

Chaque point de la Figure 4 possède non seulement son ordonnée (un taux d'imposition), mais aussi son abscisse (une valeur de centile de revenu). À la différence de la figure de LPS, celle-ci utilise un repère Cartésien classique avec une métrique bien définie (en l'occurrence un pourcentage) sur chacun des deux axes. Les points étant maintenant à leurs places, on remarque immédiatement que le maximum **b**, correspondant à la classe P90-95, n'apparaît plus au milieu de l'axe horizontal mais tout à fait à droite, comme il se doit. Notre premier constat est donc que, selon les chiffres de LPS, l'impôt reste progressif à peu près jusqu'au centile 92,5% (pour être tout à fait précis, le taux maximum d'imposition est atteint quelque part entre les centiles 90% et 95%). En d'autres termes, les chiffres de LPS nous enseignent que la régressivité de l'impôt concerne, non pas la moitié des contribuables, mais à peine 10% des contribuables les plus riches.

On constate ensuite que la curieuse convexité finale a disparu, en même temps que la fallacieuse suggestion visuelle d'une stabilisation de l'impôt à 33% au sommet de l'échelle du revenu : la vraie courbe de LPS est concave de bout en bout.

Enfin et surtout, on découvre que la courbe se termine sur un épisode aussi spectaculaire que systématique : tout en haut de l'échelle du revenu la courbe *plonge* littéralement, indiquant, chez les plus fortunés, une désertion massive de la contribution à l'effort collectif (voir l'Encadré 2). S'agissant de l'ultime décile de revenu, la loi quantitative révélée par les données de LPS est très simple : plus le contribuable est riche, plus la régressivité s'accentue. On est loin de la courbe du livre avec son apaisante stabilisation à 33%.

2. Le graphique des auteurs : une surestimation non-négligeable de la progressivité fiscale et une considérable sous-estimation de la régressivité fiscale

Progressivité. la Figure 4 et les chiffres de LPS (en annexe) montrent que jusqu'au maximum d'imposition (c'est-à-dire du centile 5% au centile 92.5% inclus), le taux d'imposition progresse quasi-linéairement de 40.5 à 49.0%. La pente moyenne est donc de (49 – 40.5) / (92,5 - 5) = +0,1 (une valeur dont l'unité est le % par %). Cette valeur montre fort éloquemment à quoi se réduit aujourd'hui la progressivité de notre système fiscal. Si nos auteurs insistent fortement sur ce point, observons seulement que la progressivité est plus proche de zéro dans leurs chiffres que dans leur graphique, lequel nous donne à voir une pente de 0,17. Le lecteur a l'impression, en effet, que les 8,5% d'augmentation du taux d'imposition correspondent à un écart de 50% sur l'axe du revenu. De 0,1 à 0,17, le graphique surestime de 70% le niveau moyen de progressivité de notre système.

Régressivité. À droite du maximum on ne peut guère calculer de pente moyenne car, comme la Figure 4 le montre de manière évidente, la valeur de la pente s'effondre de manière hautement non-linéaire. La régressivité démarre tout doucement avec une pente de 0,02, puis on passe à 0,34, puis à 0,44, puis à 1,0, etc., pour terminer sur une valeur de… 278. Ainsi, entre la progressivité du lot commun, de l'ordre de +0,1, et la régressivité extrême caractéristique des derniers centiles de revenu, de l'ordre de 300, le rapport est de 3000 (en réalité une très probable sous-estimation puisque la non-linéarité doit se poursuivre et s'amplifier au sein de l'ultime classe de revenu de cette étude ; regroupant les 500 plus gros revenus, cette classe est à l'évidence la plus hétérogène).

Conclusion

La raison d'être du présent travail est d'offrir une visualisation sans distorsion des données chiffrées de Landais, Piketty et Saez (2011) sur la progressivité et la régressivité de l'impôt dans ce pays, données précieuses dont le graphique central du livre n'a pas permis aux lecteurs de prendre véritablement connaissance.

Considérée dans l'absolu, la progressivité de l'impôt en France constitue, au vu des données chiffrées de LPS, un phénomène quantitativement négligeable et dont l'importance est plus symbolique qu'économique. C'est vrai à plus forte raison si l'on compare l'ampleur de la progressivité à celle de la régressivité. Sauf à accepter de confondre les millions avec les milliards ou les pièces de 1 centime avec les billets de 100€, le fait quantitatif majeur qui se dégage des données de LPS c'est l'existence, à l'extrémité du système, d'une formidable régressivité de l'impôt. Les auteurs ne nous ont pas signalé la chose mais leurs données numériques parlent d'elles-mêmes : cette régressivité dont on fait si peu état dans la discussion publique est supérieure d'au moins trois ordres de grandeurs à la progressivité au chevet de laquelle on se penche si volontiers.

Notre résultat principal est donc la découverte d'un authentique effondrement contributif au sommet de l'échelle de la richesse dont le graphique des auteurs ne permettait pas de soupçonner l'existence. Mais parler de taux d'imposition c'est parler de ressources déclarées. Par construction, aussi impressionnante soit-elle, la régressivité fiscale révélée par les chiffres de LPS ne tient aucun compte de l'évasion fiscale. Or il en va de l'évasion fiscale comme de l'optimisation fiscale — elle ne concerne que les plus riches et elle progresse de manière accélérée avec le niveau de richesse (Henry, 2012 ; Peillon, 2012 ; Zucman, 2013). Si les chiffres de LPS, auxquels les lecteurs de *Pour une révolution fiscale* n'ont malheureusement pas eu accès, dressent un état des lieux véritablement accablant de la fiscalité française, on gardera à l'esprit que ces chiffres n'offrent probablement qu'une image substantiellement atténuée de la réalité.

La popularisation de résultats scientifiques est toujours un exercice délicat, mais on voit mal quel scrupule pédagogique peut avoir empêché LPS de commenter leurs vraies données converties graphiquement en une courbe correcte. Quand, pour quelque raison, les données d'une recherche scientifique sont hautement sensibles, rien n'interdit de les garder pour soi. C'est tout à l'honneur de LPS d'avoir décidé de communiquer leurs données au grand public, invitant leurs lecteurs à la discussion et à la critique. Mais les chiffres sont les chiffres, et les données numériques ne sont intelligibles que si on les a correctement représentées.

Références

Henry, J.S. (2012). The price of offshore revisited. TJN.
http://www.taxjustice.net/cms/upload/pdf/Price_of_Offshore_Revisited_120722.pdf.

Landais, C., Piketty, T. & Saez, E. (2011). Pour une révolution fiscale. Paris : Seuil.
http://www.revolution-fiscale.fr/

Peillon, A. (2012). Ces 600 milliards qui manquent a la France. Paris : Seuil.

Zucman, G. (2013). La richesse cachée des nations. Paris : Seuil.

Annexe :
Illustrer avec MS Excel les données de Landais, Piketty et Saez (2011)

Voici les données du site http://www.revolution-fiscale.fr/.

	A	B	C	D
	Nom de la classe	Nombre de contribuables dans la classe	Centile moyen de la classe	Taux global d'imposition (tous prélèvements)
1	P0-10	5 000 000	5%	40.5%
2	P10-20	5 000 000	15%	43.5%
3	P20-30	5 000 000	25%	45.5%
4	P30-40	5 000 000	35%	46.7%
5	P40-50	5 000 000	45%	47.5%
6	P50-60	5 000 000	55%	48.1%
7	P60-70	5 000 000	65%	48.4%
8	P70-80	5 000 000	75%	48.6%
9	P80-90	5 000 000	85%	48.9%
10	P90-95	2 500 000	92.5%	49.0%
11	P95-96	500 000	95.5%	48.9%
12	P96-97	500 000	96.5%	48.6%
13	P97-98	500 000	97.5%	48.1%
14	P98-99	500 000	98.5%	47.1%
15	P99-99,9	450 000	99.45%	43.7%
16	P99,9-99,99	45 000	99.945%	38.1%
17	P99,99-99,999	4 500	99.9945%	34.8%
18	P99,999-100	500	99.9995%	33.4%
	Total	50 000 000		

Pour construire leur courbe uni-variée de la page 50 du livre (une « courbe » selon la terminologie de MS Excel) les auteurs ont utilisé les colonnes A et D. Remarquer que la colonne A contient des noms de classes, c'est-à-dire du texte plutôt que des nombres. Il s'ensuit que les points de la courbe vont avoir des ordonnées (les taux d'imposition donnés dans la colonne D) mais pas d'abscisse. La technique est utilisable sans distorsion graphique à condition que les intervalles soient égaux sur l'axe des abscisses. Dans le cas d'espèce on satisfait cette condition si l'on ne représente graphiquement que les neuf premières lignes du tableau, chacun des neuf points obtenus correspondant en effet à cinq millions de contribuables (colonne B).

Incluant les 18 lignes du tableau sans tenir compte du rétrécissement des classes de revenus, la courbe uni-variée de la page 50 du livre distord sévèrement les données. En particulier le maximum d'imposition (soit 49%, 10ème ligne du tableau) tombe à peu près à mi-chemin de l'échelle des revenus, suggérant bien à tort que ce maximum est atteint au niveau des classes moyennes. En réalité ce maximum n'est atteint qu'aux environs du centile 92,5%.

Quand on s'intéresse à la relation entre deux grandeurs numériques, ici le taux d'imposition en fonction du niveau de revenu, la meilleure option est de tracer une courbe bi-variée, ce que MS Excel appelle un « nuage de points », assignant à chaque point du graphique son ordonnée et son abscisse. Mais on a besoin d'un nombre pour repérer le niveau de revenu de chaque classe. Une solution raisonnable, bien qu'imparfaite, consiste à prendre comme indicateur de revenu le centile moyen de chaque classe, soit 5% pour la classe P0-10, 15% pour la classe P10-20, etc. (colonne C, calculée par mes soins). Cette méthode permet de représenter sans distorsion toutes les lignes du tableau de Piketty et collègues.

« Chacun est libre de tracer les graphiques comme cela lui semble préférable »

Thomas PIKETTY

Professeur à l'Ecole d'économie de Paris

Un commentaire très bref. Il existe bien sûr plusieurs façons de représenter les axes sur un graphique. La méthode proposée par Yves Guiard consiste à représenter les centiles en leur accordant sur l'axe horizontal une place proportionnelle à leur part dans la population. Une autre solution serait de leur accorder une place proportionnelle à leur part dans le revenu ou le patrimoine. Nous avons choisi une troisième méthode, qui nous semble-t-il permet d'exprimer de façon plus claire le fait saillant sur lequel nous insistons, à savoir la décroissance des taux effectifs d'imposition au sommet de la hiérarchie des revenus, fait que Yves Guiard ne paraît pas remettre en cause. Chacun est libre de tracer les graphiques comme cela lui semble préférable, le plus important est que cela soit fait en toute transparence : c'est ce que nous faisons, et c'est bien pour cela que toutes les données sont disponibles en ligne.
CL, TP, ES

France, que fais-tu de tes sols ?
Compte rendu d'un Café de la statistique

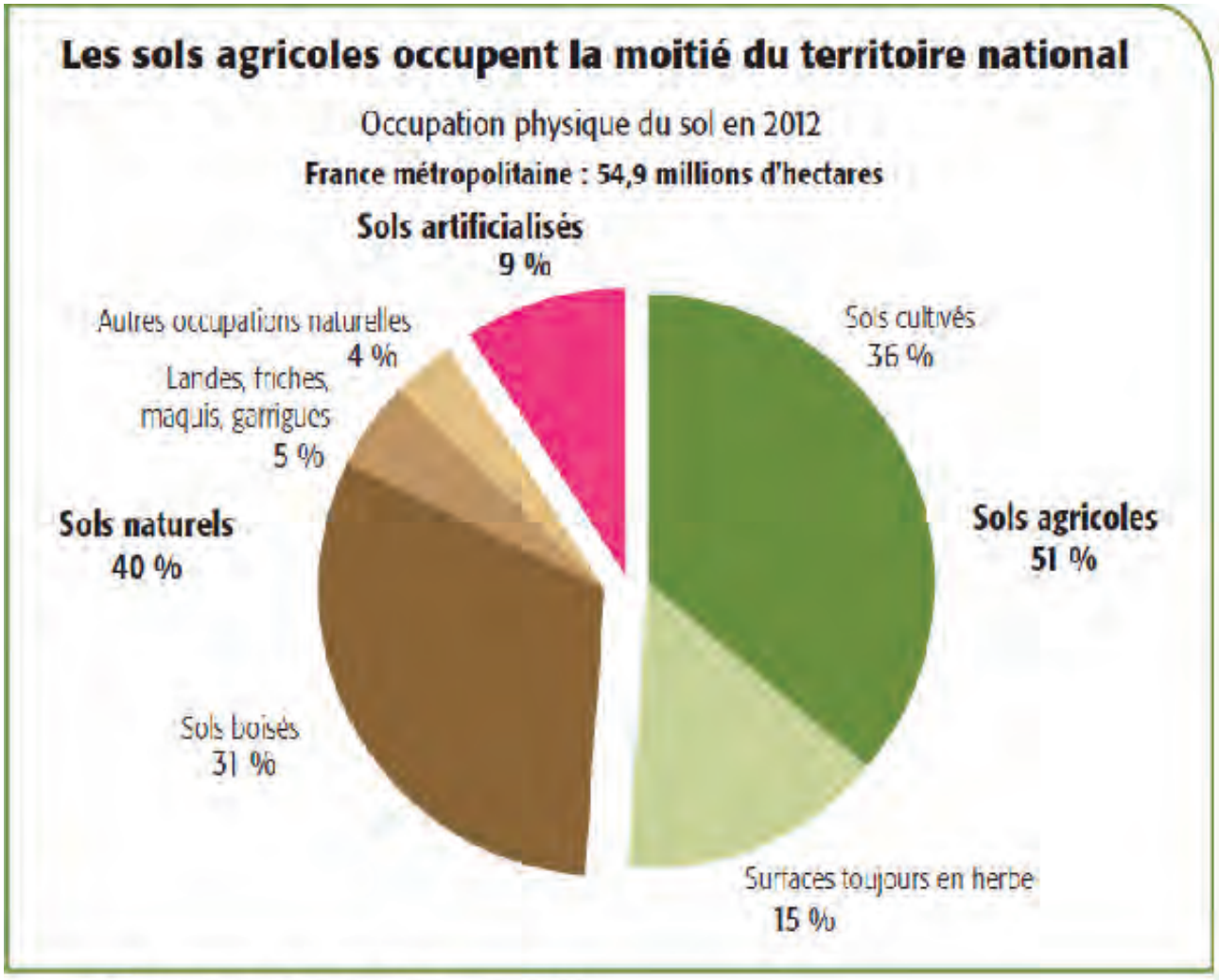
Jean-François ROYER
SFdS

Chaque année, l'emprise des zones urbaines sur le territoire français s'accroît : en dix ans, entre 2000 et 2010, c'est l'équivalent d'un gros département qui est passé de l'état de sol agricole à l'état de sol artificialisé. Cette transformation massive, lourde de conséquences économiques et écologiques, mérite d'être observée pour être régulée. La mesure statistique de l'occupation des sols met en œuvre des outils très divers, depuis les fichiers administratifs jusqu'aux photos satellites. Les résultats diffèrent selon les outils : leur comparaison permet de mieux comprendre quelles sont les tendances établies solidement et quels sont les domaines que l'observation doit encore approfondir.
Un Café de la Statistique s'est tenu sur ce thème en juin 2014 à Paris. Les intervenants étaient Valéry Morard et Michel David, du Service de l'Observation et des Statistiques du Ministère de l'Écologie, du Développement durable et de l'Énergie. Vidéos et compte rendu détaillé sont disponibles sur le site de la SFdS [1]

Vous l'avez appris à l'école : la France métropolitaine a une superficie de 550 000 kilomètres carrés – soit 55 millions d'hectares. Voilà au moins une donnée qui ne change guère ! Mais ce qui change, c'est la manière dont cette superficie est occupée. Davantage de terrains bâtis, davantage de routes, d'un côté ; davantage de forêts d'un autre ; entre les deux, à peu près autant de terres cultivées (mais pas les mêmes), et beaucoup moins de prairies. Les changements sont énormes :

- entre 2006 et 2012, en moyenne 68 000 hectares supplémentaires ont été « artificialisés » chaque année, c'est-à-dire sont passés à l'état de sols bâtis ou revêtus ou stabilisés ; ce rythme équivaut en dix ans à la superficie d'un grand département français ;
- depuis 1950, l'agriculture a perdu entre 80 000 et 90 000 hectares de terre par an, par artificialisation ou par déprise ; la surface agricole utilisée (SAU) n'occupe plus que 51 % du territoire en 2012 contre 63 % en 1950 ;
- ce sont surtout les prairies permanentes, qui représentent ¼ de la SAU, dont la surface a diminué au cours des trente dernières années (baisse de 1/5e).
- la forêt française a doublé de surface au cours des deux siècles passés.

1. Encore que ? Si le niveau des mers s'élève…
2. La déprise agricole désigne l'opération par laquelle un territoire anciennement agricole cesse d'être exploité, sans pour autant être artificialisé.



Source : SSP - Agreste - Enquête Teruti-Lucas 2012

Ces changements se réalisent à travers un marché foncier très actif : en 2013, 330 000 hectares de terres agricoles ont changé de propriétaire. Les zones périurbaines et littorales sont des zones de concurrence entre les usages agricoles et les usages liés à l'urbanisation : le marché peut y connaître des phénomènes de spéculation faisant monter les prix. Ce ne sont pas nécessairement les mauvaises terres, du point de vue agricole, qui sortent de la SAU !

Outre ses impacts sur l'agriculture, l'artificialisation des sols a aussi des effets environnementaux : sur la consommation d'énergie, sur la biodiversité, sur les risques naturels (inondations…). En matière de biodiversité, par exemple, plus que la destruction d'espaces naturels, au demeurant de mieux en mieux protégés, c'est leur fragmentation qui est à craindre, car elle est fatale à la conservation de certaines espèces.

Les enjeux publics de la « consommation d'espace » sont donc grands, tant du point de vue économique que du point de vue écologique. Les pouvoirs publics ne sont pas dépourvus de moyens d'action. Le code de l'urbanisme indique que « le territoire français est le patrimoine commun de la nation[3]» ; diverses lois[4] ont fixé des orientations et organisé la régulation de la consommation de terres agricoles, en s'appuyant sur des outils comme les « SAFER »[5] .

3. Article L110 du Code de l'Urbanisme
4. Par exemple, récemment la loi Grenelle II – loi sur l'environnement n°2010-788 du 12/7/2010, article 14 ; ou la loi « ALUR » du 24/3/2013
5. Sociétés d'aménagement foncier et d'établissement rural

Pour mettre en œuvre ces politiques publiques, une information statistique détaillée sur l'utilisation des sols est nécessaire. Détaillée, parce que la situation peut changer du tout au tout d'une localité à une localité voisine, tant sur le plan économique que sur le plan environnemental. Même un champ de blé ne vaut pas un autre champ de blé ! « Rien ne distingue un blé bio d'un blé conventionnel vu de très haut, et pourtant ils ne rendent pas les mêmes services de régulation écologique[6]» . Cette exigence de précision pose des problèmes redoutables aux statisticiens.

Les outils d'observation sont variés. Outils traditionnels : l'exploitation de fichiers administratifs, comme les fichiers fiscaux de la propriété foncière, et l'interrogation des agriculteurs sur leur utilisation du sol au travers d'enquêtes ou de recensements agricoles. Outil apparu en 1990, renouvelé en 2000, 2006 et 2012 : l'exploitation d'images prises par des satellites. « Corine Land Cover » est un programme de l'Agence Européenne de l'Environnement qui décrit par ce moyen l'occupation bio-physique des sols. Outil récemment perfectionné par les statisticiens publics du Ministère de l'Agriculture : l'enquête au sol. L'enquête annuelle « Teruti-Lucas » échantillonne le sol lui-même : 450 000 « points-échantillons » de 3 mètres de diamètre sont déterminés en bureau, puis visités sur le terrain. Ces différents outils sont complémentaires : par exemple, Corine permet des comparaisons internationales et peut servir de « couche de fond » dans des projets à grande échelle, mais sa nomenclature n'est pas très détaillée. L'enquête Teruti permet de mettre en œuvre une nomenclature plus précise, et produit des résultats assortis d'intervalles de confiance, ce qui constitue un atout incontestable.

Réconcilier au niveau global ces différentes sources ne va pas tout seul : le tableau ci-dessous en témoigne. La tendance générale n'est pas remise en question, mais les évaluations varient du simple au triple. Les statisticiens ont encore du pain sur la planche !

6. Exposé de Valéry Morard

Tableau : Consommations d'espace selon différentes sources pour la période 2000-2012

Source des données	Moyenne annuelle des consommations des surfaces agricoles, sur des périodes variables selon les sources	Moyenne annuelle d'augmentation des surfaces artificialisées, sur des périodes variables selon les sources
1 - Fichiers fonciers de la Direction Générale des Finances Publiques (DGFIP) sur 11 années 2000/2010 sur ensemble du territoire national y compris les DOM – données traitées par la DGFIP	28 910 ha/an	21 200 ha/an
2 - Fichiers fonciers de la Direction Générale des Finances Publiques sur 11 années 2000/2010 – données traitées par le Ministère en charge de l'égalité des territoires et du logement	40 100 ha/an	33 300 ha/an
3.1 - Enquêtes Statistique Agricole Annuelle sur 2000-2010 (série définitive calée sur les recensements agricoles de 2000 et 2010) Données Agreste-Ministère de l'Agriculture	89 300 ha/an	Non renseigné
3.2 - Enquêtes Statistique Agricole Annuelle sur 2010-2011 (série provisoire) Données Agreste-Ministère de l'Agriculture	20 830 ha/an	Non renseigné
4 – Enquêtes Teruti-Lucas sur 6 ans 2006-2012 Données Agreste-Ministère de l'Agriculture	69 200 ha/an	70 300 ha/an
5 – Corine Land Cover sur 6 ans de 2000 à 2006 – Données Ministère de l'Ecologie, du développement durable et de l'Energie – Service de l'observation et des statistiques	13 055 ha/an	14 455 ha/an
6 – Fédération nationale des SAFER sur 12 ans de 2000 à 2012 (à partir des déclarations d'intention d'aliéner ; il s'agit donc du marché foncier, pas de la consommation réelle)	83 981 ha/an	83 981 ha/an

Source : [3] page 25

Références

[1] Compte rendu du Café de la statistique du 10 juin 2014 « La consommation d'espace ». Sur le site web de la SFdS, ce compte-rendu est accompagné de la vidéo des exposés des deux intervenants.

[2] Deux notes du Commissariat général du développement durable (CGDD) parues dans la série "Le point sur" :
*le numéro 10 de 2009 "La France vue par Corine Land Cover, outil européen de suivi de l'occupation des sols" (4 pages)
http://www.developpement-durable.gouv.fr/IMG/spipwwwmedad/pdf/BAT_PointSurCorineBD-1_cle7ca19f-1.pdf
complété pour l'outre-mer par :
*le numéro 89 de juin 2011 "L'occupation des sols dans les départements d'outre-mer http://www.developpement-durable.gouv.fr/IMG/pdf/LPS89.pdf (6 pages)

[3]- Rapport de l'Observatoire national de la consommation des espaces agricole (ONCEA) "Panorama de la quantification de l'évolution nationale des terres agricoles" paru en mai 2014 (126 pages) http://agriculture.gouv.fr/IMG/pdf/140514-ONCEA_rapport_cle0f3a94.pdf

[4] Numéro de mars 2012 de la Revue du CGDD "Urbanisation et consommation de l'espace" (106 pages)
http://www.developpement-durable.gouv.fr/IMG/pdf/Revue_CGDD_etalement_urbain.pdf

[5] Revue « Agreste-Primeur » n°313 – Juin 2014 – « Utilisation du territoire en France métropolitaine : moindres pertes de terres agricoles depuis 2008, après le pic de 2006-2008 » José Masero, Camille Fontes-Rousseau, Didier Cébron, Service de la statistique et de la prospective, Ministère de l'Agriculture