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.