Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

PREMIER MINISTRE

Commissariat général
à la stratégie
et à la prospective

LA NOTE D'ANALYSE

11/2013
N°08

Marie-Pierre Hamel et David Marguerit,
département Questions sociales

Analyse des big data
Quels usages, quels défis ?

La multiplication croissante des données produites et le développement d'outils informatiques permettant de les analyser offre d'innombrables possibilités tant pour l'État que pour les entreprises. Il ne fait aucun doute que le traitement de ces masses de données, ou *big data,* jouera un rôle primordial dans la société de demain, car il trouve des applications dans des domaines aussi variés que les sciences, le marketing, les services client, le développement durable, les transports, la santé, ou encore l'éducation. Par ailleurs, le potentiel économique de ce secteur est indéniable et les retombées en termes d'emploi et de création de richesse seront non négligeables.

Son développement nécessite toutefois de bien comprendre les enjeux qui y sont liés. C'est l'objectif de cette note, qui s'attache à détailler ce qu'est l'analyse des *big data* et présente les usages possibles de ces technologies, qu'il s'agisse de rendre la gestion plus efficace, d'améliorer les services rendus ou de prévenir des phénomènes nuisibles (épidémies, criminalité, etc.). Elle expose les principales difficultés associées à ces usages : garantir la confidentialité et le respect de la vie privée. Enfin, elle montre comment différents pays et entreprises ont d'ores et déjà investi dans ce secteur. ■

LES ENJEUX

L'accroissement des données produites par les entreprises, les particuliers, les scientifiques et les acteurs publics, couplé au développement d'outils informatiques, offre de nouvelles perspectives d'analyses. Ces dernières ont des répercussions importantes en termes de création d'emploi, de recherche et développement ou d'amélioration des services et de leur gestion[1].

Cette note définit tout d'abord ce qu'est l'analyse des *big data.* Elle montre en quoi c'est un phénomène nouveau et à quelles évolutions sociales et techniques il est lié.

Elle détaille ensuite les usages et les possibilités offertes par les analyses de masses de données et leurs applications concrètes.

Puis elle s'attache à signaler les principaux risques associés à ces usages. L'analyse des *big data* peut engendrer des inquiétudes du fait du croisement d'un grand nombre de données. Ainsi, se pose la question des conditions nécessaires au respect de la vie privée et à la sécurité des données.

Enfin, cette note présente les grandes stratégies mises en œuvre par le secteur privé et les gouvernements de différents pays et détermine quelles sont les conditions indispensables au développement de l'analyse des *big data*.

DÉFINIR L'ANALYSE DES *BIG DATA*

Big data et 5 V

Le **volume** de données numériques augmente de manière exponentielle : 90 % de l'ensemble des données aujourd'hui disponibles ont été créées ces deux dernières années[2]. Alors que l'on parlait il y a peu de gigaoctets (10^9 octets), on parle maintenant plutôt de téraoctets (10^{12} octets), de pétaoctets (10^{15} octets), d'exaoctets (10^{18} octets) et même de zettaoctets (10^{21}octets)[3].

Cette augmentation s'explique principalement par les évolutions techniques et d'infrastructures. Entre 1990 et 2011, le pourcentage des utilisateurs d'internet et de téléphones mobiles au niveau mondial est passé respectivement de 0,05 % à 32,7 %[4] et de 0,21 % à 85,5 %[5]. Entre les troisièmes trimestres de 2011 et de 2012, les ventes mondiales de tablettes numériques et de *smartphones* ont pour leur part augmenté de 45,2 %[6]. Ericsson prédit qu'il y aura 50 milliards d'objets connectés (encadré 1) dans le monde d'ici à 2020, contre environ 12 milliards aujourd'hui[7]. Le développement d'applications et de réseaux sociaux liés à ces nouvelles technologies explique aussi la création de données.

L'avènement d'outils comme le *cloud computing* (encadré 1) permet par ailleurs de stocker des données à moindre coût. Globalement, le prix d'un gigaoctet pour un disque dur est passé d'environ 16 USD (12,30 euros) en février 2000 à 0,10 USD (0,07 euros) en août 2010[8].

1. World Economic Forum (2012), *Big Data, Big Impact: New Possibilities for International Development.*
2. Brasseur C. (2013), *Enjeux et usages du* big data. *Technologies, méthodes et mises en œuvre,* Paris, Lavoisier, p. 30.
3. 1 téraoctet représente par exemple 6 millions de livres, 1 pétaoctet représente 2 milliards de photos numériques de résolution moyenne, et 1,8 zettaoctets représentent toutes les informations enregistrées en 2011.
4. Banque mondiale (2013), *World Development Indicators.*
5. *Ibid.*
6. IDC – Press Release (2012), *Smartphones Drive Third Quarter Growth in the Worldwide Mobile Phone Market, According to IDC,* 25 octobre.
7. Ericsson White Paper (2011), *More than 50 Billion Connected Devices.*
8. http://ns1758.ca/winch/winchest.html.

ENCADRÉ 1. ÉLÉMENTS DE DÉFINITION

Big data : Énormes volumes de données structurées et non structurées, difficilement gérables avec des solutions classiques de stockage et de traitement[9]. Ces données proviennent de sources diverses et sont (pour la plupart) produites en temps réel.

Cloud computing : Désigne des prestations à distance – logiciels, stockage de données – physiquement réparties dans des *data centers*[10] et non pas sur le terminal de l'utilisateur.

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

Internet des objets : Désigne les objets connectés à internet qui transmettent des données numériques par le biais de puces radiofréquences (RFID). Ces objets peuvent communiquer entre eux. On les retrouve dans la grande distribution, dans les objets du quotidien (podomètres connectés, domotique, compteurs électriques intelligents), dans les avions, les voitures, dans le monde médical, etc.[11].

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

Tout l'intérêt des masses de données ne réside pas uniquement dans leur quantité. Le volume à partir duquel il est possible de parler de *big data* ne fait d'ailleurs pas l'unanimité. L'analyse des *big data* comprend quatre autres critères que l'on retrouve de façon plus ou moins simultanée : vitesse, variété, véracité, valeur.

La **vitesse** réfère aux délais d'actualisation et d'analyse des données numériques. Les données ne sont plus traitées en différé, mais en temps réel (ou quasi réel). Selon les cas, il est même possible de ne plus stocker les informations, mais de les analyser en flux *(streaming).* Cette rapidité peut être primordiale. Au Canada, en analysant en temps réel les informations sur l'état de santé de bébés prématurés avec un logiciel d'aide au diagnostic (encadré 2), des infections ont pu être détectées vingt-quatre heures avant la manifestation de symptômes visibles[12]. Autre exemple : en croisant les données de capteurs installés sur des éoliennes avec celles relatives à la météo ou aux marées, il est possible d'optimiser leur orientation en temps réel, de mieux prévoir les temps de maintenance, etc. De la même façon, des voitures autopilotées, "communicantes" entre elles et avec l'environnement, sont aussi en développement pour éviter les accidents (données venant de capteurs des voitures, de capteurs sous les routes, données de prévision météo, données historiques/statistiques de densité de trafic, etc.)[13].

Autre caractéristique, les données analysées ne sont plus forcément structurées comme dans les analyses antérieures, mais peuvent être du texte, des images, du contenu multimédia, des traces numériques, des objets connectés, etc. (**variété**). Par exemple, alors qu'il n'existait auparavant pas de systèmes permettant d'analyser automatiquement du texte, il est aujourd'hui possible d'étudier l'état de l'opinion *via* les *tweets (social medias analysis)*, ou encore de proposer une aide au diagnostic en se basant sur la littérature médicale (encadré 2).

Les acteurs du secteur mentionnent aussi la **véracité** ou la qualité des données. Par exemple, comment l'analyste peut-il s'assurer que les données de réseaux sociaux comme *Facebook* ne sont pas des rumeurs ou des diffusions malveillantes ? En 2013, une information d'un faux compte Twitter a dégradé le cours de l'action d'une société cotée au NASDAQ. Le tweet a entraîné l'échange de 300 000 actions en deux minutes, et une baisse de 25 % de la valeur de l'action[14]. Autre exemple : un capteur défectueux utilisé dans un système de conduite assistée peut causer un accident.

À ces quatre V s'ajoute souvent un cinquième, qui désigne la **valeur** qu'il est possible de tirer de ces données, les usages qu'elles produisent[15].

Outils et méthodes

Pour répondre aux besoins provenant d'entreprises comme Google ou Facebook, des logiciels capables de traiter de gigantesques volumes de données structurées et non structurées ont vu le jour, pour la plupart il y a moins de cinq ans. Ces logiciels, souvent *open source* comme Hadoop, peuvent distribuer des données simultanément sur plusieurs serveurs. D'autres logiciels, à

9. Brasseur C. (2013), *op. cit.*, p. 30.
10. *Data center :* en français, "centre de traitement de données". Il s'agit d'un site physique sur lequel se trouvent regroupés des équipements constituants du système d'information d'une entreprise ou d'une institution, que ce stockage soit interne et/ou externe à l'entreprise, exploité ou non avec le soutien de prestataires.
11. Gille L. et Marchandise J.-F. (dir.) (2013), *La dynamique d'Internet. Prospective 2030,* étude réalisée pour le Commissariat à la stratégie et à la prospective, Paris, Études, n° 1.
12. Mayer-Schönberger V. et Cukier K. (2013), *Big Data. A Revolution That Will Transform How We Live, Work, and Think,* Boston, New York, Eamon Dolan, Houghton Mifflin Harcourt, p. 60.
13. http://www.smartplanet.fr/smart-technology/fin-des-embouteillages-lautoroute-du-futur-plus-efficace-a-273-17768/.
14. http://talkingtechno.com/2013/02/26/un-faux-tweet-plombe-le-cours-de-bourse-dune-entreprise-le-web-et-la-folie-des-rumeurs/.
15. GFII (2012), *Dossier de synthèse de la journée d'étude du GFII "Big data : exploiter de grands volumes de données : quels enjeux pour les acteurs du marché de l'information et de la croissance ?".*

l'image de MapReduce, servent à effectuer des calculs en parallèle avec ces données distribuées. On bénéficie ainsi de la puissance de calcul concomitante de multiples serveurs banalisés en *cluster* (secteurs).

Pour améliorer le traitement des données, les logiciels doivent être capables de détecter l'information intéressante : on parle alors de *datamining*[16] (encadré 1). De plus, l'analyste utilise une méthode inductive et non plus déductive : il cherche à établir des corrélations entre plusieurs informations sans hypothèses prédéfinies. Le projet BrainsSCANr a permis la fabrication d'un logiciel qui, en s'appuyant sur 3,5 millions de résumés d'articles scientifiques, fait automatiquement un lien entre des parties du cerveau et certaines maladies. Les corrélations faibles sont les plus intéressantes, car elles représentent celles qui n'ont pas souvent fait l'objet de recherches. Ainsi, un lien entre "migraine" et "striatum" a été mis au jour, l'ordinateur ouvrant de lui-même une nouvelle piste de recherche[17].

Les logiciels, évolutifs, peuvent aussi appréhender l'environnement des données et apprendre des résultats antérieurs. On parle alors de *machine learning* ou d'apprentissage automatique (encadré 2).

ENCADRÉ 2. WATSON-IBM

Watson est un programme informatique d'intelligence artificielle conçu par IBM dans le but de répondre à des questions formulées en langage naturel. Pour développer ce programme, IBM s'est donné un objectif ludique.

Il s'agissait de remporter le jeu télévisé américain Jeopardy! contre des champions. Ce jeu consiste en l'énoncé de réponses pour lesquelles les candidats ont à trouver les questions correspondantes. Le programme informatique doit pouvoir comprendre l'énoncé (en langage naturel) et trouver la question dans un temps de réflexion comparable à celui des humains.

Ce programme d'intelligence artificielle utilise le logiciel Hadoop (voir *supra*) afin de parcourir une grande quantité de contenus (200 millions de pages lors de sa victoire à Jeopardy!) très rapidement (en moins de trois secondes pour Jeopardy!). Watson évalue la probabilité que la réponse qu'il trouve soit la bonne, répondant seulement si celle-ci est jugée assez élevée.

IBM cherche maintenant à commercialiser l'utilisation de Watson. Watson est par exemple utile dans le domaine du diagnostic médical. En analysant les symptômes et les données médicales fournis par un médecin (en langage naturel), et les connaissances emmagasinées (dictionnaires médicaux, littérature scientifique, études de cas, etc.), selon un modèle de *machine learning* qui lui permet d'apprendre des diagnostics antérieurs, Watson propose un diagnostic. Ce dernier est évalué selon une probabilité et le raisonnement est explicité.

D'autres applications sont envisagées : dans les métiers du droit (étant donné l'importance des législations, des réglementations, etc.), l'analyse de dossiers, le conseil financier, etc.

La technologie *big data* est également associée au développement de logiciels capables de rendre intelligibles les résultats – rendu possible par l'émergence de nouveaux outils de visualisation (images, diagrammes, animations).

COMPRENDRE LE POTENTIEL DES ANALYSES DES *BIG DATA*

Simplifier et adapter les services

L'analyse des *big data* permet tout d'abord de mieux écouter les usagers, de mieux comprendre leurs modes d'utilisation des services et d'affiner l'offre. Google Analytics propose par exemple aux entreprises d'améliorer la conception de leur site internet par l'analyse des visites des internautes. Ces applications ont aussi leur utilité dans le secteur public.

Avec l'éducation en ligne (dont les *Massive Open Online Courses* – MOOC), on peut analyser les activités des élèves (temps consacré, façon de suivre les programmes, arrêt-retour dans les vidéos pédagogiques, recherches internet parallèles, etc.) pour améliorer les modes d'enseignement.

Dans le domaine des transports, on modélise les déplacements des populations pour adapter les infrastructures et les services (horaires des trains, etc.). À cette fin, les données provenant des *pass* de transports en commun, des vélos et des voitures "communes", mais aussi de la géolocalisation (données cellulaires et systèmes de localisation par satellites) de personnes ou de voitures, sont utilisées.

Dans un autre domaine, celui de la logistique, à la suite du séisme qui s'est produit en Haïti en 2010, les mouvements de foule ont été analysés à l'aide des données cellulaires pour faciliter la distribution de l'aide[18]. Toujours en Haïti, l'épidémie de choléra qui s'est développée après le tremblement de terre a été mieux combattue grâce à l'étude des déplacements des personnes contaminées[19].

Autre exemple, les analyses peuvent faciliter la recherche d'emploi. Il s'agit de combiner les qualifications des indi-

16. Brasseur C. (2013), *op. cit.*

17. Fischmann S. (2013), *Sciences et technologies de l'information et de la communication.* Big data, *partie 2 : le quatrième paradigme de la science,* Bulletins électroniques États-Unis, n° 336, Ambassade de France aux États-Unis / ADIT.

18. World Economic Forum (2012), *op. cit.*, p. 5.

19. *Ibid.*

vidus avec les offres d'emploi (issues des sites internet de type Le Bon Coin, des sites d'entreprises, des sites administratifs, etc.). Les analyses permettent aussi d'identifier les formations pertinentes, d'anticiper les reconversions, d'adapter la recherche aux besoins du marché[20]. L'entreprise Monster.fr utilise ainsi un logiciel, conçu sur le modèle d'un site de rencontre, qui vise à trouver l'employeur idéal en croisant les compétences, mais aussi les affinités "psychologiques", les caractéristiques des individus embauchés, etc.

L'analyse de masses de données permet également de mieux comprendre les sentiments ou les besoins des citoyens. Pour la campagne de réélection de Barack Obama en 2012, les conseillers ont analysé localement les messages sur Twitter pour adapter en direct le discours du président. Autre exemple, en France, la mairie de Toulouse a demandé en 2013 à la société Apicube d'analyser 1,6 million de documents (*tweet,* Facebook, blogs, forums, etc.) pour mieux connaître les sujets de préoccupation des citoyens. Ces analyses ont toutefois leurs limites en termes de représentativité de la population.

Elles permettent encore d'envoyer à un usager des informations sur des services publics ou privés suivant l'évolution, en temps quasi réel, de sa situation. On peut imaginer qu'une information fournie par un employeur déclenche l'octroi d'une prestation sociale ou en facilite le calcul[21].

L'amélioration des services publics passe aussi par la limitation des demandes de pièces justificatives, la majorité des informations se trouvant déjà dans les masses de données détenues par les institutions publiques. Au Royaume-Uni, pour les demandes de nouveaux permis, l'agence en charge des permis de conduire et de l'immatriculation des véhicules peut récupérer les photographies et les signatures nécessaires dans les données en ligne du service en charge des passeports (si le demandeur a un passeport)[22]. Les analyses permettent également de préremplir les formulaires administratifs en croisant les données.

Améliorer les performances gestionnaires

Les analyses de données massives peuvent accroître la transparence administrative, faciliter l'évaluation des services, assister la prise de décision, ou permettre d'économiser des ressources[23]. Une entreprise peut, par exemple, suivre ses ventes en temps réel pour mieux réapprovisionner ses stocks. De même, une administration publique peut suivre l'activité des agents, le versement de prestations, l'accroissement des demandes, etc.

Les possibilités sont multiples et s'appliquent à une infinité de secteurs. Le *datamining* (encadré 1) de masses de données est par exemple très performant pour détecter les fraudes. En analysant quantité de données sur des populations de fraudeurs, on découvrira certains profils types qui n'étaient pas "visibles". Il est alors possible de mieux cibler les contrôles (l'administration douanière française commence à utiliser ces technologies). Le traitement en temps réel permet aussi de lancer des alertes : l'émission d'une contravention pourrait par exemple permettre de détecter qu'un individu en congé maladie ne devrait pas se trouver dans un département différent de celui où il réside.

En matière d'énergie et de développement durable, les systèmes de compteurs intelligents (électricité, gaz, eau) rationalisent la consommation énergétique. En plus d'offrir aux citoyens la possibilité de mieux contrôler leur consommation, ils permettent de couper à distance, avec l'accord des clients, l'alimentation d'équipements pour éviter les surcharges du réseau.

De même, en analysant les données provenant de capteurs sur les avions et en les associant à des données météo, on modifie les couloirs aériens pour réaliser des économies de carburant, on améliore la conception, la maintenance des avions ou leur sécurité[24].

Prédire et prévenir

L'analyse des masses de données permet plus spécifiquement d'anticiper, avec un certain degré de certitude, des comportements ou des besoins[25]. La société Critéo vend, par exemple, des services de publicités ciblées sous forme de bannières affichées sur les sites consultés. Il s'agit d'analyser une importante quantité d'informations sur les habitudes de consommation des internautes pour établir des corrélations, et ainsi prévoir leurs achats.

De la même façon, l'enseigne américaine Target parvient à identifier les femmes qui attendent un enfant pour leur proposer des produits pour nourrisson. À cette fin, les analystes ont corrélé des millions de données à l'aide de cartes de fidélité de femmes ouvrant une liste de cadeaux

20. TechAmerica Foundation (2012), *Demystifying Big Data: A Practical Guide To Transforming The Business of Government,* p. 15.
21. Hamel M.-P. (2012), "Fraude, indus, non-recours : comment faciliter le juste paiement des prestations sociales ?", *La Note d'analyse,* Centre d'analyse stratégique, n° 306, novembre ; Hamel M.-P. (2013), "Comment utiliser les technologies numériques pour poursuivre l'amélioration des relations entre l'administration et ses usagers ?", *La Note d'analyse,* Centre d'analyse stratégique, n° 317, janvier.
22. Yiu C. (2012), *The Big Data Opportunity: Making Government Faster, Smarter and More Personal,* Policy Exchange, p. 13.
23. TechAmerica Foundation (2012), *Demystifying Big Data: A Practical Guide To Transforming The Business of Government,* p. 12 ; McKinsey Global Institute (2011), *Big Data. The Next Frontier for Innovation, Competition, and Productivity,* 156 p.
24. Jouniaux P. (2013), "*Big data* au service de la sécurité du transport aérien : l'analyse des données de vol", *Télécom,* n° 169, juillet.
25. À ce sujet, voir Siegel E. (2013), *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die,* John Wiley & Sons.

de naissance. Ils ont observé qu'elles commençaient à acheter des crèmes sans parfum à environ trois mois de grossesse, puis certains suppléments alimentaires à un stade de grossesse plus avancé. Ces profils de comportements ont ensuite été étendus à toute la clientèle. Target s'est toutefois retrouvé au cœur d'un scandale, un père ayant découvert la grossesse de sa fille mineure parce qu'elle recevait ces publicités ciblées[26].

Dans le domaine de la santé, il est possible de mieux prévenir certaines maladies ou épidémies, ou d'améliorer le traitement des patients. En analysant les recherches des internautes sur Google, une équipe est parvenue à détecter plus rapidement l'arrivée des épidémies de grippe[27]. Autre exemple, en s'intéressant aux données disponibles sur Facebook, des chercheurs ont détecté les adolescents ayant des comportements à risque pour cibler les campagnes de prévention[28].

Les technologies associées aux *big data* permettent aussi des avancées spectaculaires dans l'analyse du génome humain. Alors qu'il a fallu dix ans et 3 milliards USD (2,3 milliards d'euros) pour réaliser le premier séquençage humain complet, il est maintenant possible d'en réaliser un en quelques jours et pour environ 1 000 USD (760 euros)[29]. Ces connaissances, couplées à d'autres informations, permettent de mieux comprendre l'évolution de pathologies, d'améliorer les mesures de prévention ou encore les protocoles de soins (encadré 3).

ENCADRÉ 3. COHORTE CONSTANCES

La cohorte Constances est une enquête épidémiologique ayant pour objectif de suivre à long terme un échantillon représentatif de 200 000 personnes affiliées au régime général de la Sécurité sociale[30]. Elle est menée en partenariat par l'Institut national de la santé et de la recherche médicale (INSERM), l'université Versailles-Saint Quentin, la Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS), et la Caisse nationale d'assurance vieillesse (CNAV), avec le soutien du ministère de la Santé.

Les personnes enquêtées, âgées de 18 à 69 ans à l'inclusion, ont été sélectionnées par tirage au sort.

Les volontaires doivent répondre annuellement à un questionnaire et passer un examen de santé tous les cinq ans. Ces données sont ensuite appariées tous les ans avec celles de la CNAMTS (SNIIRAM[31] et PMSI[32]), de la CNAV (SNGC[33]) et de l'INSERM (données sur les causes de décès).

L'équipe de recherche souhaite intégrer ultérieurement dans la cohorte des données sous forme d'images, par exemple des résultats de résonance magnétique ou de séquençage du génome. Le croisement des données sur la séquence d'ADN, les pathologies déclarées et l'environnement de vie (type de profession, lieu d'habitation, etc.) permettra notamment d'améliorer la compréhension des mécanismes de l'épigénétique[34].

La prévention des crimes est l'une des applications possibles de l'analyse des masses de données. La police et l'université de Memphis ont développé un programme *(Blue Crush),* maintenant utilisé par de nombreuses villes, qui permet d'identifier les zones et les heures où des délits sont le plus à même d'avoir lieu, afin d'optimiser l'affectation des services[35].

De la même façon, la ville de New York a développé un système pour détecter les logements où des incendies sont le plus susceptibles de se produire (squats, taudis, appartements surpeuplés, découpés en plusieurs "lots", ne respectant pas les règles de sécurité). Il s'agit de croiser quantité de données issues de différents services et agences municipaux portant sur les cinq dernières années : informations sur les logements, procédures d'expulsion, impayés de gaz, d'électricité, de taxes municipales, visites d'ambulance, taux de criminalité, historique des incendies, etc. En appliquant des techniques de *datamining* à ces masses de données, on détermine des profils types de plaintes reçues sur la *"hotline"* de la ville (à propos de nuisances sonores, de troubles du voisinage, sur des suroccupations présumées). Lors de contrôles, ces profils sont le plus susceptibles de déboucher sur la détection de logements où les normes de sécurité ne sont pas respectées. Avant les analyses, 13 % des inspections donnaient finalement lieu à des évacuations pour des raisons de sécurité, contre environ 70 % aujourd'hui[36].

26. Mayer-Schönberger V. et Cukier K. (2013), *op. cit.*, p. 58.
27. Ginsberg J. *et al.* (2009), "Detecting influenza epidemics using search engine query data", *Nature,* n° 457, p. 1012-1014.
28. Moreno M. *et al.* (2012), "Associations between displayed alcohol references on facebook and problem drinking among college students", *Archives of Pediatrics & Adolescent Medicine,* 166(2), p. 157-163.
29. Fischmann S. (2013) *op. cit.*
30. http://www.constances.fr/fr/.
31. Le Système national d'informations inter régimes d'assurance maladie (SNIIRAM) donne des informations sur les remboursements de l'assurance maladie aux particuliers.
32. Le Programme de médicalisation des systèmes d'information (PMSI) : il renseigne les séjours hospitaliers des patients à des fins de remboursements.
33. Le Système national de gestion des carrières (SNGC) regroupe les informations sur la carrière des assurés : salaires, emplois occupés, congé maternité, invalidité, chômage, etc.
34. L'épigénétique est l'expression différenciée des gènes en fonction de l'environnement. En d'autres termes, deux personnes porteuses d'un même gène peuvent, ou non, développer une maladie selon l'influence de l'environnement sur ce gène.
35. http://www.memphispolice.org/blue%20crush.htm.
36. Mayer-Schönberger V. et Cukier K. (2013), *op. cit.*, p. 185-189.

PRENDRE EN COMPTE LES RISQUES

Le développement de l'analyse de masses de données doit s'accompagner d'un questionnement relatif à la protection des données. Le récent scandale "Prism" sur la transmission de données d'utilisateurs d'internet à des fins de surveillance – entre des compagnies comme Google, Yahoo !, Microsoft, Apple, Aol, You Tube, Skype, Paltalk ou Facebook et les services de renseignements américains *(National Security Agency)* – a d'ailleurs donné une attention considérable à cette problématique et pourrait avoir de lourdes conséquences pour le secteur[37].

Respecter la vie privée

Traiter les données à caractère personnel

En France, l'usage des données à caractère personnel est réglementé par la loi "Informatique et Libertés[38]". Dans sa version modifiée, cette loi transpose directement la directive européenne de 1995 relative à la protection des données que le projet de règlement européen du 25 janvier 2012 doit réviser[39]. Pour la loi, la donnée personnelle concerne toutes les informations relatives à une personne physique identifiée ou qui peut être identifiée par des éléments qui lui sont propres.

Pour déterminer si une personne est identifiable, tous les moyens auxquels l'analyste peut avoir accès sont pris en considération. Beaucoup de données peuvent alors permettre cette identification – comme un numéro de téléphone, des données de géolocalisation ou une adresse IP –, et surtout lorsqu'elles sont combinées à d'autres[40]. Leur utilisation peut aussi se faire à la suite d'une anonymisation qui suppose de détruire le lien entre l'information et l'identité.

Cependant, compte tenu des possibilités de croisement des données permises par l'analyse des *big data,* cette anonymisation est quasiment impossible à obtenir. Toutes les données doivent-elles pour autant être considérées comme personnelles[41] ? Leur utilisation étant souvent fort utile, comme dans le champ de la santé, cette voie serait contreproductive si elle conduisait à empêcher toute exploitation de données imparfaitement anonymisées. L'anonymisation a, en tout cas, le mérite de compliquer la tâche de ceux qui seraient mal intentionnés[42].

La loi "Informatique et Libertés" précise par ailleurs que ces données personnelles doivent être collectées et traitées pour des finalités déterminées, explicites et légitimes. Seules les données pertinentes pour un usage défini peuvent donc être collectées. Leur durée de conservation ne doit pas excéder le temps nécessaire à l'atteinte des objectifs pour lesquels elles sont collectées (passé ce délai, prévaut le "droit à l'oubli" ou l'obligation de destruction des données). Même si les données ne sont pas enregistrées mais traitées en temps réel, la loi s'applique.

Dans le cadre des débats européens sur le projet de règlement européen du 25 janvier 2012, la position de la Commission est que la finalité de l'utilisation des données personnelles devrait être clairement établie. Avec l'analyse des *big data,* il est cependant difficile d'anticiper quel usage il en sera fait. La collecte ciblée et le principe de suppression entrent par ailleurs en contradiction avec la nécessité d'un volume de données le plus important possible[43].

La loi "Informatique et Libertés" reconnaît aussi le droit d'être informé de la collecte et de l'utilisation des données, et en particulier de la finalité du traitement, de l'identité du responsable du traitement ou des destinataires des données et des droits dont ils disposent (des droits d'accès, de rectification, d'opposition peuvent être exercés pour motif légitime)[44]. Selon le même principe, la Commission européenne veut demander un consentement explicite par type de données. La législation est cependant allégée lorsque les données collectées sont très vite anonymisées. Le consentement, lorsqu'il est requis par la loi, est en tout cas supposé offrir à la personne un pouvoir. Il désigne toute manifestation de volonté libre, spécifique ou informée[45].

Selon la CNIL et pour plusieurs autorités européennes de protection des données personnelles, cette autorisation donnée dans un contexte spécifique ne correspond pas aux pratiques actuelles de Google. On reproche ainsi à l'entreprise un manque de transparence et d'information envers les usagers concernant l'utilisation de leurs données et la maîtrise de celles-ci[46].

37. http://www.washingtonpost.com/wp-srv/special/politics/prism-collection-documents/.
38. Loi 78-17 du 6 janvier 1978 modifiée.
39. Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, JOCE n° L 281 du 23/11/1995, p. 31. Proposition de règlement du Parlement européen et du Conseil relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (règlement général sur la protection des données), Bruxelles, le 25 janvier 2012, COM(2012) 11 final, 2012/0011 (COD).
40. Levallois-Barth C. (2013), *Big data* et protection des données personnelles : un défi (quasi) impossible ?, *Télécom,* n° 169, juillet.
41. À ce sujet, voir le premier cahier "Innovation et prospective" de la CNIL, *Vie privée à l'horizon 2020,* p. 32-33.
42. Levallois-Barth C. (2013), *op. cit.*
43. *Ibid.*
44. Pas de motif légitime à invoquer dans le cadre de la prospection commerciale.
45. Levallois-Barth C. (2013), *op. cit.*
46. CNIL, Décision n° 2013-025 du 10 juin 2013 de la présidente de la CNIL mettant en demeure la société GOOGLE INC.

Même s'il est explicitement demandé, le consentement peut être biaisé ou manipulé – la personne pouvant être poussée à le donner[47]. Cependant, dans certains cas, par exemple pour détecter plus rapidement des épidémies, le champ du consentement pourrait être élargi aux fins de l'intérêt général[48].

La question du consentement rejoint celle du "détenteur de la donnée". Qui peut avoir accès aux données disponibles sur Internet, comme les données publiques des réseaux sociaux ? D'abord gratuites, la plupart sont maintenant payantes et constituent l'actif principal d'entreprises comme Facebook ou Google, d'où leur opposition à la législation européenne. Alors que les organisations produisaient et utilisaient jusqu'à maintenant leurs propres données, des *data brokers* revendent aujourd'hui les données d'entreprises ou encore de l'État à divers acteurs[49]. On estime ainsi que la société américaine Acxiom, spécialisée dans le recueil et la vente d'informations, et qui a dégagé un revenu de 1,15 milliard de dollars en 2012, posséderait en moyenne 1 500 données sur 700 millions d'individus dans le monde[50].

Traiter les données administratives

En ce qui concerne la collecte et le traitement des données personnelles dont dispose l'administration, l'individu bénéficie là encore du droit d'en être informé et de donner son consentement. Toutefois, dans le cadre des procédures administratives, de nombreuses obligations légales restreignent ses droits. Le consentement n'est par exemple pas requis lorsqu'une autorité administrative est légalement habilitée à obtenir, dans le cadre d'une mission particulière ou de l'exercice d'un droit de communication, la transmission directe d'informations par une autre autorité administrative[51]. L'usager n'a alors aucun recours : c'est plutôt la CNIL qui autorise en amont les échanges.

Le consentement à la transmission d'informations peut, par ailleurs, être difficile à maîtriser : pour un patient, le fait de remettre sa carte vitale à un médecin revient par exemple à consentir à ce que ce dernier ait accès aux données relatives à l'historique de ses remboursements[52]. La transmission de données à des personnes extérieures à l'administration n'est en principe pas permise, mais des exceptions apparaissent, comme l'accès à des données de géolocalisation lorsqu'un usager utilise des services comme Proxima mobile[53]. L'administration peut aussi – dans certains cas spécifiques prévus par une loi – vendre des données, comme les données de carte grise, sauf opposition de l'automobiliste qui peut cocher (s'il la remarque) une case sur son certificat de demande. Intéressants dans le cadre des *big data,* certains de ces usages reflètent une moins bonne prise en compte de la protection des données personnelles par l'administration. La CNIL réfléchit aux réglementations qui pourraient encadrer les analyses.

Plus généralement, à côté des risques liés au traitement des données à caractère personnel, les progrès importants qui peuvent en résulter – pour le traitement de pathologies, l'octroi de droits sociaux ou encore la protection de l'environnement par exemple – doivent être mis dans la balance.

Assurer la sécurité des données

Au-delà des règles de traitement, se pose la question de la sécurité des outils utilisés pour traiter ces données.

Les masses de données sont généralement stockées dans des *clouds* (encadré 1). Toutefois, les créateurs de ces technologies instaureraient régulièrement des *backdoors*[54] leur permettant d'avoir accès à l'ensemble des données stockées. Ainsi, quelles que soient ces données, elles seraient théoriquement accessibles par le fournisseur du service. De plus, le Patriot Act, mis en place aux États-Unis après les attentats du 11 septembre 2001, accorde aux autorités américaines le droit d'accéder directement aux données *cloud* stockées sur les serveurs des sociétés américaines (ou des entreprises étrangères ayant des intérêts économiques dans le pays), et ce quel que soit leur lieu d'implantation.

Pour assurer la sécurité de ces données personnelles, alors que les principaux *clouds* utilisés en France sont étrangers et que le recours à ces technologies de stockage s'est accru de 30 % en 2012[55], l'État français finance, à hauteur de 150 millions d'euros, deux *clouds computing* nationaux dans le cadre d'un partenariat

47. Levallois-Barth C. (2013), *op. cit.*

48. *Ibid.*

49. Dans un rapport publié en 2012, la Commission fédérale américaine du commerce s'est souciée de l'essor de la profession d'*information broker*. Elle demande entre autres à ce que les citoyens puissent avoir accès aux informations que ces "vendeurs d'informations" ont sur eux. Federal Trade Commission (2012), *Protecting Consumer Privacy in an Era of Rapid Change,* mars.

50. http://www.zdnet.fr/actualites/data-brokers-aux-etats-unis-votre-vie-privee-est-en-vente-39789295.htm.

51. Article 6 de l'ordonnance du 8 décembre 2005, loi Informatique et Libertés. Cluzel-Métayer L. (2013), "Les téléservices publics face au droit à la confidentialité des données", *Revue française d'administration publique,* n° 146, 2013/2, p. 405-418.

52. Cluzel-Métayer L. (2013), *op. cit.*, p. 405-418.

53. Proxima mobile, disponible depuis mars 2010, est le portail des services aux citoyens sur terminal mobile. Cet outil permet d'identifier des services d'intérêt général, gratuits et sans publicité, accessibles sur terminaux mobiles, qui cherchent à faciliter la vie quotidienne de tous les citoyens. Diverses applications pour *smartphones,* dont une application du service des impôts, sont par exemple disponibles à partir de ce portail.

54. Les *backdoors* sont des points d'accès confidentiel à un système d'exploitation, à un programme ou à un service en ligne installés par le concepteur.

55. http://blog.markess.fr/2013/05/barometre-markess-des-prestataires-du-cloud-computing-2013.html.

public-privé[56]. Il est essentiel de poursuivre ces initiatives, tout en sensibilisant les acteurs privés aux risques sur les libertés et la vie privée[57].

Pour protéger les données, des recommandations de la CNIL portent par ailleurs sur la sécurité des systèmes d'information et la gestion des risques liés au traitement des données personnelles[58]. On vise à protéger les ordinateurs et les données stockées contre les intrusions, les virus, ou les dommages causés aux données. Le projet de règlement européen en cours d'adoption vise aussi à mettre à la charge des responsables de traitement informatique des obligations, comme celle de prendre en compte la protection des données dès la conception des systèmes *(privacy by design).*

Concernant les échanges d'informations entre administrations, des outils de sécurisation sont mis à disposition par le Secrétariat général pour la modernisation de l'action publique (SGMAP). Le Référentiel général de sécurité (RGS) veut sécuriser, en fixant des règles, les échanges électroniques entre les usagers et les autorités administratives et entre les autorités administratives[59]. La Plateforme d'échange de confiance (PEC) met, pour sa part, en place un système d'intermédiation entre les administrations et les partenaires/usagers pour leur permettre de communiquer en confiance.

Protéger les libertés individuelles

Certains usages des *big data* posent par ailleurs des risques pour les libertés individuelles. Comme cela a été dit plus haut, de nombreux États américains utilisent des logiciels qui permettent de prédire les moments ou les lieux où des crimes sont les plus à même d'être commis. Cela signifie que les individus les plus susceptibles de commettre un crime à un moment et un lieu donnés pourront sans doute être identifiés avec beaucoup de précision[60]. Comment tirer profit de ces connaissances sans mettre en péril les libertés individuelles ?

En permettant de mieux anticiper les comportements, mais aussi l'apparition de maladies associées à des profils génétiques, ces technologies pourraient aussi être utilisées par les services de santé ou les compagnies d'assurance pour refuser des traitements ou des clients, encadrer les comportements des assurés, etc.[61]. Quels garde-fous mettre en place ?

À un autre niveau, la connaissance des comportements permettra sans doute de créer des produits de consommation, mais aussi des produits "culturels" (téléséries, cinéma, etc.) ou des services correspondant, au plus près, aux attentes, aux goûts et aux désirs des individus[62]. Des questions comme celle du libre choix se posent alors.

PROMOUVOIR LES *BIG DATA*

Dans le secteur privé

Les pays qui instaurent une stratégie nationale pour encourager l'analyse des *big data* font figure d'exception. Les sommes investies par le secteur public (voir *infra*) sont d'ailleurs minimes par rapport aux investissements privés. Selon une enquête internationale réalisée en 2012-2013 auprès de 1 217 entreprises ayant un chiffre d'affaires supérieur à 1 milliard USD (759,6 millions d'euros), 643 entreprises ont eu une stratégie *big data* en 2012 ; parmi celles-ci, 7 % ont investi au moins 500 millions USD (379,8 millions d'euros) et 15 % au moins 100 millions (75,9 millions d'euros)[63].

Les États-Unis sont sans doute le pays le plus avancé en termes de stratégie *big data.* En mars 2012, l'administration américaine a annoncé un investissement de 200 millions USD (154 millions d'euros) pour améliorer les technologies (stockage, analyse, collecte des données), accélérer la recherche en science et en ingénierie, renforcer la sécurité nationale, transformer l'enseignement et l'apprentissage, et développer une main-d'œuvre qualifiée dans le secteur.

L'Irlande aspire pour sa part à devenir le pays de référence des technologies *big data.* Le Plan d'action pour l'emploi de 2013 prévoit le développement d'une filière *big data* depuis la formation de la main-d'œuvre jusqu'à la création ou l'installation d'entreprises. S'y ajoute un investissement de 1 million d'euros pour développer un centre de recherche dont les grandes orientations seront définies par un consortium d'entreprises privées.

De son côté, la Commission européenne a entre autres mis en place le programme Big Data Public Private Forum (2012). Sur une période de vingt-six mois, 3 millions d'euros seront au total investis pour la création d'un forum internet visant à définir les grandes orientations en

56. Le projet de *cloud* public Andromède s'est concrétisé en 2012 au travers de la création de *Numergy* et de *Cloudwatt,* deux sociétés nées de partenariats public-privé avec SFR et Bull d'un côté, et Orange et Thales de l'autre.
57. Achiary A., Hamelin J. et Auverlot D. (2013), "Cybersécurité, l'urgence d'agir", *La note d'analyse,* Centre d'analyse stratégique, n° 324, mars.
58. Voir les guides de sécurité sur la méthode de gestion des risques "IL" et sur le catalogue de mesures de sécurité à mettre en place, édités récemment par la CNIL.
59. Premier ministre, ANSSI, ministère du Budget, des Comptes publics et de la Réforme de l'État (DGME), *Référentiel général de sécurité.* Version 1.0 du 6 mai 2010.
60. Mayer-Schönberger V. et Cukier K(2013), *op. cit.*, p. 157-163.
61. Reynaudi M. et Sauneron S. (2012), "Médecine prédictive : les balbutiements d'un concept aux enjeux considérables", *La note d'analyse,* Centre d'analyse stratégique, n° 289, octobre.
62. Voir par exemple au sujet de la création de séries télévisées : http://www.salon.com/2013/02/01/how_netflix_is_turning_viewers_into_puppets/.
63. Tata Consultancy Service (2013), *The Emerging Big Returns on Big Data. A TCS 2013 Global Trend Study.* http://www.lesechos-conferences.fr/data/classes/produit_partenaire/fichier_5183_540.pdf.

matière d'analyse des *big data* au sein de l'Union européenne. Ce projet veut fournir une plateforme de discussion sur l'émergence d'une économie de la donnée pour l'industrie, la recherche et les décideurs politiques.

La France est entrée très récemment dans la course à l'analyse des *big data*. Dans le cadre des investissements d'avenir, sept projets traitant des *big data* ont été sélectionnés pour recevoir 11,5 millions d'euros, quatre autres projets sont en cours d'instruction et un nouvel appel à projets devrait avoir lieu avant la fin de 2013. Les projets financés rassemblent une grande diversité d'acteurs : concepteurs de systèmes informatiques, éditeurs de logiciels, intégrateurs de technologies, laboratoires de recherche et un nombre important de *start-up*. Les retombées attendues sont multiples, touchant au marketing ou à la recherche génétique.

En parallèle, la ministre déléguée auprès du ministre du Redressement productif chargée des petites et moyennes entreprises, de l'innovation et de l'économie numérique a créé une mission ayant pour but de définir les grandes orientations nécessaires à l'émergence d'une filière *big data*. Pour la période 2013-2018, cette mission préconise de créer un "incubateur" parisien avec un investissement de 300 millions d'euros provenant de fonds publics-privés. Elle estime que la valeur générée par cet investissement pourrait atteindre 2,8 milliards d'euros et créer dix mille emplois directs sur la période[64]. Environ cent *start-up* spécialisées dans les applications *big data* doivent ainsi être financées. Plusieurs autres travaux insistent sur le développement de la filière en France. Citons le rapport de la Commission innovation 2025 qui fait du développement des analyses *big data* l'une des "sept ambitions pour une France innovante et dynamique[65]", mais aussi le rapport *La nouvelle France industrielle* présenté par Arnaud Montebourg en septembre 2013, qui désigne les *big data* comme l'un des 34 plans prioritaires[66].

La France investit également dans le développement d'outils spécifiques de stockage de données (voir *supra*) ou encore dans la création de moteurs de recherche nationaux comme Quaero[67]. La majorité des données créées sur internet sont en effet détenues par des entreprises étrangères, et principalement états-uniennes. En France, la part de marché de Google sur les moteurs de recherche était estimée à 90,9 % en avril 2013[68]. Cette situation est problématique, d'une part en termes de compétitivité et de création de "richesse", d'autre part en termes de relations stratégiques : il peut être préoccupant que des acteurs étrangers et/ou privés en sachent davantage que l'État français quant aux habitudes de vie, aux comportements, aux préoccupations, etc. des citoyens.

Précisons finalement que la demande en spécialistes de l'analyse de données massives est en plein essor. Il n'existe pas d'estimation nationale officielle, mais, à titre indicatif, l'institut Mc Kinsey Global estime que les besoins en analyses de masses de données induiront, aux États-Unis d'ici à 2018, le recrutement de 140 000 à 190 000 spécialistes[69].

Ces technologies nécessitent la maîtrise d'outils mathématiques et statistiques de très haut niveau. Des compétences dans le domaine de l'informatique, et notamment en programmation, sont également requises. Élément important, les spécialistes doivent pouvoir travailler main dans la main avec les services commerciaux et avec les gestionnaires, et être au fait des règles concernant la sécurité et le respect de la vie privée. Pour l'heure, la plupart des analystes de données massives ont suivi une formation soit en informatique, soit en mathématiques-statistiques, puis se sont formés en autodidacte. Une première formation de niveau master a toutefois ouvert ses portes en septembre 2013 à Paris[70]. Elle ne pourra cependant pas répondre à toute la demande et aux besoins.

Dans le secteur public

Bien qu'il soit difficile d'appréhender l'ensemble des usages amenés à se développer, l'analyse des *big data* est un atout important pour l'administration. Peu de pays ont cependant mis en place des stratégies spécifiques en la matière.

L'Australie fait figure d'exception en voulant améliorer la gestion et les services publics à l'aide des analyses de masses de données. Dans son plan stratégique concernant les Technologies de l'information et de la communication (TIC) pour la période 2012-2015, le bureau de la gestion de l'information du gouvernement préconise par exemple d'établir un centre d'excellence pour l'analyse et la gestion des *big data* rattaché à l'ensemble du gouvernement, ou encore de rendre les données accessibles entre administrations.

D'autres pays, tel le Royaume-Uni, concentrent plutôt leurs efforts dans des secteurs spécifiques comme celui

64. http://www.afdel.fr/actualites/categorie/actualite-afdel/article/big-data-filiere-d-avenir-pour-la-france-les-propositions-de-l-afdel.
65. Commission innovation 2025 (2013), *Un principe et sept ambitions pour l'innovation,* commission présidée par Anne Lauvergeon, Paris, La Documentation française, octobre.
66. Ministère du Redressement productif, *La nouvelle France industrielle,* Paris, 2013.
67. À l'origine un projet franco-allemand, puis seulement français.
68. http://www.atinternet.fr/documents/barometre-des-moteurs-avril-2013/.
69. McKinsey Global Institute (2011), *Big Data. The Next Frontier for Innovation, Competition, and Productivity,* 156 p.
70. Télécom ParisTech.

de la santé. Les pouvoirs publics doivent financer, à hauteur de 90 millions de livres sterling (106 millions d'euros), l'institut *Big data* de l'université d'Oxford. Cet institut réalisera des analyses pour améliorer la détection, la surveillance, le traitement et la prévention d'un large éventail de maladies.

En France, que ce soit au niveau de la conception, de la mise en œuvre ou de l'évaluation des politiques publiques, mais aussi dans la gestion quotidienne des administrations, les analyses empiriques sont globalement peu utilisées. En ce sens, au-delà des contraintes associées à la protection de données, l'analyse des *big data* nécessite d'instaurer une culture de la donnée qui fait encore défaut. Il existe ainsi des quantités énormes de données "publiques" qui ne sont pas valorisées.

Étroitement liée à ce manque de recours aux analyses empiriques, la difficulté pour les administrations est d'investir dans des technologies dont les retombées sont difficilement chiffrables et dont la mise en œuvre peut s'avérer délicate (contraintes juridiques, partage des données entre administrations, etc.). Alors que les logiciels de type *open source* existent, des investissements sont nécessaires pour normaliser les données, pour sécuriser les échanges, mais aussi pour recruter ou former des analystes (voir *supra*).

Le peu de recours aux analyses de données dans la gestion et la prise de décision s'explique en partie par le cloisonnement des données. Le partage d'informations entre administrations et avec des acteurs externes est indispensable pour donner plus de valeur à l'analyse des *big data,* la richesse des analyses résidant essentiellement dans le rapprochement des données entre lesquelles on n'avait pas présupposé de relations.

Pour encourager les échanges, des normes de sécurité des échanges ont cependant été mises en place (voir *supra*). Des outils comme le Référentiel général d'interopérabilié (RGI), qui fixe les règles techniques permettant d'assurer l'interopérabilité des systèmes d'information, encouragent aussi le partage. Le mouvement d'*open data* (encadré 1) doit par ailleurs contribuer à ce décloisonnement. Autre exemple, depuis 2010, le Centre d'accès sécurisé distant (CASD) donne accès, de façon très encadrée, aux chercheurs (publics-privés) à des données individuelles (INSEE et Services statistiques ministériels).

Les responsables prévoient d'intégrer à ce dispositif des outils informatiques permettant des analyses de type *big data*[71].

Bien que divers formats de données puissent être croisés, il est par ailleurs important de faire en sorte que les données soient le plus harmonisées possible. Les données récoltées par deux administrations, à des niveaux géographiques différents ou pour des temporalités variables, ne seront par exemple pas ou difficilement compatibles. Cette incompatibilité s'explique par le fait que les données administratives ne sont généralement pas recueillies à des fins d'analyse, mais pour la gestion interne. Dans la mesure du possible, une réflexion sur la compatibilité des données entre administrations devrait être menée.

CONCLUSION

Le nombre de données continue à croître et les outils d'analyse vont se perfectionner. Sans présager des futurs usages, l'analyse des *big data* est sans aucun doute vouée à gagner en importance, certains parlant même de révolution[72].

Loin d'être un simple effet de mode, l'analyse permet de traiter des pathologies, de créer de nouvelles technologies, d'accroître nos connaissances, de prévenir des catastrophes, d'organiser les services, etc. D'un autre côté, l'analyse des données massives comporte des risques liés au respect de la vie privée, à la confidentialité, au libre-arbitre, auxquels il convient de réfléchir dès maintenant[73].

Mots clés : masse de données, analyse, données personnelles, administration électronique, prédiction.

71. CNIL (2013), *Workshop OpenCNIL Open Data,* Paris, 4 juillet.
72. Mayer-Schönberger V. et Cukier K. (2013), *op. cit.*, p. 60.
73. Les auteurs tiennent à remercier pour leur aide précieuse : Agnès Benassy-Quéré et Antoine Bozio (Conseil d'alayse économique), Denis Berthault (LexisNexis), Rémi Bilbault et Ruth Martinez (GFII), Pascal Caillerez (Décideur public – Systèmes d'information), Jean-Pierre Camilleri, Mehdi Benchoufi, Alexandre Bredimas et Christian Delom (Club Jade), Christine Chambaz, Alain Folliet et Marie-Noëlle Séhabiague (CNAF), Stéphan Clemençon (Telecom Paris Tech), Julien Damon (Sciences Po), Bertrand Diard (Talend), Joël Hamelin et Antton Achiary (CGSP), Charles Huot (TEMIS), Mathieu Jacomy (Médialab), Mathieu Jeandron et Annelise Massiera (DISIC), Nadia Joubert, Philippe Louviau, Rémi Favier et Bruno Nicoulaud (DNLF), Maxime Lesur et Bernard Ourghanlian (Microsoft), Claire Levallois-Barth (Institut Mines-Télécom), André Loth (DREES), Hammou Messatfa, Christophe Burgaud, David Kerr et Laura Haas (IBM), Philippe Niewbourg (Decideo), Judicaël Phan, Geoffrey Delcroix et Delphine Carnel (CNIL), Vincent Poubelle (CNAV), Pascal Saubion et Jean-Paul Leroux (Orange), Henri Verdier (Etalab), Marie Zins et Marcel Goldberg (INSERM).

DERNIÈRES PUBLICATIONS À CONSULTER

www.strategie.gouv.fr, rubrique publications

Notes d'analyse :

N° 01 ■ **Un fonds européen pour l'emploi des jeunes - Proposition pour une initiative** (juin 2013)

N° 02 ■ **Internet : prospective 2030** (juin 2013)

N° 03 ■ **Approvisionnements en métaux critiques : un enjeu pour la compétitivité des industries française et européenne ?** (juillet 2013)

N° 04 ■ **Les compagnies aériennes européennes sont-elles mortelles ? Perspectives à vingt ans** (juillet 2013)

N° 05 ■ **Pour un secteur des semences diversifié et innovant** (octobre 2013)

N° 06 ■ **Intensifier et réorienter les transferts de technologies bas carbone pour lutter contre le changement climatique** (octobre 2013)

N° 07 ■ **Doha, Varsovie, des conférences de transition vers un accord climatique mondial en 2015** (octobre 2013)

Retrouvez les dernières actualités du Commissariat général à la stratégie et à la prospective sur :

- www.strategie.gouv.fr
- CommissariatStrategieProspective
- @Strategie_Gouv

Commissariat général à la stratégie et à la prospective

La Note d'analyse n° 08 - novembre 2013 est une publication du Commissariat général à la stratégie et à la prospective
Directeur de la publication : Jean Pisani-Ferry, commissaire général
Directeur de la rédaction : Hervé Monange, adjoint au commissaire général
Secrétaires de rédaction : Delphine Gorges, Valérie Senné
Impression : Commissariat général à la stratégie et à la prospective
Dépôt légal : novembre 2013 - N° ISSN : 1760-5733
Contact presse : Jean-Michel Roullé, responsable de la communication - 01 42 75 61 37 / 06 46 55 38 38
jean-michel.roulle@strategie.gouv.fr

Créé par décret du 22 avril 2013, le Commissariat général à la stratégie et à la prospective se substitue au Centre d'analyse stratégique. Lieu d'échanges et de concertation, le Commissariat général apporte son concours au Gouvernement pour la détermination des grandes orientations de l'avenir de la nation et des objectifs à moyen et long termes de son développement économique, social, culturel et environnemental. Il contribue, par ailleurs, à la préparation des réformes décidées par les pouvoirs publics.