Le moment big data des sciences sociales

Gilles Bastin, Paola Tubaro

Le moment big data *des sciences sociales*

Gilles BASTIN
Paola TUBARO

Peu de sujets ont suscité dans les années récentes autant d'intérêt dans le débat public et dans les sciences sociales que celui des *big data*. La numérisation d'un nombre toujours plus grand d'activités sociales produit régulièrement de nouvelles données et alimente aussi une réflexion intense sur le fonctionnement des sociétés contemporaines ainsi que sur les modalités de la production du savoir à leur propos. De ce fait, une grande part de la littérature consacrée aux *big data* dans les sciences sociales oscille entre deux approches. La première vise à caractériser de manière instrumentale ces données dites massives par opposition aux données d'enquête, traditionnellement utilisées par les chercheurs. Elle questionne l'usage qui en est fait (Kitchin, 2014) comme celui qui pourrait en être fait (Beer et Burrows, 2007 ; Boullier, 2015 ; Varian, 2014). La seconde approche oppose à l'enthousiasme suscité par les *big data*, principalement dans le domaine des activités marchandes, une analyse des risques induits par le recours à ce type de données : la crainte d'obsolescence de la méthode scientifique d'analyse des données, appelée à être remplacée par des méthodes algorithmiques sans lien fort avec les théories sociales (Anderson, 2008) mais aussi, et surtout, l'apparition d'une nouvelle forme de société « dirigée par les données » (Pentland, 2012), bouleversant nos façons de vivre, de travailler et de penser (Mayer-Schönberger et Cukier, 2013), soulevant des questions éthiques inédites (Boyd et Crawford 2012) et annonçant l'avènement d'une nouvelle « gouvernementalité » du social (Rouvroy et Berns, 2013).

Les sociologues des techniques de l'information et de la communication ont été parmi les premiers à s'investir dans l'exploration de ces nouveaux matériaux, mais ils n'ont pas été les seuls. Ce changement spectaculaire du paysage des données qui touche la société entière a suscité la mobilisation de chercheurs dans tous les domaines des sciences humaines et sociales. Témoignent de cet engagement de la communauté académique la création de centres de recherche comme le Data & Society Research Institute à New York (2014), celle de nouvelles revues savantes comme *Big Data & Society* (2013), et de nombreux numéros spéciaux de revues généralistes. On peut citer entre autres ceux de l'*International Journal of Sociology* (2016), de *Sociological Methodology* (2015), des *ANNALS of the American Academy of Political and Social Science* (2015), du *Journal of Communication* (2014), de l'*International Journal of Communication* (2014), du *Journal of Economic Perspectives* (2014) et, dans l'espace francophone, d'*Économie et statistique* (à venir), de *Sociologie et sociétés* (à venir), de *Statistique et société* (2014).

À la suite de ce foisonnement d'initiatives, il a paru utile à la *Revue française de sociologie* de proposer à la communauté académique dans cette discipline de dresser un premier bilan des effets des *big data* sur ses pratiques, ses objets et ses résultats. Quels avantages concrets les sociologues qui ont commencé à travailler avec ces données ont-ils pu en tirer pour leurs recherches ? Quels écueils ont-ils rencontrés en chemin ? Plus généralement, qu'a-t-on appris sur les *big data* tout au long de ces premières années d'expérimentation ? L'ambition de ce numéro spécial est d'apporter quelques éléments de réponse à ces questions afin d'évaluer la portée et l'ampleur des changements qui se sont

produits d'ores et déjà dans le monde académique et d'anticiper, autant que faire se peut, des scénarios possibles pour l'avenir.

À cette fin, nous avons mobilisé la communauté des sociologues autour de deux grandes questions qui ne nous paraissaient pas devoir être séparées : comment les *big data* transforment-elles la société ? Comment ces données affectent-elles la pratique de la sociologie (et plus généralement, des sciences sociales) ? Cette double approche – que nous avons tenu à conserver dans la préparation de ce numéro – manifeste le fait que la sociologie est aujourd'hui confrontée à une véritable mutation des processus de *datafication* des sociétés qui la touche aussi directement. En somme, ce numéro aspire à réitérer à propos des *big data* le questionnement qu'Alain Desrosières avait appliqué aux statistiques et à la *quantification* du social, considérant sans jamais les séparer « leurs apports de connaissance et les circuits sociaux de leur mise en forme et de leurs usages » (2005, p. 6).

Parce qu'elles sont étroitement entrelacées, ces deux dimensions du phénomène des *big data* posent des défis inédits à la sociologie. Si l'on adopte le point de vue réflexif de Savage et Burrows (2007) dans leur fameux article sur la « *coming crisis* » de la recherche empirique dans cette discipline, la sociologie serait en train de perdre sa « juridiction » sur tout un pan de la connaissance de la société. L'entretien et l'enquête par questionnaire qui lui ont longtemps assuré cette juridiction seraient en effet dépassés par de nouveaux modes de représentation de la société, sans lien évident avec les connaissances sociologiques acquises. C'est à la fois parce que ces représentations sont issues de transformations sociales dont l'origine se situe généralement en dehors de toute démarche de recherche (comme la mise en chiffre de toute action, opinion ou comportement par des plateformes numériques qui visent avant tout à commodifier et à monétiser ces informations), et parce que les propriétés de ces modes de représentation de la société s'écartent des critères chers à l'analyse sociologique traditionnelle (en termes de représentativité des échantillons par exemple), qu'elles posent avec force la question de la légitimité de toute une discipline et, par conséquent, celle des chercheurs qui s'en revendiquent.

Dans cet esprit, nous avons sollicité des contributions portant autant sur les problèmes publics qui ont émergé dans le sillage des *big data* (comme la surveillance, la propriété privée des données, le *digital labor*, les formes de discrimination algorithmique, mais aussi les formes de réappropriation des données dans le cadre de mouvements comme celui du *civic tech* ou de l'*open data*), que sur les enjeux méthodologiques propres aux sciences sociales (par exemple les éléments de continuité et rupture par rapport aux méthodes traditionnelles de la recherche en sciences sociales, le besoin de mise à niveau des compétences, le renouveau du dialogue interdisciplinaire). Nous avons aussi étendu le périmètre de l'appel à soumission d'articles jusqu'aux questions épistémologiques qui se posent aujourd'hui, comme par exemple celle de savoir dans quelle mesure les *big data* bouleversent l'espace de l'enquête et les modalités du « raisonnement sociologique » (Passeron, 1991). Le débat sur le remplacement de l'analyse causale par la combinaison de corrélations efficaces en termes de prédiction mais très peu en termes d'explication, l'émergence dans le domaine des *big data* de notions comme celle de « trace » (Merzeau, 2009) qui fait écho à des préoccupations anciennes de la sociologie sur la nature des matériaux qu'elle utilise ou encore le développement de modèles d'enquête inspirés de la police scientifique dans la « *forensic social science* » (Goldberg, 2015) illustrent quelques-unes des préoccupations qui nous animaient. La *Revue française de sociologie*, qui s'attache à publier aussi bien des contributions à la connaissance du monde social que des articles de réflexion théorique et méthodologique sur la sociologie, était le lieu idéal pour lancer un tel chantier.

Les *big data*, quelle définition ?

Avant de présenter le contenu de ce numéro et la façon dont nous avons travaillé avec les auteurs des articles, il est nécessaire de définir plus précisément le périmètre scientifique que nous délimitions dans l'appel à contributions. Qu'entend-on par *big data* aujourd'hui ? L'expression anglaise est de loin la plus utilisée, y compris en France[1]. Nous l'avons donc adoptée malgré l'existence d'une traduction française recommandée par le *Journal officiel*, celle de « mega-données »[2] et d'une alternative non officielle et moins employée, l'expression « données massives ». Dans la communauté académique l'origine de l'expression a été identifiée dans un article de l'économiste américain Francis Diebold intitulé « "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting » (2003). F. Diebold lui-même a précisé ultérieurement ce que le terme devait à des débats dans d'autres milieux comme ceux des ingénieurs de l'entreprise informatique Silicon Graphics, Inc. Il a aussi expliqué avoir cherché un terme évocateur, susceptible de résumer la révolution en cours dans les pratiques de modélisation économétrique d'une part et dans le volume des données disponibles pour cette modélisation d'autre part. La connotation orwellienne du terme *Big Data*, alors employé avec des majuscules, lui avait semblé donner encore plus de poids à l'expression (Diebold, 2012).

Des nombreuses définitions qui ont depuis été proposées des *big data*, aucune n'est réellement consensuelle. Des différences d'usage persistent, qu'elles soient liées au contexte (entreprises commerciales *vs.* recherche publique) et à la discipline (informatique *vs.* autres sciences). Selon une première caractérisation assez basique, sont qualifiées de « *big* » les données dont le traitement et le stockage dépassent les capacités des outils informatiques classiques de gestion de bases de données ou de l'information. Ces données nécessitent l'usage d'instruments informatiques sophistiqués comme ceux qui permettent le calcul parallèle à haute performance, par exemple avec un ensemble d'ordinateurs dédiés, reliés par un réseau local rapide, ou un réseau de cartes graphiques (GPU) détournées pour du calcul scientifique. Mais à elle seule, cette caractérisation parait faible. Beatrice Cherrier (2017) a montré que l'insuffisance des capacités de traitement et stockage n'est pas un problème fondamentalement nouveau. Cette insuffisance caractérise en effet de manière cyclique les phases postérieures à chaque innovation technologique dans le domaine de la gestion de l'information. Aux États-Unis, par exemple, elle était déjà apparue après le recensement de 1890 que les anciens instruments ne permirent pas de tabuler en dix ans (d'où l'introduction de cartes perforées), ou dans les années 1940 quand les systèmes existants de classification s'avérèrent incapables de suivre l'expansion rapide des bibliothèques et durent donc être révisés.

Une définition couramment utilisée des *big data* est celle dite des « 3 V » proposée très tôt par Doug Laney (2001) et systématisée par la suite (De Mauro *et al*., 2016). Au delà du seul volume (le premier V), elle souligne l'importance du critère de variété (les données sont hétérogènes et peuvent correspondre aussi bien à des valeurs chiffrées ou codées, comme dans les statistiques traditionnelles, qu'à des textes, des images, des vidéos, etc.) et celui de

[1] L'usage le plus courant est celui des minuscules, quoique d'aucuns mettent les initiales en majuscules (« *Big Data* ») comme s'il s'agissait d'un nom propre ou d'une entité unique reconnaissable (comme « *Big Oil* »). Nous avons décidé d'utiliser, dans tout le numéro, *big data* sans guillemets mais en italique pour signaler la locution étrangère.

[2] Voir le *Journal officiel* du 22 août 2014 : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000029388087&dateTexte=&categorieLien=id).*

vélocité de captation (celle-ci est en effet le plus souvent continue et immédiate comme dans l'enregistrement automatique des logs d'activité d'un service en ligne, celui des données de capteurs connectés au *web*, des vidéos des caméras de surveillance ou des messages postés sur les médias sociaux). Ces caractéristiques des *big data* expliquent le foisonnement de travaux informatiques de conception et exploitation de bases de données « *not only SQL* » (*NOSQL*), permettant de stocker et d'interroger des données complexes[3]. Bien qu'elle permette de prendre en compte ces aspects importants des *big data*, cette définition reste contestée. Certains auteurs lui ont par exemple ajouté un quatrième V (la véracité des données), voire un cinquième (leur valeur) et d'autres caractéristiques ne pouvant être associées à la lettre V comme l'exhaustivité, la résolution ou le caractère relationnel des données (Kitchin, 2014). Certains considèrent finalement que ces grandes caractéristiques relèvent plus des effets des données massives que de leur nature.

Les définitions les plus récentes associent les données aux algorithmes capables de les traiter. Le mot algorithme, pourtant ancien, a d'ailleurs connu un regain de popularité après les premiers engouements pour les *big data*. En effet, l'utilité et la valeur des données ne peuvent être séparées des opérations algorithmiques appliquées à ces données (Cardon, 2015). Au plus simple, un algorithme est défini comme une séquence d'instructions pour résoudre un problème (Abiteboul et Dowek, 2017), qu'il s'agisse de la collecte ou du traitement des données, ou bien encore de la restitution des résultats.

De fait, le débat sur les *big data* n'est pas séparable de l'intérêt croissant pour les algorithmes relevant du « *machine learning* » ou, en français, « apprentissage automatique », dans les sciences sociales[4]. Au croisement de l'informatique et des statistiques, le *machine learning* est un ensemble de méthodes faisant évoluer une machine par entrainement, de sorte qu'elle trouve seule des solutions à partir des données fournies, sans que chaque étape ne doive être explicitement programmée. L'entrainement exige de grandes quantités de données pour que la solution puisse s'ajuster et aboutir à un réglage toujours plus fin. Ainsi, le *machine learning* bénéficie du volume des *big data* et dépasse la performance des techniques statistiques classiques, souvent mises à mal par la variété de ces données. Resté longtemps dans l'ombre, il s'est récemment imposé à la suite de réussites importantes, notamment dans la reconnaissance d'images numérisées (2012), puis dans les jeux comme Go (2016). À leur tour, ces succès ont insufflé une nouvelle vie à l'intelligence artificielle, un domaine qui a pu puiser dans d'autres approches scientifiques et techniques au cours de son histoire, mais dont les avancées actuelles reposent surtout sur le *machine learning* et les *big data*. Cette intelligence artificielle dont le champ d'application s'étend désormais du diagnostic médical à la voiture autonome et à la distribution d'énergie, pour ne citer que quelques exemples, est aujourd'hui au cœur de préoccupations industrielles et politiques majeures (Villani, 2018).

Il nous a paru important de prendre en compte ces aspects dans leur complémentarité. Les succès autant scientifiques que techniques et commerciaux des *big data* ne peuvent se comprendre sans rappeler l'essor parallèle du *machine learning*. De même, les inquiétudes formulées sur les *big data* sont reliées aux questions de société soulevées par l'intelligence artificielle. Méthodologiquement et épistémologiquement, il nous a donc paru important de

---

[3] Le langage *SQL* (*Structured Query Language)* a très longtemps été utilisé pour interroger les bases de données relationnelles classiques mais il est de plus en plus remplacé par des langages non structurés permettant une plus grande souplesse dans l'analyse des données.

[4] À partir de mai 2017, l'expression « *machine learning* » est systématiquement plus recherchée que « *big data* » dans Google, tous pays confondus (*Source* : Google Trends, consulté le 10 mai 2018). Voir aussi : https://www.kdnuggets.com/2017/05/machine-learning-overtaking-big-data.html

faire de la place à une discussion de ces nouvelles techniques, quelles que soient d'ailleurs les données auxquelles elles s'appliquent : qu'est-ce que le *machine learning* peut apporter à la recherche en sciences sociales ? Va-t-il mettre à mal les méthodes quantitatives classiques ? Ces questions sont d'autant plus pertinentes que ces techniques se font déjà une place dans des disciplines voisines de la sociologie comme l'économie[5].

Données massives et données personnelles

Couplées aux algorithmes de *machine learning*, ces nouvelles données ouvrent la voie à des analyses ciblées, qui ne se contentent plus de résultats agrégés, fondés sur des moyennes. Se référant à la recherche clinique, Tim O'Reilly et ses co-auteurs (2012) proposent l'exemple d'un médicament dont il était auparavant connu que, chez un patient moyen, il était efficace à environ 80 %. Maintenant, nous savons qu'il est efficace à 100 % chez 70 % à 80 % des patients, et inefficace chez les autres. Ce ne sont pas des jeux de mots, car nous pouvons dire très précisément si le médicament est susceptible d'être efficace pour un patient spécifique. Ce simple exemple révèle une autre facette des *big data* qui, au-delà de leur volume, permettent de pénétrer le niveau micro, la personne dans sa singularité. Ce n'est alors pas une coïncidence si les enjeux de la protection des données personnelles interviennent transversalement dans tous les débats que nous venons d'évoquer – car l'intrusion dans la vie de l'individu, qui peut parfois sauver des vies, a aussi un côté sinistre. L'essor des *big data* dès 2011-2012 a pu en effet être associé aux préoccupations qui se font jour dans la société en matière de protection de la vie personnelle et de clarification des frontières entre sphère publique et sphère privée sur le *web*.

Dans ce domaine, les réactions les plus fortes sont surtout le fait de groupes d'usagers engagés, comme le montre la tentative de recours collectif d'*Europe v Facebook* (2014). Mais les scandales fortement médiatisés autour de la NSA (2013) et plus récemment de Cambridge Analytica (2018) ont mis les questions des données personnelles et de la protection de la vie privée sur le devant de la scène (Tubaro *et al*., 2014). Après avoir attiré l'attention du grand public, le débat s'est progressivement institutionnalisé et a conduit les autorités à légiférer, notamment en Europe avec le Règlement général sur la protection des données (RGPD) récemment entré en vigueur.

Si l'on doit saluer ces actions qui visent à protéger davantage les usagers des services numériques, on peut aussi s'interroger, en tant que sociologues, sur leurs implications en termes d'accès aux données pour la recherche. Actuellement, les lois sur la protection des données personnelles de la plupart des pays reconnaissent la finalité de recherche et autorisent l'accès ou la manipulation de données personnelles à cette fin, généralement sous des conditions strictes de publicité, de rétractabilité et d'anonymisation. Pour des données individuelles très détaillées, des structures d'accès hautement sécurisé ont aussi été mises en place comme le Centre d'accès sécurisé aux données (CASD) en France. Les chercheurs se plaignent parfois des contraintes imposées par ces services (interdiction de copier les données, export de résultats seulement après validation, etc.). Cependant, ceux-ci ont établi une procédure transparente et égalitaire d'accès aux données et élargi le champ des données publiques accessibles pour la recherche.

[5] À titre d'exemple, la base de données EconLit maintenue par l'American Economic Association recense plus de 300 articles, 20 thèses et 60 documents de travail en économie portant la mention *machine learning*, tous datant de 2005 à aujourd'hui (consulté le 11 mai 2018).

Le problème est que les données numériques et « massives » ne rentrent pas nécessairement dans le périmètre de ce qui est géré par ces structures. Si les principes fondamentaux de la protection des données personnelles s'appliquent toujours en théorie dans le domaine des *big data*, leur opérationnalisation peut être plus complexe et conduire le chercheur dans une zone grise du point de vue du droit. Les informations qu'un usager publie volontairement sur son profil LinkedIn ou Facebook constituent-elles des données personnelles au même titre que ses réponses à un enquêteur de l'Insee ? Il est difficile de trancher sur ce type de question, où le point de vue du chercheur, celui de l'usager, celui de la plateforme et celui du régulateur divergent souvent (Bastin et Francony, 2016). La décision sur le statut des données se complique encore davantage si l'on tient compte de la structure en réseau des profils (les « *retweets* » que suscitent mes propres *tweets* sont-ils mes données personnelles, ou celles des usagers qui m'ont *retweeté* ?) et de la dynamique temporelle des identités sur le *web* (les données collectées sur un réseau social à un temps *t* doivent-elles être supprimées *ex post* si à $t + n$ l'utilisateur décide de transformer son compte public en compte privé ?).

Ces difficultés de définition risquent aujourd'hui d'être exacerbées par l'attention croissante que reçoit la question de la protection des données personnelles. La réaction contre les abus avérés, commis le plus souvent par les plateformes elles-mêmes ou par des entreprises vendant des services fondés sur l'exploitation de données à des fins économiques ou politiques peut se retourner contre le chercheur, si l'intérêt public de son activité n'apparait pas clairement et n'est pas explicitement défendu. Le chercheur se trouve en effet dans une position délicate puisqu'il contribue d'une part à l'analyse des problèmes de surveillance et de violation de la vie privée comme enjeux sociétaux majeurs liés aux *big data* et que, d'autre part, il doit pour cela – ainsi que dans un but purement descriptif – s'assurer un accès à ces données.

Ces questions fondamentales seront sans aucun doute à l'avenir au centre des nombreux débats que devra susciter la communauté académique si elle ambitionne de continuer à utiliser les *big data*. Mais elles ne sont pas les seules. La question de la collecte matérielle des données ouvre un autre champ immense de réflexion (mais aussi d'inquiétudes) pour les sociologues. Comment négocier avec les plateformes de réseaux sociaux un accès à leurs données qui ne soit pas biaisé par les caractéristiques d'une *API* (*application programming interface*) créées par leurs soins et impossible à contrôler ? Comment éviter le risque juridique auquel expose le fait de récolter des données sur des plateformes qui limitent de façon drastique les possibilités de le faire de façon automatisée (par *scraping*) dans leurs conditions générales d'utilisation ? Comment, enfin, développer des compétences informatiques qui garantissent au chercheur de pouvoir mener ce travail en dépit des barrières posées dans le code lui-même pour empêcher son activité ?

La construction du numéro

Tous ces questionnements sont à l'origine du projet de ce numéro spécial dont l'appel à contributions a été diffusé, en langues française et anglaise, en novembre 2016. La réponse de la communauté académique a de loin dépassé nos attentes les plus optimistes : à l'échéance (février 2017), nous avons reçu 41 soumissions, couvrant un éventail très large d'aspects, de problèmes et de nuances. Le choix a donc été difficile. Dans notre travail de présélection, nous avons été guidés par notre objectif de mettre en avant des travaux qui interrogent les effets sociaux et les implications scientifiques des *big data* à partir d'une expérience concrète

de recherche dans ce domaine, plutôt que des textes programmatiques. Les auteurs des 15 pré-propositions retenues ont été invités à présenter leurs projets d'articles à un atelier que nous avons organisé à Grenoble en juin 2017[6]. Hormis un désistement, tous les auteurs ont joué le jeu, et nous avons été ravis d'observer un bel esprit d'entraide et de coopération, visant à faire progresser les articles de tous les participants à l'atelier. Cette stratégie a été productive car 14 textes complets ont été soumis en septembre 2017. La sélection finale est le résultat de la procédure de lecture habituelle de la revue, qui a eu lieu entre l'automne 2017 et le printemps 2018.

En soi, et avant même de pouvoir nous exprimer sur les articles qui ont franchi toutes les étapes de la relecture par les pairs, ce processus a été très instructif. Le nombre de pré-propositions qui nous ont été adressées témoigne du grand intérêt que cette thématique suscite au sein de notre discipline. Si l'on considère le nombre élevé de jeunes chercheurs qui ont participé à notre atelier grenoblois, cet intérêt ne pourra que grandir davantage dans le temps. Certaines des propositions n'étaient sans doute pas encore assez mûres pour une publication dans ce numéro (s'agissant, dans certains cas, de doctorants au début de leur thèse), mais nous avons apprécié les fortes potentialités des recherches en cours sur lesquelles elles étaient fondées, et nous nous attendons à une vague de travaux de très bonne qualité sur ces thèmes d'ici quelques années. L'émergence de ces nouveaux travaux est aussi attestée par la production éditoriale dans ce domaine. La revue s'en est d'ailleurs aussi fait l'écho dans ce numéro avec un effort spécial mené par les responsables de la rubrique Livres pour solliciter des compte rendus d'ouvrages qui complètent utilement le panorama de la recherche sociologique actuelle liée aux *big data*.

Au total, en plus des comptes rendus, ce numéro consiste en quatre articles et deux notes critiques. Le premier article est une contribution de Marie Bergström à la sociologie de la formation des couples dans laquelle sont utilisées et confrontées deux sources de données : des données classiques d'enquête sociologique et des données tirées du réseau social Meetic. L'article est intitulé « De quoi l'écart d'âge est-il le nombre ? L'apport des *big data* à l'étude de la différence d'âge au sein des couples ». Il traite avec un regard neuf une question classique de la sociologie, à savoir celle de l'écart d'âge qu'on observe dans les couples hétérosexuels entre l'homme (statistiquement plus âgé) et la femme. M. Bergström utilise les différentes formes d'enregistrement des préférences des hommes et des femmes en matière d'âge de leur conjoint : d'un côté des préférences déclarées face au sociologue dans l'enquête classique portant sur 7 800 individus ; de l'autre des préférences exprimées par plus de 400 000 membres de Meetic afin de paramétrer les offres de rencontres qui seront proposées et les 25 millions de messages échangés entre ces individus en 2014. L'écart entre ces deux séries de données permet à M. Bergström une contribution forte à l'étude de la mise en couple et des préférences sexuelles : l'écart d'âge n'est en effet pas seulement la résultante d'une domination masculine intériorisée par les femmes mais aussi de stratégies masculines explicites. Il conduit aussi à une réflexion intéressante sur la prudence avec laquelle certaines déclarations faites lors d'enquêtes sociologiques classiques doivent être considérées, par exemple pour ce qui est de la tolérance des hommes à un écart d'âge en leur défaveur dans le couple. Celle-ci est exprimée très majoritairement dans l'enquête sociologique mais elle est démentie par les pratiques de contact sur Meetic et ce d'autant plus que les hommes vieillissent.

[6] Cet atelier a bénéficié du soutien du Data Institute de l'université Grenoble Alpes financé par l'ANR dans le cadre du programme « Investissements d'avenir » (ANR-15-IDEX-02).

Dans le deuxième article, intitulé « Plateforme, *big data* et recomposition du gouvernement urbain. Les effets de Waze sur les politiques de régulation du trafic », Antoine Courmont s'intéresse au dispositif de guidage automobile Waze qui propose à ses usagers une optimisation en temps réel de leur temps de parcours grâce aux informations sur l'état de la circulation que les mêmes personnes communiquent à la plateforme par leur géolocalisation. A. Courmont analyse finement les perturbations que Waze introduit dans le système de gestion publique de la circulation automobile. Du fait que cette plateforme cherche avant tout à optimiser les temps de parcours, elle oriente en effet ses usagers vers des itinéraires qui ne respectent pas les principes usuels de la régulation du trafic, notamment en termes de hiérarchie des voies empruntées. Deux « réalités » concurrentes d'un même « monde » se font face pour A. Courmont : d'un côté celle des élus, des techniciens et des opérateurs de la gouvernance urbaine traditionnelle, qui cherchent à concentrer le trafic sur les voies les plus importantes, de l'autre celle de l'algorithme qui dévie les automobilistes sur le réseau secondaire dès lors que cette déviation leur fait gagner du temps. Mais l'enquête de terrain révèle surtout que ces acteurs et l'algorithme, bien loin de représenter deux formes totalement antagonistes de gouvernance par les données, savent aussi « faire réalité commune » en échangeant des informations et en collaborant.

Le troisième article s'intitule « Le tout plutôt que la partie. *Big data* et pluralité des mesures de l'opinion sur le *web* ». Baptiste Kotras exploite les résultats d'une enquête par entretiens menée dans le monde des intermédiaires des données d'opinion en ligne, un secteur d'activité florissant souvent identifié sous l'appellation *social media analysis*. Alors que de plus en plus d'individus partagent sur le *web* des opinions très diverses dans leurs billets de blog, leurs posts sur les réseaux sociaux, leurs statuts ou les commentaires laissés sur des sites, de nombreuses entreprises ont développé des algorithmes permettant de synthétiser ces opinions, de les mesurer et de les représenter. Les promesses des études d'opinion en ligne sont en effet considérables, tant en termes d'instantanéité que de spontanéité, et semblent parfois faire vaciller l'ancien monde des sondages d'opinion. B. Kotras observe dans le cas français la lutte entre deux modalités de la mesure d'opinion en ligne : la première obéit à un principe d'échantillonnage et procède par sélection de sources fiables et influentes ; la seconde repose en principe sur une aspiration exhaustive du *web*, sur l'indexation et la typification du plus grand nombre possible de traces laissées par ses utilisateurs. Ce conflit semble aujourd'hui tranché en faveur de la seconde approche. B. Kotras en retrace finement les enjeux sur un plan technique, économique mais aussi épistémologique dans la mesure où, dans ce conflit, ce sont deux représentations de l'espace public numérique qui sont aussi mobilisées.

L'article d'Étienne Ollion et Julien Boelaert intitulé « *The Great Regression: Machine Learning, Econometrics, and the Future of Quantitative Social Sciences* » présente de manière convaincante et pédagogique les méthodes de *machine learning* et les compare avec les méthodes classiques de la statistique « paramétrique », notamment la régression. Il se centre non pas sur toutes les techniques de *machine learning* mais sur l'utilisation de celles-ci à des finalités d'exploration scientifique pour la prédiction et l'explication. Il montre comment cette approche peut être efficacement mobilisée pour l'exploration de la complexité des relations entre variables dans un jeu de données. Au-delà de l'exercice méthodologique, l'article revient sur les enthousiasmes et les craintes qui entourent le *machine learning* et défend une thèse provocatrice : le *machine learning* n'est ni une illusion, ni une révolution, et surtout ne produira pas de changement de paradigme dans l'immédiat. Il a plutôt le mérite de faire avancer une réflexion méthodologique poussée, tenant compte des limites avérées des méthodes quantitatives couramment utilisées en sociologie. Les auteurs prévoient une

concurrence accrue entre différentes formes de quantification du monde social qui, dans leur esprit, ne peut que renforcer la recherche sociologique en imposant des exigences plus strictes de rigueur.

Enfin, nous publions dans ce numéro deux notes critiques qui nous ont paru offrir un point de vue intéressant sur deux domaines particulièrement actifs de la recherche fondée sur les *big data* : l'exploitation des données tirées du réseau social Twitter d'une part, qui a focalisé l'attention de tout un pan des études d'opinion depuis une dizaine d'années, et le renouveau de la statistique lexicale permis par l'application de techniques nouvelles aux corpus conséquents numérisés depuis les années 2000. Pour ce qui est de Twitter, Marta Severo et Robin Lamarche-Perrin proposent dans leur article intitulé « L'analyse des opinions politiques sur Twitter : défis et opportunités d'une approche multi-échelle » une relecture des travaux menés dans les sciences sociales dans la décennie 2010. Ils exposent les différentes façons de conceptualiser les données du réseau de micro-blogging et les rattachent à des conceptions de l'opinion. Ils rappellent les résultats de ces recherches et proposent un cadre analytique pour une appréhension complexe de ces données permettant d'en explorer aussi bien les contenus que les structures relationnelles révélées par le phénomène des *retweets* et des réponses suscités par un message. Les différentes méthodes, plus ou moins supervisées, d'analyse de ces données sont aussi présentées.

Dans le domaine de la statistique lexicale, Jean-Philippe Cointet et Sylvain Parasie proposent dans leur article intitulé « Ce que le big data fait à l'analyse sociologique des textes : un panorama critique des recherches contemporaines » de faire le point sur la façon dont les sociologues – ou d'autres spécialistes des sciences sociales – traitent aujourd'hui les matériaux textuels. Le volume de ces documents a beaucoup augmenté avec la numérisation de pans considérables du patrimoine culturel et des médias d'une part, avec la collecte de données plus hétérogènes liées à l'activité des usagers du web d'autre part. Ce champ de recherche, qui avait donné lieu dès les années 1970 au développement de méthodes originales et de solutions informatiques, en France notamment, a été profondément bouleversé ces dernières années. L'irruption des études fondées sur l'analyse d'occurrences dans de gros corpus numérisés (*culturomics*), des humanités numériques ou de l'étude de sentiments illustrent assez bien ce changement du paysage académique. J.-P. Cointet et S. Parasie synthétisent l'ensemble des nouvelles méthodes employées depuis une dizaine d'années et proposent d'illustrer les grandes tendances observables à partir d'exemples qui permettront à tous les sociologues de se faire une idée à la fois précise et évocatrice des potentialités ouvertes en matière d'enquête sociologique par des méthodes comme la modélisation thématique (*topic modeling*) ou les plongements de mots (*word embedding*) en matière d'enquête sociologique.

Des données numériques mais pas nécessairement massives

Les profils et les messages des usagers d'un site de rencontres (M. Bergström), les interactions sur Twitter (M. Severo et R. Lamarche-Perrin), les corpus textuels indexés par Google ou les grandes bibliothèques du monde (J.-P. Cointet et S. Parasie) sont de très bons exemples des données nouvelles que les sociologues peuvent s'approprier aujourd'hui pour éclairer de manière originale des phénomènes sociaux jusqu'ici mal compris. Ces données mettent les sociologues face à des défis importants bien identifiés dans les articles regroupés dans ce numéro, comme le manque de représentativité de ces données qui doit inviter à la prudence avant toute généralisation des résultats à des populations dont des couches parfois

importantes restent encore aujourd'hui peu numérisées. Pour cette raison, de nouvelles méthodes d'analyse, principalement issues de l'informatique, doivent être mobilisées.

Malgré leur originalité et ces changements méthodologiques majeurs en matière d'analyse, les données utilisées par nos auteurs ne sont pas toujours massives au sens strict du terme. La taille des données numériques utilisées dans notre discipline ne dépasse pas encore de façon significative celle des grandes bases de données de la statistique publique classique, comme le recensement ou l'enquête emploi, et le temps ne semble pas encore arrivé où les sociologues devront recourir à des outils nouveaux en matière de traitement des données comme le calcul parallèle qui occupe une partie des collègues qui se reconnaissent dans d'autres disciplines comme des praticiens des *big data*. Le cas des données administratives utilisées par É. Ollion et J. Boelaert illustre ce principe de continuité : le recours à des registres administratifs arrivant parfois à achever une couverture exhaustive de la population au lieu d'enquêtes sur échantillons est une pratique bien enracinée dans la tradition de la statistique publique des pays de l'Europe du Nord. Ces données ne sont pas nouvelles et ne remplissent pas tous les critères des « V » des *big data*. Elles sont notamment assez homogènes et bien structurées. La disponibilité de ce type de données est destinée à augmenter à l'avenir, y compris dans les pays qui n'y avaient pas recours, du fait de la diminution des ressources que les États consacrent à la statistique publique, et de la recherche de moyens de connaissance de la société supposant un moindre effort de la part des enquêtés. Les sociologues trouveront sans doute avantage à exploiter ces données qui supposent cependant de leur part un plus grand investissement en termes de méthodes mais aussi d'accès[7]. De même les données textuelles dont J.-P. Cointet et J. Parasie analysent l'usage dans les sciences sociales – qu'elles soient issues de l'enregistrement de conversations en ligne (par exemple, sur un forum) ou de la numérisation de documents divers (livres, journaux, archives, registres comptables, actes juridiques, etc.) – ne sont pas complètement inédites et ont donné lieu à des innovations méthodologiques dès les années 1970 avec le développement de l'informatique personnelle et son accessibilité pour les chercheurs en sciences humaines et sociales. Elles connaissent cependant un renouveau aujourd'hui grâce à la disponibilité d'outils novateurs issus de la recherche informatique dans le domaine du traitement automatique de la langue, et à l'augmentation du volume de textes numérisés.

Le volume ne peut donc pas servir à caractériser ce que les sociologues appellent *big data*. D'une part, des sources classiques de données sociales sont déjà volumineuses, d'autre part, les études fondées sur des données nativement numériques ne reposent pas toujours sur de grands volumes dans notre discipline. Il y a là un mystère persistant dans la mesure où la société vit, pour sa part, dans ce que la littérature de vulgarisation tend à qualifier de « régime de l'abondance » en matière de données (Shadbolt et Verdier 2015). On rappelle souvent par exemple que près de 500 millions de *tweets* sont envoyés chaque jour dans le monde, que 600 ventes sont réalisées en moyenne chaque seconde sur Amazon, que 250 millions de comptes LinkedIn sont utilisés au moins une fois par mois et que plus de 5 milliards de personnes produisent aujourd'hui des données quotidiennement en appelant ou envoyant des textos depuis leurs téléphones mobiles. Les articles publiés dans ce numéro témoignent d'ailleurs de la taille des bases de données dont disposent les grandes entreprises du numérique, à l'image de Waze qui géolocalise et oriente quotidiennement 100 millions d'usagers (A. Courmont), Linkfluence, qui aspire à fournir un accès « exhaustif » aux expressions d'opinion sur le *web* (B. Kotras), Meetic, qui offre à un très grand nombre d'individus des possibilités nouvelles

[7] Les conditions d'accès à ces données peuvent être très strictes. L'accès aux données suédoises utilisées par É. Ollion et J. Boelaert ne se fait par exemple que sur le territoire du pays, et même une correction mineure au modèle nécessite de s'y rendre.

d'entamer une relation amoureuse (M. Bergström) ou évidemment Twitter, qui héberge un volume considérable de conversations politiques utilisées par de nombreux acteurs pour prédire des résultats électoraux (M. Severo et R. Lamarche-Robin).

Le problème auquel les sociologues sont aujourd'hui confrontés est celui de l'accès à ces données hébergées par des entreprises de plus en plus réticentes à les partager avec les acteurs de la recherche publique. Ce qui change en effet fondamentalement entre le monde de la statistique publique et des enquêtes sociales qui ont accompagné le développement des sciences sociales au XX^{e} siècle, d'une part, et celui des *big data* d'autre part, se trouve davantage là que dans la taille des bases de données : les données nouvelles sont le plus souvent l'apanage d'acteurs du secteur privé pour lesquels elles constituent des ressources compétitives cruciales. Elles servent notamment à se protéger de la concurrence (actuelle ou potentielle), à attirer des clients et à convaincre les investisseurs. Les plateformes du *web* sont de ce fait prises dans des logiques contradictoires : d'un côté assurer la visibilité publique des données (qui est la raison pour laquelle les utilisateurs les déposent en règle générale) et de l'autre limiter cet accès pour mieux le monétiser auprès du public et entraver toute mise à disposition massive qui pourrait conduire à une perte de l'avantage compétitif si des concurrents pouvaient les exploiter. Les articles d'A. Courmont et B. Kotras, dans ce numéro, mettent bien en avant cette tension et les jeux qui en découlent avec les pouvoir publics et les concurrents moins bien équipés. Les conséquences de cette situation ne peuvent être que néfastes pour la recherche (surtout la recherche publique). Celle-ci ne dispose pas à ce jour d'un statut privilégié lui permettant de réclamer une forme d'accès à ces données. Dans ce domaine, la différence par rapport aux données « classiques » de la statistique publique est donc frappante.

La question de l'accès aux données peut aussi se comprendre comme un problème technique. Les entreprises qui gèrent ces bases de données géantes contrôlent l'accès de trois manières principales. Certains sites *web* et plateformes mettent à disposition du public ou des développeurs des interfaces de programmation (*API*, voir ci-dessus) grâce auxquelles un usager extérieur peut accéder à certains contenus, généralement dans des formats assez structurés et aisément réutilisables. L'*API* de Twitter permet par exemple d'envoyer des requêtes à la base de données en fonction des *tweets* à collecter sur la base de mots-clés (*hashtag*), d'identifiants d'utilisateurs ou d'autres variables comme la localisation géographique. Cette collecte est en revanche limitée en volume et toujours susceptible d'être bridée par Twitter. Lorsqu'aucune *API* n'est mise à disposition par l'entreprise qui contrôle les données, il est possible de programmer des logiciels de *web scraping* et/ou *web crawling* qui parcourent et aspirent les éléments constituants d'une ou plusieurs pages *web*. Ces sources étant généralement peu (ou non) structurées, la difficulté technique est de les transformer en un corpus cohérent et susceptible d'être traité analytiquement. Une troisième méthode (celle utilisée par exemple par M. Bergström) consiste à négocier avec l'entreprise une extraction de sa propre base de données, par définition plus complète et plus structurée que tout ce à quoi une *API* ou du *web scraping* peuvent donner accès.

Les contraintes qui pèsent sur ces trois méthodes principales tendent à s'accentuer dans le temps. Depuis la mi-2012, Twitter a par exemple progressivement limité l'accès à son *API* publique. On ne peut aujourd'hui accéder par cet intermédiaire qu'à un nombre réduit de *tweets* très récents (ceux de la semaine écoulée). Des jeux de *tweets* historiques et volumineux ont parfois été mis à disposition de la recherche par ceux qui les ont collectés, mais ce sont des cas rares car Twitter restreint fortement le droit à partager des données collectées via l'*API*. Par ailleurs, ces jeux de données collectés dans des contextes très

variables et avec des objectifs différents de ceux des chercheurs qui envisagent de les analyser, se prêtent moins bien à la réutilisation en sociologie qu'en informatique. L'alternative existe d'acheter les données auprès de Twitter ou d'un vendeur agréé mais à un coût souvent prohibitif pour la recherche publique.[8] D'autres entreprises proposent de vendre des données, comme Google Maps qui, au moment où nous écrivons ces lignes, a annoncé une hausse importante des tarifs pour l'accès à son *API* [9]. Des obstacles juridiques peuvent aussi empêcher l'utilisation de certaines données comme celles de la plateforme TripAdvisor qui met son *API* à disposition des établissements de tourisme mais exclut expressément son utilisation à des finalités de recherche scientifique[10].

La pratique du *web scraping* ne va pas non plus de soi. Les conditions générales d'usage de certains sites et plateformes interdisent explicitement cette pratique (LinkedIn, Yelp, La Fourchette, etc.). Quant à la négociation d'un accès privilégié aux données, son résultat est évidemment très variable. Elle aboutit parfois à un accord (comme dans le cas de Meetic analysé par M. Bergström) et parfois non (comme dans le cas de Waze dans l'article d'A. Courmont). Même lorsque le résultat final d'une négociation de ce type est positif, celle-ci prend beaucoup de temps au chercheur, suppose une contractualisation complexe et s'accompagne souvent de restrictions en termes d'usage et de compensations financières.

La protection de données personnelles est quelquefois mise en avant pour justifier toutes les limites mises à l'accès aux données du *web*. Les scandales récurrents relatifs à des fuites de ces données – dont celui impliquant Cambridge Analytica est le plus récent au moment où nous écrivons ces lignes – conduisent à douter du bien-fondé de ces explications. Les entreprises du *web* peuvent être plus ou moins vigilantes en matière de protection des données personnelles, mais elles cherchent évidemment aussi à préserver leur avantage compétitif et à monétiser leurs données.

Nombreuses sont les voix qui s'élèvent aujourd'hui en faveur d'une législation qui, tout en respectant le nouveau Règlement général sur la protection des données, favorise le partage de ces informations (Verdier, 2018). La solution envisagée est liée à la réforme en cours du cadre juridique relatif au droit d'auteur : il s'agirait d'autoriser une exception aux droits d'auteur et des producteurs de bases de données pour permettre le *web scraping* (ou d'autres formes de fouille de données et de textes), sans qu'une négociation explicite avec leur détenteur ne soit nécessaire. Le débat sur cette proposition, lancé en France au moment des discussions autour de la loi pour une République numérique de 2016, n'a pas encore abouti et s'est aujourd'hui déplacé au niveau européen. Il n'est pas encore clair de savoir si cette exception potentielle se limiterait aux seules finalités de recherche ou s'étendrait à un éventail plus large de situations. Le « déluge de données » (Hey et Trefethen, 2003) existe sans doute. Mais il ne touche pas toutes les zones de l'espace social de la même façon et profite aujourd'hui plus aux entreprises commerciales qu'à la recherche publique.

Méthodes qualitatives, méthodes quantitatives ?

[8] Voir Justin Littman, « Where to Get Twitter Data for Academic Research », *Social Feed Manager Blog*, George Washington University, September 14, 2017 : *https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data* (consulté le 12 mai 2018).

[9] Google Maps Platform Team, « Introducing Google Maps Platform », Googleblog, *https://mapsplatform.googleblog.com/2018/05/introducing-google-maps-platform.html* (consulté le 2 mai 2018).

[10] TripAdvisor, Content API FAQS : *https://developer-tripadvisor.com/content-api/FAQ/* (consulté le 12 mai 2018).

Les articles parus dans ce numéro utilisent ou évoquent des données très hétérogènes (données administratives, échanges de messages sur des plateformes numériques, données textuelles) mais aussi des méthodes diverses (méthodes qualitatives, quantitatives, revue de littérature). Cette variété peut paraitre surprenante à la lumière de certains des premiers travaux sur les *big data* qui, il y a déjà plus de dix ans, prophétisaient la disparition des outils classiques du sociologue comme l'entretien et l'enquête par questionnaire. Il n'en est rien manifestement. La recherche sociologique s'enrichit de nouvelles données et de nouvelles méthodes, sans pour autant renoncer aux anciennes. C'est la complexité de la situation actuelle qui invite à tirer parti du bagage méthodologique hérité du passé. Puisque l'accès aux données numériques contrôlées par des entreprises privées est limité, il faut en effet s'appuyer sur ce que l'on a pour pouvoir encore faire entendre la voix des sciences sociales. A. Courmont n'aurait pas pu faire apparaitre les enjeux de pouvoir, économiques et politiques, liés à l'usage de Waze sans des entretiens de recherche « classiques ». De même, B. Kotras n'aurait pas pu conclure sur les tensions internes au marché de l'opinion en ligne sans les outils de la recherche sociologique traditionnelle.

D'une certaine manière, on pourrait même conclure que les méthodes qualitatives sont peu mises en cause dans le paysage des *big data*. Elles constituent la seule manière d'accéder à des coins de la réalité sociale peu numérisés, ou dont les traces numériques sont privatisées par de grandes entreprises commerciales réticentes à les partager. Les méthodes qualitatives conservent également un grand intérêt en complément de travaux utilisant, eux, des données numériques : elles constituent une fenêtre dans le ressenti des individus (Kennedy, 2018), dans leur vécu et leur interprétation des pratiques dont les données numériques sont la trace, et qu'il est souvent difficile de comprendre autrement. Aussi, les méthodes qualitatives peuvent nous aider à trancher sur la généralité des résultats obtenus d'une analyse de données numériques qui, on le sait, ne répondent que rarement aux critères classiques de représentativité statistique. Les méthodes qualitatives et mixtes ont donc encore des beaux jours devant elles.

Le socle des méthodes quantitatives utilisées dans les sciences sociales est plus fortement impacté par le développement des *big data*. Ce ne sont pas tellement les « masses » de données en tant que telles qui mettent en cause l'inférence statistique traditionnelle et l'économétrie ; c'est surtout l'entrée sur la scène du *machine learning* comme instrument d'analyse, se prêtant bien au traitement non seulement des données numériques (par exemple celles issues de Twitter, comme en témoigne la note critique de M. Severo et R. Lamarche-Robin) mais aussi de données d'origine administrative (article d'É. Ollion et J. Boelaert dans ce numéro) et même de données d'enquête, pourvu que la taille de l'échantillon soit assez grande (quelques milliers d'observations au moins). Le *machine learning* attire d'autant plus l'attention que les modèles de régression bien connus font l'objet de critiques majeures, surtout en raison des conditions qu'elles imposent pour la validité de l'inférence (l'indépendance des observations par exemple). Comme le disent É. Ollion et J. Boelaert, cette remise en cause n'implique pas nécessairement la disparition des anciennes façons de faire, mais oblige les chercheurs à s'interroger plus en profondeur, à faire des choix méthodologiques moins conventionnels et plus réfléchis. Une piste de recherche prometteuse consiste à explorer les intersections possibles entre *machine learning* et autres méthodes, comme ont commencé à le faire les économistes (Mullainathan et Spiess, 2017), pour systématiser, par exemple, le processus de sélection des variables et de spécification des modèles (Belloni *et al*., 2014 ; Varian, 2014).

Le *machine learning* est d'ailleurs en pleine évolution, et certains de ses développements récents peuvent particulièrement bénéficier aux sciences sociales. On a pu lui reprocher son insistance sur la prévision plutôt que l'explication, comme le rappellent bien M. Severo et R. Lamarche-Perrin. En effet, ses critères de validation reposent non pas sur des tests statistiques (les *p-values*), mais essentiellement sur sa capacité à prédire le résultat d'un modèle, construit sur des données dites d'entrainement, avec des données nouvelles, dites de test. Toutefois, des solutions hybrides se développent aujourd'hui qui intègrent le *machine learning* à l'économétrie pour identifier des effets causaux (Athey, 2017, 2018), voire des modèles qui « apprennent » la causalité (Guyon, 2014 ; Lopez-Paz *et al*., 2015 ; Mooij *et al*., 2016 ; Goudet *et al*., 2017). La crainte de voir les objectifs des sciences (sociales et autres) détournés par les *big data* et le *machine learning* est donc à relativiser. Ce n'est sans doute pas à la fin de la juridiction du sociologue que nous assistons, contrairement à ce que d'aucuns avaient pu prédire, mais à sa profonde transformation dans le champ quantitatif, par hybridation d'approches et de méthodes différentes. La recherche ne peut que bénéficier de ces croisements, qui pourraient conduire à dépasser certaines des limites du passé, et peut-être même à accélérer le rapprochement du qualitatif et du quantitatif par les méthodes mixtes.

Concurrence entre disciplines, ou entre public et privé ?

Les deux notes critiques publiées dans ce numéro font bien apparaitre un autre aspect important de la révolution des *big data* : l'importance de fertilisations croisées entre disciplines. J.-P. Cointet et S. Parasie montrent par exemple comment l'analyse textuelle, déjà bien connue en sociologie, s'est récemment enrichie au contact de l'informatique. M. Severo et R. Lamarche-Perrin mettent quant à eux au jour l'apport des disciplines relevant de l'informatique (surtout dans le champ de la *sentiment analysis*, mais aussi de l'analyse des réseaux) dans la compréhension des mouvements d'opinion sur Twitter. De manière plus générale, les sociologues se trouvent aujourd'hui de plus en plus conduits à partager leurs objets de recherche avec des collègues informaticiens, statisticiens, physiciens. Il est donc légitime de se demander dans quelles conditions ces échanges prennent la forme d'une véritable collaboration (ou a minima de la « coopétition ») plutôt que de la concurrence entre eux.

Les cas observés par J.-P. Cointet et S. Parasie nous amènent à croire que l'arrivée des informaticiens dans le champ de l'analyse textuelle ne visait pas délibérément à concurrencer ou à délégitimer les sciences sociales. Dans une certaine mesure, l'intérêt des informaticiens tient en effet aux opportunités de gain économique pouvant dériver d'une exploitation commerciale de données textuelles (et, plus généralement, numériques) et du développement de services payants sur le marché. Les grandes entreprises du *web* ont fortement investi dans le développement et l'application de ces techniques, et les chercheurs en informatique, même dans le secteur public, sont aujourd'hui ouvertement incités à valoriser leurs résultats.

Cette présence ne nuit pas à la sociologie, et peut même l'enrichir : les deux notes critiques que nous publions montrent bien que des collaborations ont pu avoir lieu, débouchant sur des publications parfois de très grande qualité. Par ailleurs, les sciences sociales pourraient avoir davantage besoin, à l'avenir, de systèmes de calcul parallèles pour traiter des volumes réellement « massifs » de données que des institutions de recherche publique interdisciplinaires permettront de mettre en place alors que seuls les grands acteurs privés en disposent aujourd'hui (Villani, 2018, p. 88-89). La co-écriture peut aussi aider le

sociologue à se positionner rapidement dans le champ, sans nécessairement attendre de se former aux techniques du *machine learning*. Quoique anecdotiques, nos expériences respectives – de sociologues impliqués dans des instituts interdisciplinaires de *data science*, voire dans des laboratoires d'informatique – témoignent aussi de l'ouverture de ces autres disciplines vers la nôtre, malgré les différences de repères, de langage et de pratiques qui nous séparent.

La menace pour la sociologie, s'il y en a une, est plutôt un effet des disparités de pouvoir entre entreprises privées et recherche publique. Les entreprises du numérique se disputent aujourd'hui les meilleurs chercheurs en *machine learning* et en intelligence artificielle, au point de faire redouter une pénurie dans l'enseignement supérieur et la recherche publique (Villani, 2018), alors que peu d'entre elles s'entourent de spécialistes des sciences sociales[11]. C'est dans ce contexte que les craintes de délégitimation de la discipline se concrétisent, par exemple devant la demande de brevet déposée par Facebook et récemment dévoilée d'un « classifieur » (un outil servant à regrouper des données en des catégories homogènes) qui assignerait chacun des usagers de cette plateforme à une classe sociale en croisant un ensemble de données le concernant, mais sans mobiliser les connaissances et théories des sociologues à ce sujet[12]. Le risque est double : que les résultats de ces opérations algorithmiques occultent ceux de la recherche sociologique et qu'ils exercent à terme des effets performatifs aux conséquences imprévues et imprédictibles. Tout changement du fonctionnement des algorithmes de Facebook est en effet susceptible d'affecter l'activité et les comportements de ses millions d'usagers, enclenchant des mécanismes sociaux que même l'analyse *data-driven* des traces du passé peinerait à mettre au jour, le contexte de l'action étant inédit. Aux acteurs de la recherche publique reste alors un devoir de veille et d'alerte, dans la mesure où leur voix peut encore se faire entendre.

Data as labor

Ces évolutions récentes nous interrogent non seulement sur la pertinence de l'adjectif « *big* » mais aussi sur le substantif « *data* ». Littéralement, le mot fait référence à ce qui nous aurait été donné, et ne devrait qu'être saisi et utilisé. La réalité est autre, car loin de se trouver (pour ainsi dire) dans la nature, la donnée est produite. Une enquête statistique est par exemple issue du travail de ses concepteurs, enquêteurs et codeurs, soutenu par des investissements parfois conséquents. Il en va de même pour les données numériques, qui n'existeraient pas sans une importante activité productive humaine. La note critique de M. Severo et R. Lamarche-Robin mentionne des « annotateurs », chargés d'étiqueter manuellement des données d'entrainement pour que la machine puisse s'en servir pour « apprendre » à catégoriser ensuite seule le corpus (en l'occurrence, en assignant chacun des *tweets* d'un corpus à l'une parmi plusieurs catégories de sentiments). Ce n'est qu'après cette phase fortement *labor-intensive* que l'opération de catégorisation devient automatisable (la machine devenant alors capable d'appliquer les mêmes principes de catégorisation à d'autres jeux de données). Les annotateurs ne sont pas toujours des assistants recrutés par les analystes : il existe des plateformes digitales qui proposent ces activités à des foules d'usagers sous la forme de micro-tâches rémunérées à la pièce, comme par exemple Amazon

[11] Avec des exceptions notables, par exemple celle de Microsoft Research à Cambridge (États-Unis).
[12] Demande de brevet US 20180032883 A1, publiée le 1er février 2018.

Mechanical Turk (Gray et Suri, 2017) ou FouleFactory en France. Il est donc nécessaire de considérer aussi la donnée *comme* un travail (Arrieta-Ibarra *et. al*., 2018)[13].

Ce travail de production des données n'est pas toujours rémunéré, ni même reconnu formellement comme tel. À l'apport des annotateurs payés s'ajoute en effet celui de bénévoles comme les usagers de Waze (article d'A. Courmont) qui nourrissent volontairement les bases de données de la plateforme. Cet apport des usagers n'est pas non plus toujours conscient. Google possède par exemple de grandes masses de données tirées des requêtes effectuées sur son moteur de recherche, des clics sur certaines annonces publicitaires, des alternatives suggérées à son service de traduction ou encore de la reconnaissance de caractères dans reCaptcha (von Ahn *et al*., 2008). L'idée selon laquelle cette activité des utilisateurs de plateformes du *web* est un véritable travail (c'est-à-dire une activité productrice de valeur) gagne aujourd'hui du terrain dans les sciences sociales (Cardon et Casilli, 2015 ; Scholz, 2012 ; Terranova, 2000)[14].

Au constat que nous faisions plus haut d'une flagrante inégalité dans la distribution des données il faut donc ajouter que la production de ces données est le résultat d'un système complexe de rapports socio-économiques fortement asymétriques. Cette prise de conscience ne peut qu'interpeller le sociologue dans son rôle d'observateur des transformations du monde du travail, comme en témoigne la littérature en plein essor sur le *digital labor* (Dujarier, 2014 ; Graham *et al*., 2017 ; Neff, 2012 ; Scholz, 2012 ; Vendramin et Valenduc, 2018). Elle l'interpelle aussi en tant que producteur et usager de données. Se constituer une base de données numériques adaptée pour la recherche exige en effet du temps, des compétences (notamment en programmation), et un certain bagage de *soft skills* comme la patience et la créativité requises pour imaginer des moyens d'arpenter ce nouveau terrain qu'est le code, la capacité de négociation ou de collaboration avec des informaticiens et avec les propriétaires des plateformes. La lecture et la discussion des travaux des sciences connexes de la sociologie qui manipulent au moyen d'outils nouveaux des données de grand intérêt pour notre discipline en requièrent tout autant. Aucun des auteurs de ce numéro spécial n'a échappé à cette loi nouvelle des sciences sociales dans leur moment *big data.*

Gilles BASTIN
PACTE
Université Grenoble Alpes, CNRS, Sciences Po Grenoble
BP 48 – 38040 Grenoble cedex 9

gilles.bastin@sciencespo-grenoble.fr

Paola Tubaro
Laboratoire de recherche en informatique
CNRS-Université Paris-Sud
Bât 660 Claude Shannon
Rue Noetzlin – 91190 Gif-sur-Yvette

[13] On pourra se référer à deux numéros spéciaux de revues consacrés récemment au problème général de la production et de l'entretien des données et des bases de données : *Réseaux*, « Sociologie des bases de données », 2013, 178-179 ; *Revue d'anthropologie des connaissances*, « Ce que les data font faire aux SHS (et vice-versa) », 2016, *10*, 4.

[14] Étant donné l'importance de ces questions, leur traitement fréquent dans la littérature internationale et leur politisation récente, nous nous attendions à recevoir de nombreuses propositions d'articles qui les abordent, ce qui n'a pas été le cas. Nous souhaitons évidemment que ce champ de recherche soit à l'avenir davantage exploré par les sociologues.

paola.tubaro@inria.fr

RÉFÉRENCES BIBLIOGRAPHIQUES

ABITEBOUL S., DOWEK G., 2017, *Le temps des algorithmes*, Paris, Éditions Le Pommier.

ANDERSON C., 2008, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired*: *https://www.wired.com/2008/06/pb-theory/* (consulté le 6 juillet 2018).

ARRIETA IBARRA I., GOFF L., JIMÉNEZ-HERNÁNDEZ D., LANIER J., GLEN WEYL E., 2018, « Should We Treat Data as Labor? Moving beyond "Free" », *AEA Papers and Proceedings*, 108, p. 38-42.

ATHEY S., 2017, « Beyond Prediction: Using Big Data for Policy Problems », *Science*, *355*, 6324, p. 483-485.

ATHEY S., 2018, « The Impact of Machine Learning on Economics » dans A. K. AGRAWAL, J. GANS, A. GOLDFARB, *The Economics of Artificial Intelligence: An Agenda*, Chicago (IL), University of Chicago Press [à paraitre].

BASTIN G., FRANCONY J.-M., 2016, « L'inscription, le masque et la donnée. *Datafication* du web et conflits d'interprétation autour des données dans un laboratoire invisible des sciences sociales », *Revue d'anthropologie des connais*sances, *10*, 4, p. 505-530.

BELLONI A., CHERNOZHUKOV V., HANSEN C., 2014, « High-Dimensional Methods and Inference on Structural and Treatment Effects », *Journal of Economic Perspectives*, *28*, 2, p. 1-23.

BEER D., BURROWS R., 2007, « Sociology and, of and in Web 2.0: Some Initial Considerations », *Sociological Research Online*, *12*, 5: *http://www.socresonline.org.uk/12/5/17.html.*

BOULLIER D., 2015, « Les sciences sociales face aux traces du *big data*. Société, opinion ou vibrations ? », *Revue française de science politique*, *65*, 5/6, p. 805-828.

BOYD D., CRAWFORD K., 2012, « Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon », *Information, Communication & Society*, *15*, 5, p. 662-679.

CARDON D., 2015, *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Le Seuil.

CARDON D., CASILLI A. A., 2015, *Qu'est-ce que le Digital Labor* ? Bry-sur-Marne, INA Éditions.

CHERRIER B., 2017, « Big Data in Social Sciences: A Promise Betrayed? », *The Undercover historian blog*: *https://beatricecherrier.wordpress.com/2017/03/22/are-the-promises-of-big-data-in-social-sciences-being-betrayed/* (consulté le 6 juillet 2018).

DE MAURO A., GRECO M., GRIMALDI M., 2016, « A Formal Definition of Big Data Based on its Essential Features », *Library Review*, *65*, 3, p. 122-135.

DESROSIÈRES A., 2005, « Décrire l'État ou explorer la société : les deux sources de la statistique publique », *Genèses*, 58, p. 4-27.

DIEBOLD F., 2003, « "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting: A Discussion of the Papers by Lucrezia Reichlin and by Mark W. Watson » dans M. DEWATRIPONT, L. P. HANSEN, S. J. TURNOVSKY (eds.), *Advances in Economics and Econometrics. Theory and Applications, World Congress, Volume III*, Cambridge, Cambridge University Press, p. 115-122.

DIEBOLD F., 2012, « The Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline », document de travail: *https://www.sas.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf*.

DUJARIER M.-A., 2014, *Le travail du consommateur. De Mac Do à eBay : comment nous coproduisons ce que nous achetons*, Paris, La Découverte.

GOLDBERG A., 2015, « In Defense of Forensic Social Science », *Big Data & Society*, *2*, 2: *https://doi.org/10.1177/2053951715601145*.

GOUDET O., KALAINATHAN D., CAILLOU P., LOPEZ-PAZ D., GUYON I., SEBAG M., TRITAS A., TUBARO P., 2017, « Learning Functional Causal Models with Generative Neural Networks » : *arXiv:1709.05321v2* [stat.ML]: *https://arxiv.org/pdf/1709.05321.pdf*.

GRAHAM M., HJORTH I., LEHDONVIRTA V., 2017, « Digital Labour and Development: Impacts of Global Digital Labour Platforms and the Gig Economy on Worker Livelihoods », *Transfer: European Review of Labour and Research*, *23*, 2, p. 135-162.

GRAY M. L., SURI S., 2017, « The Humans Working behind the AI Curtain », *Harvard Business Review*: *https://hbr.org/2017/01/the-humans-working-behind-the-ai-curtain.*

GUYON I., 2014, Chalearn Fast Causation Coefficient Challenge: *https://www.codalab.org/competitions/1381* (consulté le 6 juillet 2018).

HEY H., TREFETHEN A., 2003, « The Data Deluge: An e-Science Perspective » dans F. BERMAN, G. FOX, T. HEY (eds.) *Grid Computing – Making the Global Infrastructure a Reality*, Wiley and Sons, chap. 36.

KENNEDY H., 2018, « How People Feel about What Companies Do with their Data is Just as Important as What they Know about it », *LSE Impact Blog*, March 29: *http://blogs.lse.ac.uk/impactofsocialsciences/2018/03/29/how-people-feel-about-what-companies-do-with-their-data-is-just-as-important-as-what-they-know-about-it/* (consulté le 6 juillet 2018).

KITCHIN R., 2014, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, London, Sage.

LANEY D., 2001, « 3-D Data Management: Controlling Data Volume, Velocity and Variety », *META Group Research Note*, February 6: *https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf* (consulté le 6 juillet 2018).

LOPEZ-PAZ D., MUANDET K., SCHÖLKOPF B., TOLSTIKHIN I. O., 2015, « Towards a Learning Theory of Cause-Effect Inference », *ICML*, p. 1452-1461.

MAYER-SCHÖNBERGER V., CUKIER K., 2013, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston (MA), Houghton Mifflin Harcourt.

MERZEAU L., 2009, « Du signe à la trace : l'information sur mesure », *Hermès, La Revue*, 53, p. 21-29.

MOOIJ J. M., PETERS J., JANZING D., ZSCHEISCHLER J., SCHÖLKOPF B., 2016, « Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks », *Journal of Machine Learning Research*, *17*, 32, p. 1-102.

MULLAINATHAN S., SPIESS J., 2017, « Machine Learning: An Applied Econometric Approach », *Journal of Economic Perspectives*, *31*, 2, p. 87-106.

NEFF G., 2012, *Venture Labor. Work and the Burden of Risk in Innovative Industries*, Cambridge (MA), The MIT Press.

O'REILLY T., STEELE J., LOUKIDES M., HILL C., 2012, *How Data Science Is Transforming Health Care. Solving the Wanamaker Dilemma*, Sebastopol, O'Reilly Media.

PASSERON J.-C., 1991, *Le raisonnement sociologique. L'espace non-poppérien du raisonnement naturel*, Paris, Nathan.

PENTLAND A., 2012, « Reinventing Society in the Wake of Big Data », *Edge*: *https://www.edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data* (consulté le 6 juillet 2018).

ROUVROY A., BERNS T., 2013, « Gouvernementalité algorithmique et perspectives d'émancipation », *Réseaux*, 177, p. 163-196.

SAVAGE M., BURROWS R., 2007, « The Coming Crisis of Empirical Sociology », *Sociology, 41*, 5, p. 885-899.

SCHOLTZ T., 2012, *Digital Labor. The Internet as Playground and Factory*, Abingdon, Routledge.

SHADBOLT N., VERDIER H., 2015, « La révolution de la donnée au service de la croissance. Innovation, infrastructure, compétences et "Pouvoir d'agir" à l'ère numérique », *Rapport du groupe de travail franco-britannique sur l'économie de la donnée* : *https://www.economie.gouv.fr/files/files/PDF/rapport-taskforce_fr.pdf* (consulté le 6 juillet 2018).

TERRANOVA T., 2000, « Free Labor: Producing Culture for the Digital Economy », *Social Text, 18*, 2, p. 33-58.

TUBARO P., CASILLI A.A., SARABI Y., 2014, *Against the Hypothesis of the End of Privacy*, Zurich, Springer.

VARIAN H. R., 2014, « Big Data: New Tricks for Econometrics », *The Journal of Economic Perspectives*, *28*, 2, p. 3-27.

VENDRAMIN P., VALENDUC G., 2018, « Gigabits et microjobs – L'expansion des petits boulots dans l'économie digitale » dans M. SOMERS (éd.), *Vorm geven aan digitale tijden*, Antwerpen, Minerva, p. 78-95.

VERDIER H., 2018, *La donnée comme infrastructure essentielle. Rapport au premier ministre sur la donnée dans les administrations 2016-2017*, Paris, La Documentation française.

VILLANI C., 2018, « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », *Rapport de la mission confiée par le Premier Ministre*, Paris, La Documentation française.

VON AHN L., MAURER B., MCMILLEN C., ABRAHAM D., BLUM M., 2008, « reCAPTCHA: Human-Based Character Recognition via Web Security Measures », *Science*, *321*, 5895, p. 1465-1468.