Au-delà des *big data*
Les sciences sociales et la multiplication des données numériques

Beyond big data
Social sciences and the proliferation of digital data

par **Étienne Ollion*** et **Julien Boelaert****

RÉSUMÉ

Dans le débat public comme dans le monde académique, l'enthousiasme pour les *big data* n'a eu d'égal que les critiques que ce phénomène a suscité. « Opportunité empirique inouïe » *vs* « données pauvres » ; « révolution méthodologique » *vs* « fascination pour le nombre » ; « révolution scientifique » *vs* « dégradation du savoir produit » : les positions sont tranchées. À partir d'une lecture de ces débats et des travaux en sciences sociales souvent regroupés sous ce label, l'article soutient que cette situation polarisée a de fortes chances de perdurer tant que la discussion s'organise autour du concept mal défini de *big data*. Il propose de distinguer différents types de données souvent regroupées sous ce terme. Il montre ce faisant que les *big data* souvent évoquées ne sont qu'un aspect limité d'une transformation bien plus importante : la disponibilité croissante et massive de données numériques, qui pose des questions nouvelles à nos disciplines. Quatre aspects sont plus particulièrement explorés : les réorganisations disciplinaires, les transformations des méthodes quantitatives, l'accès et la gestion des données, les objets des sciences sociales et leur rapport à la théorie.

ABSTRACT

In the public sphere as in academia, the frenzy over big data has been met with equally high levels of enthusiasm and criticism. "Unprecedented empirical opportunity" vs. *"poor data"; "methodological revolution"* vs. *"fascination for large N"; "scientific revolution"* vs. *"debasement in the production of knowledge": the positions are polarized. Based on a review of the debates in the social sciences, this paper shows that this situation is likely to persist as long as the discussion is organized around the fuzzy term of "big data." Instead, this paper proposes to distinguish between different types of data, each of which raises specific questions. By so doing, it shows that big data is but one aspect of a much wider transformation –the massive availability of digital data– which in turn raises new questions for our disciplines. Four aspects are subsequently explored: disciplinary reorganizations, the evolutions of statistical methods, the access and management of data, as well as the objects of social sciences and their relationship to theory.*

MOTS-CLÉS : *big data*, méthodes, statistiques, données numériques, sciences sociales

KEYWORDS: *Big data, methods, statistics, digital data, social sciences*

* Chargé de recherche au CNRS, UMR 7363 SAGE (Sociétés, Acteurs, Gouvernement en Europe), université de Strasbourg
Laboratoire SAGE, université de Strasbourg, 5 allée du Général Rouvillois, CS 50008, 67083 Strasbourg cedex, France
etienne.ollion@unistra.fr
** Post-doctorant à l'ENS-Cachan
École normale supérieure de Cachan, laboratoire STEF, bâtiment Cournot, 61 avenue du Président Wilson, 94235 Cachan cedex, France
julien.boelaert@gmail.com

Les *big data* sont à la mode[1]. Dans les médias comme dans le débat public, le terme est évoqué de manière récurrente. Des entreprises dédiées au traitement de ces masses d'informations sont créées chaque jour, d'autres réorganisent radicalement leurs activités autour de la gestion de ces données. Le domaine semble suffisamment porteur pour que, dans de nombreux pays, les pouvoirs publics investissent massivement dans ce domaine. À la suite d'entreprises privées, les administrations de nombreux pays sont désormais dotées d'un responsable de haut rang dédié à cette thématique. *Chief data officer* aux États-Unis, *Chief digital officer* responsable du *Government Digital Service* en Grande-Bretagne, *Administrateur général des données* en France : ils sont en charge de s'assurer que le pays valorise ses données, présentées comme une source de croissance inexplorée.

Le même phénomène est à l'œuvre dans le monde académique. Dans la lignée de centres pionniers comme le *Media Lab* créé en 1985 par le *Massachussetts Institute of Technology*, de nombreuses universités ont créé des centres dédiés à l'analyse des grands volumes de données. Des conférences sur le thème sont régulièrement organisées. Des postes destinés à des spécialistes de « *big data* » sont créés chaque année, et leurs titulaires sont amenés à enseigner dans des cursus de « *data science* », qui promettent à leurs étudiants de tirer profit des opportunités de ce phénomène. Des numéros de revues sont publiés, des revues spécialisées voient aussi le jour. En sciences sociales, c'est le cas de *Big data and Society*. Lancée en 2013, son objectif est d'étudier les « *big data* et leurs implications pour les sociétés[2] ». Elle accueille des travaux dans les sciences sociales et les humanités qui s'intéressent à ces thématiques, tout comme elle se fait l'écho de controverses.

C'est que, si les *big data* intéressent, elles font aussi débat. L'enthousiasme des partisans de ces techniques n'est en effet pas toujours partagé. Suite aux révélations répétées de surveillance massive des populations, plusieurs chercheurs ont souligné les dangers de ces méthodes, qui peuvent rapidement se transformer en – voire sont consubstantiellement de – puissantes technologies de gouvernement (Harcourt, 2014). Du point de vue de la recherche même, nombreux sont ceux qui ont pointé les limites de ces approches. Anthropologue spécialisée dans le domaine des nouvelles technologies, pionnière des usages des techniques numériques, danah boyd écrivait, dès 2011, que les informations collectées avec ces méthodes étaient problématiques. Leur commensurabilité – entre elles, avec d'autres sources –, leur traitement – et les difficultés techniques que posent la masse et l'hétérogénéité des informations – ou encore les questions d'éthique étaient autant de problèmes encore non résolus (boyd & Crawford, 2012). En France, d'autres critiques provenant de chercheurs eux aussi peu suspects de défiance instinctive vis-à-vis des nouvelles technologies ont depuis vu le jour. Dominique Cardon et plusieurs collègues spécialistes de l'internet ont critiqué l'idée selon laquelle une avalanche de données pouvait en soi améliorer la connaissance (Bastard *et al.*, 2013). Les données sont toujours construites par une opération et elles doivent aussi toujours être interprétées. Pour ce faire, il faut disposer d'une connaissance précise du domaine étudié. Cette remarque dissipe le fantasme d'un outil de connaissance transposable immédiatement d'un objet à l'autre.

Une discussion s'est progressivement installée, dans les sciences sociales comme dans les sciences en général, qui interroge la pertinence du phénomène pour la recherche. De manière croissante, les chercheurs sont aussi invités à se positionner par rapport à un ensemble de questions générales : les *big data* constituent-ils une révolution pour les disciplines ? Vont-ils transformer nos manières d'étudier et de connaître ? Aussi intéressantes que soient ces discussions, aussi justes que soient certaines des remarques qui sont faites, il nous semble que la question ainsi posée ne peut recevoir de réponse satisfaisante. La raison principale tient au concept même de *big data*, tellement vague qu'il empêche toute discussion précise. Terme à l'origine contestée[3], il n'en existe pas de définition stabilisée. Une caractérisation fréquente, dite des « 3 V »,

1. Cet article a bénéficié des commentaires, critiques et suggestions d'un grand nombre de personnes, ainsi que de ceux des évaluateurs et membres du comité de rédaction de la revue *Sociologie*. Que toutes et tous soient chaleureusement remerciés de leur aide. Les erreurs comme les interprétations restent évidemment nôtres.

2. http://bigdatasoc.blogspot.fr/p/big-data-and-society.html (consulté le 30 avril 2015).

3. C'est la conclusion d'une enquête menée par un journaliste étatsunien : http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

propose de les définir comme des données qui croiseraient trois propriétés : un volume (important), une variété (des types de données) et une vélocité de la captation (permanente et/ou immédiate). À la fois trop générale et trop restrictive pour bien décrire les pratiques dans ce domaine, elle est régulièrement critiquée. D'autres définitions voient régulièrement le jour.

Ces controverses sémantiques sont le signe d'un autre investissement, pécunier cette fois, et autrement plus important. Les technologies de collecte et de traitement de données ont fait l'objet de dépenses massives ces dix dernières années, en technologie comme en publicité. Un marché à vu le jour, sur lequel sont engagés une multitude d'acteurs qui tous ont intérêt à entretenir l'idée qu'une révolution des *big data* est en cours. La polarisation récente des débats n'a fait qu'accroître cette indétermination. Terme étendard, les *big data* sont désormais une marque et une promesse commerciale avant que d'être un concept ou de désigner une réalité de la pratique des sciences.

À partir d'une revue de littérature sélective et de notre expérience d'usagers de méthodes souvent classées sous ce label englobant, cet article propose une lecture critique et une précision conceptuelle. Plutôt que de parler indistinctement de *big data*, l'article propose de distinguer différentes sources de données numériques, un terme plus large mais finalement bien plus précis pour évoquer ce phénomène. Ces distinctions ainsi opérées, on est alors en mesure de revenir sur la question des changements induits par la multiplication des données numériques pour les sciences sociales. L'article soutient que des mutations profondes sont bien en cours, et suggère certaines dimensions où la multiplication des données numériques modifie les pratiques de recherche. Quatre sont distinguées : les réorganisations disciplinaires, les transformations des méthodes quantitatives, l'accès et la gestion des données, les objets des sciences sociales et leur rapport à la théorie.

Promesses et faiblesses des *big data*

Dans leur acception la plus générale, les *big data* désignent ces vastes ensembles de données, dont l'existence comme le traitement rapide ont été rendus possibles par une série de changements technologiques. Le développement d'internet, medium par lequel de nombreuses activités humaines peuvent être enregistrées, la multiplication des capteurs de toute sorte et l'informatisation croissante des organisations font que l'on peut disposer d'informations précises, récurrentes et massives sur d'innombrables pratiques. L'augmentation de la puissance des ordinateurs et l'apparition de nouvelles méthodes de traitement font qu'il est possible de les analyser rapidement. C'est l'existence et plus encore l'abondance de ces données qui a donné lieu aux déclarations enthousiastes sur les *big data* et l'avenir de la connaissance.

Des promesses importantes

Comme dans d'autres domaines, les *big data* ont suscité un réel intérêt en sciences sociales. Nombreux sont les chercheurs qui s'en sont saisis, attirés par les promesses du phénomène. Celles-ci étaient de trois ordres : empirique, méthodologique et théorique. Du point de vue *empirique*, la possibilité d'accéder à d'immenses ensembles de données a été accueillie avec enthousiasme par beaucoup. Des informations inédites et extensives devaient permettre d'étudier des sujets jamais appréhendés jusque-là, et surtout de connaître avec une précision infiniment plus grande des aspects entiers du comportement humain. Les exemples ne manquent pas, qui montrent les applications possibles de ces données. Les études de mobilité ont ainsi bénéficié de la collecte régulière et automatique de données *via* des capteurs RFID, ces puces de radio-identification qui enregistrent les informations relatives aux objets – ou aux personnes, aux animaux – auxquelles elles sont attachées et peuvent les transmettre régulièrement vers un serveur qui les collecte. Installées sur les cartes d'abonnement de transports en commun, elles fournissent un tableau jusque-là inaccessible des mobilités urbaines. Les données de téléphone portable sont encore plus précises. Depuis quelques années, des démographes étudient les déplacements individuels par ce biais. Ils soulignent ainsi que les informations issues de ces appareils permettent d'obtenir des informations riches et fiables sur les trajets d'une population (Williams *et al.*, 2013). Des expérimentations montrent ainsi qu'on peut, en fonction du trajet et de la vitesse de déplacement, inférer le mode de transport (voir la synthèse que propose Eagle, 2010). Outre l'échelle (un pays), on peut donc connaître les mobilités de populations entières avec un luxe de détails impensable jusqu'alors. Ces données peuvent par ailleurs être actualisées, ce qui limite le risque classique de la péremption que connaissent les enquêtes par questionnaire.

D'autres domaines ont bénéficié de ces avancées empiriques. Des sociologues de la conjugalité se sont intéressées aux

pratiques sur les sites de rencontre. Suite à des partenariats avec ces prestataires, elles ont eu accès aux données d'interaction entre utilisateurs. Elles ont ainsi pu étudier non les unions réalisées, mais plutôt le couple en train de se faire (Bergström, 2014 ; Potârcă & Mills, 2015). La recherche en économie a aussi bénéficié des masses d'informations disponibles en ligne[4]. Le *Billion Prices Project*, coordonné par Alberto Cavallo, propose un indice des prix à la consommation fondé sur l'extraction automatique des prix sur des centaines de sites web. Calculé avec les mêmes méthodes que son homologue officiel étatsunien du *Bureau of Labor Statistics*, il s'appuie sur un panier de biens représentatifs pour saisir l'inflation. Aux États-Unis, les deux indices suivent des variations très semblables. Dans d'autres pays dotés d'un moindre appareillage statistique, cet indice calculé en ligne semble désormais plus performant, plus rapide et moins coûteux (Cavallo, 2013).

Du point de vue *méthodologique*, les avancées potentielles sont tout aussi significatives. La plus saillante concerne le gain de précision que ces méthodes permettent d'obtenir. L'automatisation des enregistrements et l'accélération des traitements font que sur de nombreux sujets, il est désormais quasiment aussi rapide d'étudier la population entière qu'un échantillon restreint. Le changement est significatif, puisque c'est la question de la représentativité qui cesse de se poser quand l'un et l'autre sont similaires – les statisticiens disent « N = All ». Les études qui font usage de données bibliométriques en sont un bon exemple. Acte routinier de l'écriture scientifique, la citation est depuis longtemps utilisée pour, en sociologie des sciences, analyser le champ scientifique (de Solla Price, 1986). Riches, ces données sont aussi rapidement trop nombreuses pour mener une enquête exhaustive. Les chercheurs avaient en effet jusqu'à peu à faire des choix : ils travaillaient sur un auteur, sur la diffusion des citations d'un article dans une ou deux revues. Ils devaient ensuite justifier ce choix et expliquer le biais qu'il introduisait. Quand les sources existent, ce qui est le cas pour la littérature scientifique, cette question ne se pose plus. Dans un article qui revisite un moment de l'histoire des sciences – la découverte de la double hélice de l'ADN –, Yves Gingras montre que l'article pionnier du domaine a connu des débuts bien plus rapides que ce qu'en disaient les historiens des sciences jusqu'alors. Plutôt que de se concentrer sur une revue, ou de tirer des conclusions à partir d'un nombre limité de cas, Y. Gingras a pu suivre la référence à cet article dans le temps et dans différents espaces, avant de le comparer aux débuts d'autres découvertes (Gingras, 2010). L'étude est à la fois minutieuse et exhaustive, des qualités rarement combinées mais permises par la multitude de données utilisables.

Côté *théorique* enfin, les *big data* ont suscité de nombreux espoirs. L'abondance nouvelle de données devait permettre de faire progresser la connaissance. C'est la revendication principale du courant *Culturomics,* dont les auteurs ont étudié d'immenses corpus de textes publiés grâce au travail de numérisation de trente millions de livres réalisé par Google[5]. Le volume de données, pourvu qu'il soit suffisant, doit selon eux permettre d'apporter des réponses à des questions aussi larges que l'évolution des langues ou des représentations. Dans un article manifeste du mouvement (Michel *et al.*, 2011), ses principaux promoteurs affirment qu'on peut désormais suivre l'évolution de mots dans des langues – revisitant ainsi la linguistique –, repérer des moments de censure – et ainsi quantifier l'histoire –, ou faire des recherches dans la mémoire collective en étudiant la variation des références – de Churchill aux cellules souches, en passant par Dieu).

L'ambition dépassait toutefois la simple découverte empirique ou les approfondissements théoriques locaux. Pour plusieurs auteurs, les *big data* allaient changer le regard porté sur la société en général. Erez Aiden et Jean-Baptiste Michel écrivent à propos des *big data* que « leurs conséquences vont modifier la façon dont nous nous percevons. [...] Les *big data* vont changer les humanités, transformer les sciences sociales » (Aiden & Michel, 2013, p. 8), une affirmation aussi avancée par d'autres (Mayer-Schönberger & Cukier, 2014). La raison de ce changement se trouve dans la masse de données. Les auteurs espèrent y trouver des régularités inconnues jusqu'alors. Partant, ils espèrent aussi en faire émerger des lois de fonctionnement du monde social. Invisibles des acteurs, ces lois le seraient aussi des chercheurs en sciences sociales, trop proches de leurs objets pour discerner les tendances de fond qui régissent les sociétés. Elles seraient, en revanche, accessibles à ceux capables de faire parler les données et de mettre au jour les lois cachées du monde social. Le titre de l'ouvrage

4. Voir (Einav & Levin, 2014) pour une recension des travaux récents.

5. Deux responsables du projet racontent cette expérience dans un ouvrage (Aiden & Michel, 2013).

de Sandy Pentland, qui étudie les processus de diffusion d'idées à grande échelle, exprime bien cette ambition nomothétique renouvelée : *Social Physics. How Good Ideas Spread – The Lessons from a New Science* (Pentland, 2014).

Affinités et sociabilité, mobilité, actions individuelles et dynamiques de groupe, formation des prix et transformations des marchés ne sont que quelques sujets qui ont été revisités suite à l'usage de ces méthodes. En quelques années, nombre de questions centrales aux sciences sociales trouvaient des terrains d'expérimentation nouveaux et prometteurs, et offraient de nouvelles perspectives théoriques.

Des réalisations imparfaites

Avec quelques années de recul, on ne peut que constater que les promesses n'ont pas toutes été tenues. Trop ambitieuses, mal adaptées aux informations récoltées, elles ne pouvaient d'ailleurs pas révolutionner les sciences sociales intégralement. Les mêmes critères évoqués ci-dessus peuvent être repris. Du côté empirique, les données récoltées *via* les techniques d'enregistrement automatique se sont avérées bien plus pauvres que prévues. Pour précises qu'elles sont, les données de capteurs ne mesurent souvent qu'une partie de ce qui intéresse les chercheurs en général. Produites à des fins autres que la recherche, elles n'apportent pas toujours de réponses aux questions que se posent les chercheurs. La quantité d'informations sur le suivi d'un dossier client (la dernière connexion, la dernière commande, l'accessibilité d'un historique) ne forme pas nécessairement des variables intéressantes pour le scientifique. Du point de vue de la recherche, *big data* ne veut pas nécessairement dire *rich data*. De la même manière, la connaissance des flux de mobilité urbains *via* les transports en commun ne renseigne finalement que peu sur les modes de déplacement des habitants des villes s'ils ne peuvent être croisés avec d'autres informations – lieu de travail et de résidence, usage d'autres modes de transport, propriétés sociales des enquêtes. Assez rapidement, l'information peut s'avérer plus limitée que si le chercheur avait eu recours à une enquête par questionnaire à la taille limitée. Par ailleurs, la réappropriation de ces données demande souvent, outre leur extraction et leur nettoyage, des procédures de recodage aussi longues qu'appuyées par des connaissances précises du sujet étudié. Si elles font le quotidien des chercheurs, ces contraintes limitent les grandes ambitions de développer des outils qui informeraient quasi-immédiatement sur le monde social.

Les promesses méthodologiques se sont elles aussi révélées exagérées car, en l'absence d'exhaustivité, l'abondance de données n'est pas nécessairement déterminante. À cet égard, l'exemple de George Gallup reste d'actualité. En 1936, le récent fondateur de la première entreprise de sondage avait correctement prédit l'issue de l'élection présidentielle étatsunienne à partir d'un échantillon de quelques milliers de personnes, là où la revue *The Literary Digest* annonçait le résultat inverse sur la base de plusieurs millions de réponses volontaires de ses lecteurs (Didier, 2008). La leçon de théorie des sondages vaut toujours à l'heure actuelle : on connait souvent mieux les pratiques d'une population avec un échantillon bien construit de 1000 personnes qu'avec une base de plusieurs centaines de fois ce chiffre, mais composée sans principe. Ce problème méthodologique n'est pas le seul qui a traversé les travaux qui se revendiquent des *big data*. Assez rapidement, une sorte de « quantophrénie » s'est aussi développée. Ce terme, proposé par le sociologue Pitirim Sorokin (1956) dans un autre contexte d'augmentation massive des volumes de données disponibles pour la recherche – celle de la multiplication du *survey research* dans la sociologie étatsunienne[6] –, désigne l'appétence immodérée pour les chiffres, indépendamment de ce qu'ils peuvent nous apprendre. La disponibilité croissante de données produites pour le fonctionnement de services – plutôt que pour les besoins de la recherche[7] – a conduit les chercheurs à privilégier des sources existantes à la production d'enquêtes.

Sur le plan théorique enfin, les déclarations annonçant une nouvelle science, qu'il s'agisse d'une « physique sociale » ou d'une « science des données », qui révolutionnerait les pratiques des chercheurs avaient été accueillies avec un certain attentisme. Il faut dire que les chercheurs en sciences sociales sont habitués à voir resurgir épisodiquement ce genre de prophéties : depuis Adolphe Quételet, qui proposait lui aussi un *Essai de physique sociale* (1830), de tels manifestes sont

6. Voir (Abbott & Sparrow, 2005) sur ce point.

7. Cette distinction fait écho à celle proposée par le directeur du *Census* étatsunien, qui distinguait les « *designed data* » conçues pour l'enquête aux « *organic data* », produites à des fins autres mais potentiellement mobilisées. Voir http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/ (consulté le 9 mai 2015).

régulièrement publiés sans que les sciences sociales en soient profondément modifiées. C'est que la découverte de structures cachées – si elles existent – demande bien plus que de vastes ensembles de données et des outils rapidement transposables. Elle demande aussi une connaissance de première main de leur sujet et un cadre théorique certes flexible mais déjà établi. L'histoire du décodage du génome humain le rappelle : les chercheurs n'ont pu mettre en place les procédures de décodage que parce qu'ils avaient des attentes relativement précises quant à ce qu'ils allaient trouver. De ce point de vue, les réalisations de projets comme *Culturomics* – indéniables par ailleurs –, se trouvent moins du côté de la révolution théorique qu'ils auraient produite que des informations empiriques sur tel ou tel aspect de l'évolution des langues ou de l'histoire qu'ils donnent à voir[8].

À tous les niveaux, les limites présentées par ces approches sont donc importantes, d'autant plus qu'elles s'ajoutent à d'autres, comme le fait que nombre d'objets ne sont pas justiciables d'analyses quantifiées, ou plus prosaïquement que des pans entiers de la vie sociale ne sont pas traçables. Près d'une décennie après son commencement, le débat sur la pertinence des *big data* pour les sciences sociales se poursuit. C'est, nous semble-t-il, parce que les termes du débat sont mal posés.

Des *big data* aux données numériques

Dans les différentes discussions sur les *big data*, le terme vient regrouper tout un ensemble de pratiques assez hétéroclite. En particulier, on trouve souvent rangé sous ce terme des informations aux origines diverses. Données de l'internet, données de capteurs, *open data*, données d'organisations qui collectent des traces d'activités, questionnaires auto-administrés qui permettent d'obtenir des centaines de milliers de réponses, voire archives numérisées : chacune gagne à être distinguée afin de mieux comprendre ses intérêts propres, comme les limites qu'elle pose. Cette clarification faite, il apparaît évident que les *big data*, si on doit conserver ce terme, ne sont qu'un aspect relativement limité d'une transformation plus générale : la multiplication et l'accessibilité croissantes de données numériques.

Des données et leurs sources

Un premier ensemble de données souvent rangées sous ce label englobant correspond aux *données de l'internet*, ces informations qui sont collectées, ou auxquelles on accède, *via* le Web. La distinction mérite d'être faite : sur internet, on trouve des informations relatives à des pratiques en ligne, c'est le cas de tous les sites sur lesquels les usagers laissent des traces au moment de leur passage. Elles peuvent aussi, et ce n'est pas exactement la même chose, être le reflet plus ou moins précis de pratiques qui ont lieu *offline*. Les informations relatives à la présence de commerces dans tel ou tel quartier, les annuaires des professionnels de santé – utiles pour faire des cartes de déserts médicaux –, ou encore les innombrables informations disponibles sur les personnes – et qui peuvent faire l'objet de traitements prosopographiques – n'est sont qu'un exemple.

Second type de données qui sont parfois rangées sous le label commode de *big data* : les *données produites par des organisations* (administrations, entreprises, associations) dans le cadre de leur fonctionnement. Elles recensent de très nombreuses informations : membres, commandes et clients, budgets, détails de l'activité des services, etc. De longue date, elles ont constitué une source utile pour les chercheurs en sciences sociales. Les économistes ont eux largement recours à des données commerciales pour étudier le marché d'un bien donné. Avec les historiens, ils ont depuis longtemps recours aux données fiscales pour étudier l'évolution de la richesse dans un pays. Dans certains pays, les sociologues et les démographes ont même accès aux données des organismes de protection sociale. C'est le cas en Belgique, où une collaboration entre chercheurs et administrations s'est mise en place depuis 1999 au sein de la « banque carrefour de la sécurité sociale »[9], qui regroupe les informations de plusieurs dizaines d'organismes. L'objectif de conseil et d'évaluation des politiques publiques n'empêche pas la réalisation de recherches moins appliquées, qui bénéficient de la forte granularité comme de la quasi exhaustivité de ces informations (Knapen *et al.*, 2014). Ces données sont propriétaires pour la plupart, et ne sont pas toujours disponibles en ligne – elles le sont même rarement dans le cas des données personnelles. Les sociologues de la conjugalité, citées plus haut, ont eu accès aux bases de données des

8. Paradoxalement, ces projets à l'ambition nomothétique affirmée semblent même avoir déplacé les sciences sociales vers un pôle plus idéographique.

9. Voir www.bcss.fgov.be/nl/dwh/homepage/index.html (consulté le 1er mai 2015).

sites de rencontre, mais ces dernières ne sont pas disponibles sur internet. Créées en ligne, elles sont stockées dans les serveurs des entreprises qui les gèrent.

Au sein de cet ensemble, les données qui sont mises délibérément à la disposition du public rejoignent les rangs chaque jour grossissant des *open data*. Si le label regroupe en théorie des données produites par une organisation et rendues globalement disponibles, ce sont généralement les données organisationnelles accessibles *via* internet que l'on désigne ainsi. La consultation des différents portails qui permettent d'accéder à ces informations témoignent de leur richesse. Les jeux de données qu'on y trouve sont extrêmement hétéroclites : une liste de crèches dans une ville côtoie celle des crimes au cours de la décennie précédente, ou le budget d'une administration[10]. Comme n'ont de cesse de le rappeler les défenseurs de l'*open data*, la plupart des informations produites par des organisations demeurent néanmoins inaccessibles. C'est évidemment le cas pour celles des administrations publiques, cible principale de leurs mobilisations. C'est *a fortiori* aussi le cas des données d'entreprises privées ou d'associations, dont une infime minorité seulement est accessible en ligne.

Il est un autre type de données, parfois rangées un peu vite sous le label de *big data* mais qui gagne à être précisé : toutes les *archives initialement non numériques, mais qui ont été converties*. C'est le cas des documents que l'on scanne, parfois en très grand nombre, afin de leur faire subir des traitements, statistiques ou autres. L'amélioration constante des procédures de reconnaissance automatique de texte (*Optical Character Recognition*) fait qu'il est désormais possible de transformer des sommes d'archives en autant de fichiers lisibles par un logiciel de traitement de texte. Le courant *Culturomics*, évoqué ci-dessus, s'appuie sur ce genre de matériau. Des millions de documents initialement non numériques ont été numérisés par Google – on parle de données « non nativement numériques » –, à partir desquels les études sont menées.

Il faut enfin évoquer les *données issues de questionnaires*, avec lesquels les sociologues travaillent régulièrement et dont les résultats sont souvent identifés aux *big data*. La diffusion d'internet a multiplié la possibilité de passation de questionnaires en ligne. Leur validité reste sujette à d'âpres discussions (Dilman *et al.*, 2014), mais les questionnaires auto-administrés font désormais partie de la boite à outils de plusieurs chercheurs. Ils peuvent être appliqués à une population très nombreuse. C'est par exemple le cas de l'enquête sur les classes sociales en Grande-Bretagne. Mené au début des années 2010 *via* un dispositif placé sur internet et publicisé par de nombreux médias, le *Great British Class Survey* proposait aux enquêtés de répondre à une centaine de questions relatives à leur patrimoine et leurs revenus, leurs pratiques de consommation, leur santé ou leurs relations amicales. Entre 2011 et 2013, 160 000 personnes ont participé à l'enquête, et les résultats ont donné lieu à de multiples publications (Savage *et al.*, 2013).

Faut-il vraiment qualifier toutes ces données de *big data* ? Sans entrer dans des luttes de classement qui chercheraient à en donner une définition fixe, il semble raisonnable de conserver cette appellation pour les cas où la quantité et/ou la diversité des données dépasse très largement les standards habituels de nos disciplines. Sinon, pourquoi seraient-elles *big* ? Le seuil précis importe peu, mais il est utile de conserver à l'esprit que l'Enquête Emploi de l'INSEE porte sur 67 000 foyers, soit environ 100 000 personnes qui répondent à plusieurs centaines de question, et ce de manière répétée. En termes d'informations, c'est finalement beaucoup plus que dans nombre d'études évoquées ci-dessus. En d'autres termes, si on veut conserver le terme de *big data* – le faut-il ? –, il faut probablement le restreindre à une petite partie des données évoquées ci-dessus, où les volumes et la diversité des sources sont particulièrement importants.

Parce qu'elle est sûrement moins intéressante que les recherches qui se déroulent à l'heure actuelle dans ce sous-champ, et parce qu'elle donnera lieu à des batailles sémantiques pendant longtemps, la question de la bonne définition des *big data* n'est finalement pas si centrale. Le vrai enjeu est ailleurs. L'intérêt médiatique et les mutations technologiques ne doivent, en effet, pas faire oublier les perspectives offertes à la recherche par le développement d'un autre ensemble, autrement plus vaste : celui des données numériques. Le terme de numérique, et son équivalent anglais de *digital*, ne signifie pas qu'elles ont spécifiquement trait à internet, même si le

10. Voir par exemple : http://data.gouv.fr ou https://data.cityofchicago.org/.

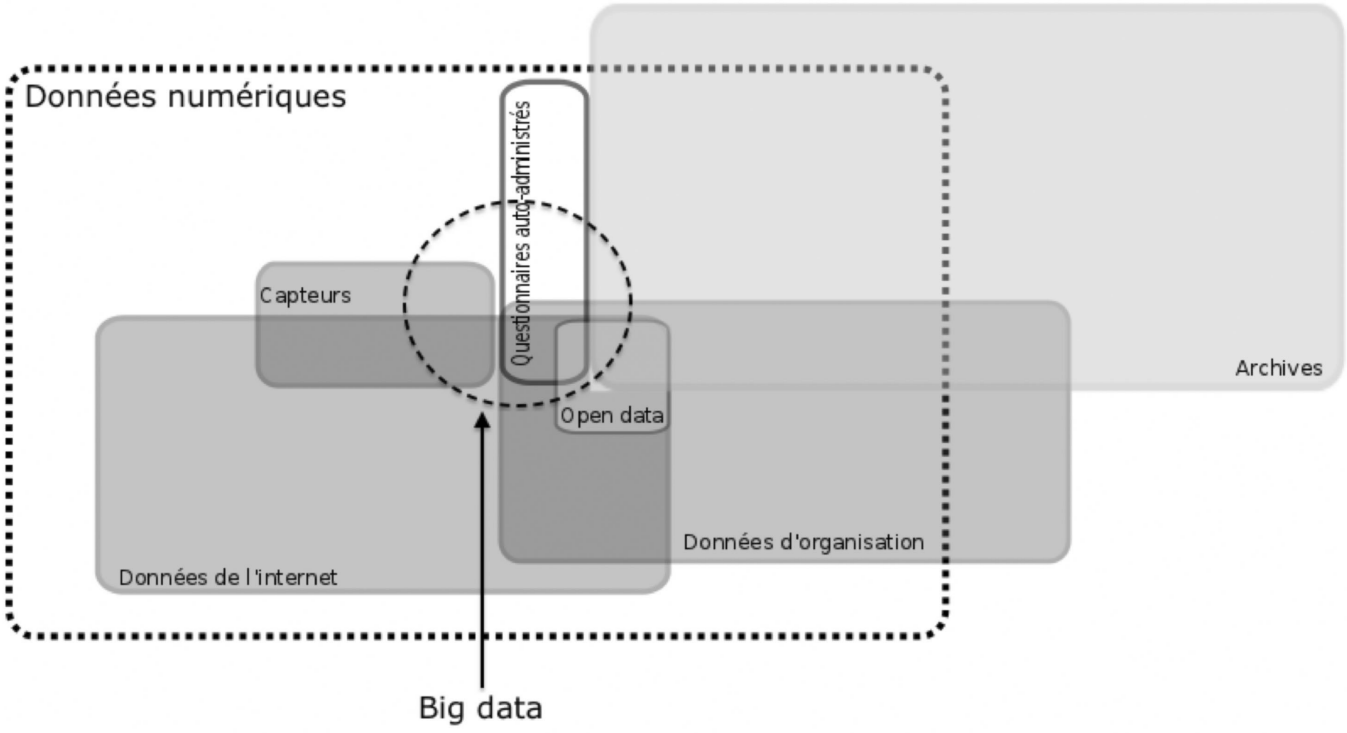
qualificatif de pratiques numériques est parfois utilisé comme synonyme de Web. Si on parle de données numériques, c'est plutôt au sens où l'entendent les informaticiens. Ces données sont dites numériques parce que l'objet qu'elles représentent est encodé sous forme de chiffres – d'où le terme numérique – qui sont interprétés par un processeur. Ces données peuvent prendre la forme de tableaux, de fichiers, de pages de traitements de texte, d'images, de sons, etc., mais elles partagent toutes une propriété : cet encodage sous forme de chiffres, qui permet à une machine de lire et d'exécuter des actions séquentielles sur ces informations. Comme on l'a dit plus haut, ces données peuvent être « massives » ou pas, « nativement » numériques ou pas, libres d'accès ou propriétaires. Les sources sont là encore nombreuses : capteurs, ordinateurs, et même chercheurs qui scannent et recodent manuellement.

Si elles ne sont pas toutes si massives, toutes les données évoquées ci-dessus sont bien des données numériques. Le schéma ci-dessous illustre les différentes sources et leurs relations.

Des données et leurs biais

Repensé à l'aune de ces distinctions, le débat relatif à l'opportunité et aux difficultés des *big data* s'éclaire. Aborder le sujet en termes de données numériques, qui sont elles-mêmes diverses, permet en effet de préciser les enjeux. Les vastes ensembles de données, ce qu'on désigne par *big data* sur la Figure 1, nécessitent pour être traités des capacités de mémoire vive, de stockage et de gestion de flux particulières. Il faut souvent des ordinateurs puissants, voir des *clusters* d'ordinateurs mis en réseau, pour arriver à faire un calcul simple. Une analyse géométrique de données ou une régression peuvent prendre des heures sur un ordinateur personnel. Alors que depuis des décennies la puissance de calcul des ordinateurs augmentait significativement plus vite que le volume de données à traiter, le phénomène inverse se produit depuis quelques années, qui oblige à repenser les stratégies pour gérer la masse et la diversité de ces données. Ces problématiques de *high-performance computing* ne se posent toutefois que rarement avec d'autres

Figure 1 : Au delà des big data, les principales sources numériques en sciences sociales*



* Ce shéma se veut illustratif et ne fournit pas de représentation à l'échelle de la situation actuelle. Tout au plus a-t-on tenté de représenter l'importance différentielle que prennent certaines sources à l'heure actuelle dans la recherche en sciences sociales.

sources. La plupart des questionnaires en ligne ne dépassent pas le millier d'individus, et beaucoup de données issues de l'internet peuvent être traitées sur un ordinateur classique. Le plus souvent, c'est aussi vrai des archives et de *l'open data*.

Les données de l'internet sont, en revanche, lourdes d'enjeux relatifs à l'échantillon récolté. La controverse qui a suivi la publication du *Great British Class Survey* évoqué ci-dessus le rappelle. La question du biais de sélection des 160 000 répondants à l'enquête et des éventuelles possibilités de contrôle a donné lieu à des débats qui préfigurent bien des futures polémiques à ce sujet[11]. Peut-on dire quelque chose d'un tel échantillon en l'absence de connaissances sur la population ? Cette question ne se pose toutefois pas quand l'enquêteur dispose d'une base exhaustive, que les méthodes numériques lui permettent de collecter immédiatement – comme dans le cas des analyses de citations évoquées ci-dessus. Beaucoup de données issues de l'*open data* posent des questions complexes quant à leur production. Dans le cas des administrations, la mise à disposition de ces informations au public passe souvent par la mise en commun des fichiers produits dans des services différents, par des personnes aux formations, aux routines et aux enjeux parfois divergents. Les opérations de codage, si importantes au travail de recherche, sont au mieux invisibles pour les utilisateurs finaux. Par ailleurs, elles sont le plus souvent réalisées sans aucune harmonisation ni contrôle.

Ces biais ne sont que quelques exemples des difficultés spécifiques qui se posent aux chercheurs qui ont recours à ces données. Ils ne doivent pas être occultés, mais ils n'empêchent pas nécessairement l'usage de ces informations non plus. Reposer le débat en termes de données numériques présente, en effet, un autre avantage : rappeler que ce dont il est question, ce sont avant tout des données, c'est-à-dire des informations partielles destinées à éclairer un phénomène particulier – mais qui ne sont jamais pures, jamais neutres, et qui ne peuvent jamais être utilisées sans une opération réflexive destinée à en connaitre les conditions de production. Le rappel de cette évidence a toutefois une vertu : il transforme la controverse souvent totalisante sur les *big data* en un débat sur des matériaux avec lesquels nos disciplines ont l'habitude pluriséculaire de travailler. Les reproches souvent faits aux *big data* ou aux données numériques apparaissent à la fois comme familiers, et parfois surmontables[12].

Enjeux contemporains

Moins qu'une « révolution des *big data* » souvent annoncée, le véritable déplacement pour les sciences sociales se situe probablement du côté de la multiplication de données numériques. Dans de nombreux aspects du travail de recherche, ces données ne constituent pas une nouveauté radicale, certaines sont même présentes depuis des décennies. Ce qui change en revanche, c'est l'augmentation exponentielle de leur nombre, permise par la diffusion de capteurs, l'usage d'internet, et la numérisation croissante du quotidien. Cette abondance nouvelle n'est pas sans conséquences pour la recherche, mais les enjeux se situent ailleurs que là où les débats polarisés sur les *big data* les situent. Dans les pages qui suivent, on évoque certains domaines qui sont dès à présent affectés par cette abondance nouvelle. La liste est, bien entendu, non exhaustive.

Réorganisations disciplinaires

Le premier changement a trait aux multiples réorganisations que produit la multiplication de données numériques pour nos disciplines. L'un des effets les plus marqués est la valorisation de nouvelles qualifications, et en premier lieu la capacité à traiter ces données numériques. Les publications comme les fiches de postes montrent un véritable intérêt pour ceux qui peuvent mobiliser ces sources. Des informaticiens sont embauchés dans des laboratoires de sciences sociales, les formations à destination des chercheurs en sciences sociales se multiplient. La clarification opérée ci-dessus est utile : dans le cas de données vraiment massives et hétérogènes, disposer d'informaticiens est certainement nécessaire afin de mettre en place les procédures qui permettent de traiter l'information de manière efficace. Ces partenariats entre science de l'informatique et sciences sociales existent déjà, et ils s'avèrent souvent

11. Voir http://soc.sagepub.com/site/British_Social_Class/British_Social_Class_Homepage.xhtml (consulté le 1er mai 2015).

12. Il y a plus d'un siècle, Charles-Victor Langlois et Charles Seignobos (1898) invitaient les historiens à pratiquer la critique des sources, c'est-à-dire à replacer les informations dont ils se saisissent dans leur contexte de production afin d'estimer leur biais. On retrouve les mêmes remarques sur la nécessaire objectivation plusieurs décennies plus tard dans *Le métier de sociologue* (Bourdieu *et al.*, 1968).

fructueux. Ce déplacement a aussi pour conséquence d'accroître la division du travail scientifique entre ceux qui extraient les données et ceux qui les traitent.

Les sciences sociales peuvent toutefois avoir bien d'autres usages des données numériques, et ne pas avoir besoin de telles compétences. Savoir lire le Web, savoir écrire du code, comprendre le fonctionnement d'un ordinateur : ces compétences ne servent pas que ceux qui travaillent sur internet et étudient les interactions en ligne. Un économiste qui travaille sur des données fiscales peut avoir à nettoyer les bases de données fournies par l'administration avant de pouvoir les traiter statistiquement – supprimer les scories, repérer les doublons, les erreurs de codage, harmoniser. Un sociologue urbain peut vouloir collecter la liste des commerces dans un quartier pour une étude sur la densité marchande, puis les géocoder pour faire de la cartographie. Un historien qui aurait patiemment compilé une base de données afin de réaliser des traitements prosopographiques peut avoir besoin de changer son format afin de s'en servir pour des traitements sous d'autres logiciels. Dans chacun de ces cas, il suffit souvent de quelques lignes de code pour réaliser des tâches autrement tellement chronophages qu'elles en sont décourageantes. La barrière technique, qui peut paraître importante au premier abord, est en fait relativement limitée. En quelques heures de cours, on peut avoir une idée précise de la manière dont s'écrit une bonne partie du web ; en quelques heures de plus, on peut se former au fonctionnement des bases de données, et avec encore quelques heures, on peut écrire des scripts capables de collecter et de reformater ces informations.

De fait, il est urgent de réaliser que nous travaillons tous, et de plus en plus, avec des ordinateurs, et le parti que nous pouvons en tirer, quel que soit le type de recherche que nous menons. Rien n'oblige en effet à traiter ces informations avec des statistiques avancées. Une fois extraites – d'un site web, d'une base de données – et sélectionnées, elles peuvent être consultées comme autant d'archives. Elles peuvent aussi donner lieu à des explorations visuelles, sans leur imposer de traitement statistique avancé (Healy & Moody, 2014). En d'autres termes, s'il est probable que l'intérêt pour ces données numériques participe à la revalorisation des méthodes quantitatives, l'effet le plus net est probablement l'émergence d'une nouvelle compétence méthodologique à côté du « quali » et du « quanti » (ordi ?)[13].

De nouvelles approches quantitatives

La disponibilité croissante de données numériques est allée de pair avec le développement de nouvelles approches quantitatives. Profitant de la multiplication des informations disponibles et des capacités de calcul informatique, les méthodes d'apprentissage automatique – *machine learning* – sont venues s'ajouter aux méthodes désormais plus classiques – statistiques descriptives, inférentielles[14] et analyse de données. Dans le cas où les volumes de données sont importants, elles permettent des améliorations sensibles. D'abord, par différence avec les statistiques inférentielles classiques, elles permettent, pour la plupart, de tirer parti d'échantillons de très grande taille. Les tests statistiques d'hypothèses présentent, en effet, cet inconvénient que lorsque l'échantillon est très grand, les valeurs-p sont systématiquement très faibles[15]. Ces techniques permettent, par ailleurs, de s'affranchir de bon nombre d'hypothèses relatives à la structure des données que l'on retrouve dans nombre de méthodes classiques, créées à une époque où les échantillons étaient – relativement – de petite taille.

Né dans la décennie 1960 à l'interface des mathématiques et de l'informatique, le domaine de l'apprentissage statistique a connu de profondes transformations depuis les années 1990, au moment où les algorithmes d'apprentissage sont devenus des objets d'étude pour les statisticiens. Les champs d'application sont très nombreux, allant de la reconnaissance vocale à la programmation de véhicules sans conducteur, en passant par la recherche en génomique ou l'entraînement de robots champions d'échecs. Le champ s'articule toutefois autour de quelques grandes tâches : l'optimisation numérique, qui consiste à chercher les points qui minimisent une fonction donnée ; l'approximation de fonctions, pour approcher la

13. Aux États-Unis, c'est d'ailleurs le choix fait pour désigner ces approches : on parle de « *computational social sciences* ».

14. La statistique inférentielle, que l'on distingue traditionnellement de la statistique descriptive, consiste à inférer d'un échantillon aléatoire des propriétés de la population étudiée, essentiellement sur le mode de tests d'hypothèses, comme celui du Khi^2.

15. Par exemple, pour un million d'observations, l'hypothèse d'indépendance de deux variables numériques sera rejetée au seuil de 5 % dès que leur coefficient de corrélation linéaire dépasse, en valeur absolue, la valeur de 0,002, ce qui est absolument sans intérêt pour l'interprétation (Saporta, 2006).

forme d'une fonction dont on n'observe que des réalisations aléatoires ; la visualisation de données multidimensionnelles ; le *problem solving*, pour choisir la meilleure solution parmi un ensemble discret de possibilités ; et quelques autres. Parmi ces nombreuses tâches, toutes ne sont pas également utiles à la recherche en sciences sociales, mais au moins trois méritent l'attention. D'abord l'apprentissage supervisé, appellation qui regroupe les tâches de régression et de classification, c'est-à-dire la prédiction d'une variable expliquée, quantitative ou qualitative, à partir d'un ensemble de variables explicatives. Ensuite la réduction de dimensionnalité, qui consiste à construire une représentation synthétique d'un nuage de points multidimensionnel, souvent sous la forme d'une visualisation graphique en deux ou trois dimensions. Enfin le *clustering*, qui consiste à former des groupes d'observations similaires. Ces trois aspects ont des applications désormais bien connues en statistique d'apprentissage. Elles se distinguent des méthodes standards sur deux points particuliers.

La première grande différence a trait au caractère restrictif des postulats de départ des méthodes classiques. Développées à une époque où les échantillons dépassaient rarement les quelques milliers d'observations, ces dernières requièrent de fortes hypothèses sur la forme des données et du modèle choisi. Ainsi, estimer un modèle de régression linéaire demande de spécifier *a priori* la forme exacte de la relation entre variable expliquée et variables explicatives. Dans une analyse en composantes principales, les seules représentations possibles sont des transformations linéaires des variables de départ. Cette hypothèse de linéarité fait qu'une variable donnée ne peut pas successivement diminuer puis augmenter le long d'une composante principale. Sous ces hypothèses, les techniques standard permettent alors de trouver, dans une classe de modèles restreinte, celui qui s'ajuste le mieux aux données. *A contrario*, les approches issues du *machine learning* peuvent être qualifiées de flexibles en ce qu'elles s'appuient sur des hypothèses moins nombreuses ou moins fortes. Elles sont alors capables d'approximer une large classe de modèles, et de choisir une bonne spécification en même temps qu'elles l'ajustent aux données[16].

Une seconde différence tient au fait que les méthodes d'apprentissage statistique sont particulièrement efficaces pour prédire et/ou décrire, mais qu'elles sont généralement moins performantes pour proposer une explication ou une interprétation. Les raisons sont à chercher dans les fondements théoriques des deux approches qui, pour accomplir des tâches similaires, adoptent des démarches souvent divergentes. Plutôt qu'une résolution exacte fondée sur un modèle formel connu, les méthodes d'apprentissage s'appuient sur une heuristique, *i.e.* une stratégie de recherche de solution dans le cadre d'un modèle formellement moins bien posé. Les modèles de forêts aléatoires, apparus au début de la première décennie 2000 et qui ont depuis rencontré un large succès dans de nombreux domaines d'application – reconnaissance d'images, génomique, recherche pharmaceutique... –, illustrent bien ce point. Ces modèles d'apprentissage supervisé sont fondés sur un principe d'agrégation de nombreux modèles à faible pouvoir prédictif, eux-mêmes construits à partir de perturbations aléatoires de l'échantillon de départ. Ils excellent dans la prédiction et la généralisation – la capacité d'un modèle à prédire correctement des observations qui n'étaient pas disponibles au moment de l'optimisation des paramètres – mais, si leur efficacité empirique est largement démontrée, leur grande complexité mathématique font que cette efficacité échappe encore, quinze ans après leur création, à toute explication formelle rigoureuse. L'analyse des résultats est à l'avenant. Là où les coefficients d'une régression linéaire sont immédiatement interprétables et répondent à des questions précises – la variable *x* contribue-t-elle significativement aux variations de la variable *y* ? Quelle augmentation moyenne une incrémentation de *x* entraîne-t-elle sur *y*, toutes choses égales par ailleurs ? –, ce n'est pas le cas des paramètres d'une forêt aléatoire. Non pas que les modèles d'apprentissage interdisent toute interprétation, mais celle-ci est fondée sur un critère autre – la contribution à la prédiction. Cette difficile interprétation est intimement liée à la flexibilité de ces modèles, qui a pour pendant une grande complexité formelle.

Les graphiques de la Figure 2 illustrent bien cette différence dans les approches, dans le cadre d'une opération de réduction de dimensionnalité. Les deux graphiques sont des visualisations en deux dimensions d'un nuage de points multidimensionnel représentant les invitations des 577 députés de l'Assemblée nationale dans différents médias au cours des trois premières

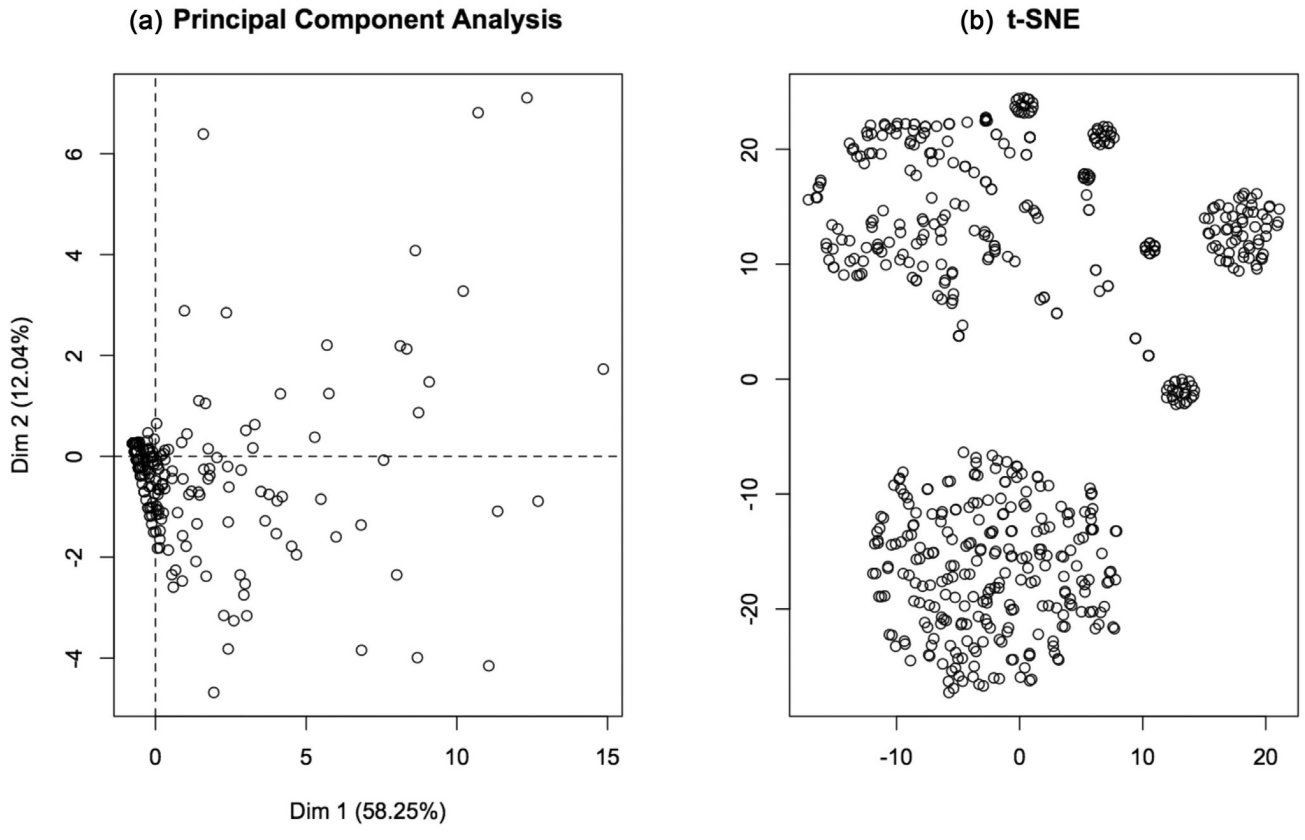
16. Le pendant de cette flexibilité est que les méthodes de *machine learning* fonctionnent généralement très mal sur des échantillons trop petits – la taille d'échantillon minimale dépendant de la complexité du modèle sous-jacent aux données.

années de la 14e législature[17]. Pour ce faire, on a recours à une analyse en composantes principales (ACP), et à une méthode nommée *t-SNE* (*t-distributed stochastic neighborhood embedding*, Van der Maaten & Hinton, 2008). Là où l'ACP permet une interprétation puissante mais reste relativement limitée pour ce qui est de produire une description fine, la situation est inverse avec t-SNE.

L'un des principaux atouts d'une représentation par ACP réside dans l'interprétabilité immédiate des axes, qui sont de simples combinaisons linéaires des variables de départ. La lecture des résultats de l'ACP nous apprend que le premier axe de la Figure 2a est fortement corrélé au nombre total d'apparitions dans les médias, et que le second distingue principalement entre médias généralistes et médias spécialisés. L'inertie du premier axe nous apprend aussi que celui-ci est crucial à la structuration des données, puisqu'il explique à lui seul près de 60 % de la variance totale du nuage. Mais l'intérêt descriptif de cette ACP est relativement limité passée cette première remarque. La majorité des points reste en effet concentrée autour de l'origine, et la représentation illustre essentiellement la différence entre cette masse d'observations (députés rarement ou pas invités) et quelques valeurs extrêmes (députés très médiatiques).

Par différence, la seconde représentation ne permet pas d'interprétation géométrique des axes, même s'il est possible de calculer leur corrélation aux variables de départ. Elle offre, en revanche, une description beaucoup plus riche des observations : les différents groupes distingués sur la Figure 2b représentent des députés aux profils médiatiques similaires. Le groupe en haut à gauche concentre les parlementaires

Figure 2 : Deux méthodes de réduction de dimensionnalité ; (a) Analyse en composantes principales ; (b) t-SNE



17. Les données ont été collectées dans le cadre d'une enquête menée sur le travail parlementaire. Elles ont été regroupées en une dizaine de variables – chaines parlementaires, radios spécialisés, télévisions spécialisées, etc. Elles sont, avec les scripts R qui permettent de reproduire et d'analyser les graphiques, disponibles dans l'annexe en ligne, publiée sur le site de la revue (https://sociologie.revues.org/2615). Pour plus d'informations sur la communication des députés, voir (Delpierre, 2015).

régulièrement invités sur différents médias, le groupe du bas ceux qui ne sont jamais invités nulle part[18]. Les petites grappes que l'on voit en haut à droite du graphique représentent ceux qui sont rarement invités, mais qui l'ont été au moins une fois dans un média ou un autre. On parvient donc à une description fine des profils d'individus, mais l'interprétation générale est moins évidente – en tout cas en première analyse. Dans le cas d'énormes bases de données, elle devient vite très difficile. En d'autres termes, ces techniques savent très bien regrouper des individus, mais elles ne permettent pas si facilement de savoir pourquoi ils sont si près les uns des autres.

La recherche en apprentissage automatique est encore très active à l'heure actuelle, à la croisée entre monde académique et entreprises qui voient dans ces méthodes des manières de mieux connaître leurs objets comme leurs clients. Ces techniques, qui commencent à se diffuser en sciences sociales, ne révolutionneront probablement pas les approches quantitatives de ces disciplines. La difficulté à interpréter certains résultats les éloigne des attentes des chercheurs dans nos domaines et la barrière mathématique reste très présente. Elles permettent toutefois d'aborder autrement les données. En particulier, leurs puissantes capacités descriptives peuvent être mises à profit. Que ce soit à des fins d'analyse exploratoire ou en tant que tel, les résultats qu'elles offrent pourraient favoriser le développement d'une description robuste, car armée sur de nombreuses observations, entre statistique inférentielle et traitements qualitatifs plus classiques.

Accès et conservation des données

Outre la collecte et le traitement, la question de l'accès aux sources et de leur gestion est un autre enjeu important que soulève la multiplication des données numériques. Si on trouve bien de plus en plus d'informations, elles ne sont pour autant pas toutes disponibles, que ce soit pour des raisons légales ou déontologiques. Les chercheurs qui travaillent sur des données numériques rencontrent deux types de problématiques : celle du droit de la propriété intellectuelle et celle de la protection de la vie privée.

Le droit des bases de données est restrictif et clairement encadré par la loi. Revenons sur le cas, évoqué ci-dessus, des chercheuses qui travaillent sur la conjugalité au XXIe siècle à partir des sites de rencontre. Une possibilité est de collecter automatiquement les informations visibles sur le site. Profils des individus, présentation de soi, type de personne recherchées : toutes ces informations sont disponibles en ligne. À partir d'un programme de quelques lignes, il serait possible en une journée de récupérer des dizaines de milliers de profils, sur lesquels on pourrait mener des analyses. C'est toutefois illégal, et la peine encourue est rédhibitoire[19], même si elle n'est finalement que très rarement appliquée. La solution pour accéder à ces données – et à d'autres, qui ne sont pas mises en ligne mais peuvent être utiles – est alors de négocier l'accès à cette base. Cela passe souvent par la rédaction d'un contrat, l'engagement d'une restitution, et parfois la communication au nom de la société. En d'autres termes, il faut mettre en place une procédure longue, chronophage. Si les données sont de bonne qualité, cette pratique peut être intéressante. À l'inverse, s'il ne s'agit que de collecter une seule information, ou de ne faire qu'un calcul afin de tester une hypothèse, la procédure est bien trop longue pour être mise en place de manière répétée dans le cadre d'une recherche.

L'autre type de droit avec lequel doit composer le chercheur a trait à la vie privée et aux libertés. Dans de nombreux pays, ces droits de la personne sont encadrés par des agences indépendantes. Les jurisprudences varient, mais la règle de base est qu'un fichier ne peut contenir des informations qui permettraient d'identifier une personne. D'autres éléments entrent aussi en ligne de compte : le fait que l'information collectée soit pertinente pour l'étude, le caractère proportionnel aux finalités de l'enquête, la protection des données amassées, l'information des personnes concernées, ou la durée de conservation. Ces dispositions imposent des contraintes fortes sur la recherche. Il est ainsi interdit de collecter l'adresse de médecins généralistes, pourtant disponible sur plusieurs sites, en vue de dresser une simple cartographie des déserts médicaux. Une telle pratique contreviendrait en effet à la loi Informatique et Libertés.

18. Nous avons préalablement ajouté de très petites perturbations aléatoires au nuage de point multidimensionnel, de façon à ne pas écraser les observations à coordonnées identiques sur un seul point. Cette opération n'affecte pas la représentation par ACP, mais améliore grandement celle par t-SNE.

19. En France, l'article 112-3 du code de la propriété intellectuelle prévoit de punir ces actes de 500 000 euros d'amende et de trois ans d'emprisonnement.

Les autres règles de protection de la vie privée, pour salutaires qu'elles sont, sont elles aussi tellement contraignantes qu'elles en paraissent déplacées. Souvent évoqué, le principe de pertinence de la collecte, qui veut que les chercheurs ne récoltent que les informations qui leurs seront utiles, peut paraître de bon sens. Mais il occulte le fait que la recherche procède au moins autant par une logique de la découverte que par une logique de la preuve. Autrement dit, l'activité scientifique ne se limite pas à tester des hypothèses bien formulées, mais elle progresse aussi par essais, échecs et reformulations. Ne collecter que des données définies comme pertinentes *ex ante*, imposer un protocole strict en amont de l'enquête, cela revient à n'utiliser les données que pour valider des hypothèses déjà présentes. Dans bien des cas, c'est simplement s'interdire de découvrir.

Dans le même temps, l'avalanche d'informations impose une réflexion sur la protection des enquêtés. Si cette question n'est pas nouvelle, l'enjeu se pose avec une ampleur plus grande qu'avant. Agrégées, des données publiées en ligne offrent des informations précises sur les pratiques d'une personne. On peut aussi collecter rapidement des centaines d'informations sur les pratiques de milliers d'individus, parfois avec une granularité bien plus fine qu'une enquête de terrain. L'anonymisation n'est alors pas une protection suffisante : les chercheurs en informatique ont, depuis longtemps, montré qu'on pouvait retrouver l'identité des personnes en croisant des données anonymisées avec des informations publiquement disponibles. C'est ce que démontrait Latanya Sweeney voilà plus de dix ans : à partir d'une base de données médicale, anonymisée, et des listes électorales du Massachussetts, elle était parvenue à retrouver l'identité de la majorité des personnes indiquées dans la première base, dont le gouverneur de l'État (Sweeney, 2002)[20]. La réidentification n'est pas le seul problème. Les données, anonymisées ou non, sont la plupart du temps stockées sur des disques durs qui peuvent être copiés, perdus ou volés. Disponibles dans un format numérique, les données collectées, qu'elles soient « big » ou non, peuvent facilement circuler. Il importe donc que les chercheurs continuent la réflexion sur la protection des enquêtés à l'aune de ces changements. Droit d'enquêter d'un côté, protection des enquêtés de l'autre, voilà les deux aspects qui se posent aux chercheurs de manière pressante, sans forcément s'opposer (Laurens & Neyrat, 2010).

Quels objets, quelles théories?

Enfin, sans succomber aux discours prophétiques qui annoncent régulièrement une révolution, il faut reconnaître que l'abondance de données tend à modifier en profondeur certains aspects de la recherche. L'histoire des sciences sociales confirme ce point. L'irruption de l'enquête par questionnaire dans la sociologie étatsunienne dans les années 1950 a produit des effets importants sur la recherche dans ce domaine. L'opposition bien connue entre Paul Lazarsfeld et Talcott Parsons, entre l'empirisme fort du premier et les volontés systématisantes du second, était largement sous-tendue par le recours aux nouvelles données, issues de la généralisation des questionnaires, et les capacités de traitement *via* les premiers ordinateurs électroniques. Elle s'est conclue par la victoire du premier (Calhoun & van Antwerpen, 2007), à tel point que la référence à T. Parsons, incontournable au milieu des années 1950, était devenue anecdotique une décennie plus tard. La recherche en économie en fournit un autre exemple. Aux ambitions théoriques générales de l'économie politique, puis au mouvement de formalisation mathématique, a clairement succédé une phase où l'investigation empirique devenait centrale (Fourcade, 2009). Ce basculement correspond assez nettement à la disponibilité croissante de données dans cette discipline. Pour certains, les changements se font déjà sentir. Dans son dernier ouvrage, Andrew Abbott soutient que l'actuelle multiplication de données donne lieu à une transformation du savoir. Partant de son expérience d'enseignant et d'encadrant, il soutient que les étudiants ont développé des compétences certaines pour aller chercher de l'information, mais que l'articulation entre questions de recherches et enquête est progressivement remplacée par un hyper-empirisme qui oublie trop souvent la construction d'un objet de recherche (Abbott, 2014, pp. 103-104).

Mais le niveau de généralisation n'est pas le seul aspect. La disponibilité de données numériques dans de nombreux domaines est aussi source de changements potentiels dans les objets mêmes qu'étudient les sciences sociales. Au-delà de la nouveauté, ces données sont utilisées simplement parce qu'elles existent et sont parfois facilement accessibles. Cette disponibilité a un revers : les domaines moins riches en

20. On pourra se référer à la présentation que fait Benjamin Nguyen des différentes techniques d'anonymisation disponibles (2014).

données numériques pourraient être délaissés au profit de ceux où l'enregistrement est permanent. Un aspect plus préjudiciable est que cette disponibilité de données pourrait inciter les chercheurs à se tourner vers les sources existantes plutôt que de prendre le temps d'en chercher, voire d'en produire de nouvelles. Les sciences sociales seraient alors tributaires des données produites par d'autres, et pourraient perdre leur inventivité, tant méthodologique que théorique et empirique.

Conclusion

Dans les années 1950, l'arrivée des premiers ordinateurs avait donné lieu à des discours enthousiastes. L'abondance de données et la capacité de traitement des machines devaient modifier la recherche en profondeur. Les sommes et les énergies investies devaient permettre un progrès technologique et social inouï jusqu'alors. Les sciences comme les sociétés allaient connaitre une révolution (Galison & Hevly, 1992). Comme c'est souvent le cas, le détour par l'histoire permet des mises en perspective utiles. Dans le cas présent, il rappelle que chaque changement technologique significatif donne lieu à des fantasmes comme à des mutations effectives. Il en va de même pour les *big data*. Les discours à leur propos restent trop généraux pour qu'on puisse saisir les véritables changements qui sont à l'œuvre. Afin de clarifier ces enjeux, l'article distingue les différentes sources regroupées sous le terme de *big data* avec lesquelles les chercheurs en sciences sociales sont amenés à travailler. Cette entrée par les sources n'est pas la seule possible, mais elle a un double mérite. Elle permet d'abord de rappeler que parmi les données numériques désormais disponibles, toutes ne sont pas volumineuses : la multiplication du nombre de bases de données ne doit pas être confondue avec leur taille, qui peut rester modeste. Elle rappelle aussi que nous travaillons toujours avec des sources qui doivent être évaluées chacune en fonction de leur intérêt propre. Pour beaucoup, leurs avantages comme leurs biais sont bien connus. Le véritable changement est finalement ailleurs. Les *big data* ne sont que la face visible d'un phénomène bien plus large, la multiplication exponentielle de données numériques. Cette transformation est lourde d'enjeux, qui travaillent déjà nos disciplines.

Bibliographie

Abbott A. (2014), *Digital Paper. A Manual for Research and Writing with Library and Internet Research*, Chicago, University of Chicago Press.

Abbott A. & Sparrow J. (2007), « Hot War, Cold War: The Structures of Sociological Action, 1940-1955 », *in* Calhoun C. (dir.), *Sociology in America: a History,* Chicago, University of Chicago Press, pp. 281-312.

Aiden E. & Michel J.-B. (2013), *Uncharted: Big Data as a Lens on Human Culture*, New York, Riverhead Books, Penguin Books.

Bastard I., Cardon D., Fouetillou G., Prieur C. & Raux S. (2013), « Travail et travailleurs de la donnée », *InternetActu.net*, 13 décembre, http://www.internetactu.net/2013/12/13/travail-et-travailleurs-de-la-donnee/ [visité le 1er mai 2015].

Bergström M. (2014), « Au bonheur des rencontres. Classe, sexualité et rapports de genre dans la production et l'usage des sites de rencontres en France », Thèse de doctorat en sociologie, Sciences Po Paris.

Bourdieu P., Chamboredon J.-C. & Passeron J.-C. (1968), *Le métier de sociologue*, Paris, Mouton.

boyd d. & Crawford K. (2012), « Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon », *Information, Communication & Society*, vol. 15, no 5, pp. 662-679.

Calhoun C. & van Antwerpen J. (2007), « Orthodoxy, Heterodoxy and Hierarchy: "Mainstream sociology and its critics" », *in* Calhoun C. (dir.), *Sociology in America: A History,* Chicago, University of Chicago Press.

Cavallo A. (2013), « Online and Official Price Indexes: Measuring Argentina's Inflation », *Journal of Monetary Economics*, vol. 60, no 2, pp. 152-165.

Delpierre A. (2015), « "Une fois l'émission faite, j'ai senti qu'on me regardait autrement ici, à l'Assemblée". Rapports à la médiatisation et pratiques de communication des élus », Mémoire de Master 2, ENS-EHESS.

Didier E. (2009), *En quoi consiste l'Amérique ? Les statistiques, le New deal et la démocratie*, Paris, La Découverte.

Dillman D., Smyth J., Christian L. (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, New-York, Wiley.

Eagle N. (2010), « Mobile Phones as Social Sensors », *The Handbook of Emergent Technologies in Social Research*, New-York, Oxford University Press.

Einav L. & Levin J. (2014), « The Data Revolution and Economic Analysis », *Innovation Policy and the Economy*, vol. 14, no 1, pp. 1-24.

Fourcade M. (2009), *Economists and Society. Discipline and Profession in the United States, Britain and France, 1890s-1990s*, Princeton, NJ, Princeton University Press.

Galison P. & Hevly B. (1992), *Big Science: The Growth of Large-scale Research*, Redwood City, CA, Stanford University Press.

Gingras Y. (2010), « Revisiting the "Quiet Debut" of the Double Helix: A Bibliometric and Methodological Note on the "Impact" of Scientific Publications », *Journal for the History of Biology*, vol. 43, no 1, pp. 159-181.

Harcourt B. (2014), « Governing, Exchanging, Securing: Big Data and the Production of Digital Knowledge », *Columbia Public Law Research Paper n°14-390*, 2014. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2443515 [consulté le 1er mai 2015].

Healy K. & Moody J. (2014), « Data Visualization in Sociology », *Annual Review of Sociology*, vol. 40, pp. 105-128.

Knapen H., Braes S., Ermans T. & Erremans W. (dir.) (2014), *Het Datawarehouse, een duizendpoot! Perspectieven van het Datawarehouse Arbeidsmarkt en Sociale Bescherming*, Gand, Academia Press.

Langlois Ch. V. & Seignobos Ch. (1898), *Introduction aux études historiques*, Paris, Hachette.

Laurens S. & Neyrat F. (2010), *Enquêter : de quel droit ? Menaces sur l'enquête en sciences sociales*, Bellecombe-en-Bauges, Éditions du Croquant.

Mayer-Schonberger V. & Cukier K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston, Eamon Dolan/Mariner Books.

Miller D. & Slater D. (2001), *The Internet: An Ethnographic Approach*, Londres, Bloomsbury.

Michel J.-B., Shen Y.K., Presser Aiden A., Veres A., Gray M., The Google Books Team, Pickett J., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M.A. & Aiden E. (2011), « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, vol. 331, pp. 176-182.

Nguyen B. (2014), « Techniques d'anonymisation », *Statistiques et sociétés*, vol. 2, no 4, pp. 43-50.

Pentland A. (2014), *Social Physics. How Good Ideas Spread – The Lessons From A New Science*, New York, Penguin.

Potârcă G. & Mills M. (2005), « Racial Preferences in Online Dating across European Countries », *European Sociological Review*, first published online January 5, 16 p., doi:10.1093/esr/jcu093.

Salganik M., Dodds P. & Watts D. (2006), « Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market », *Science*, vol. 311, pp. 854-856.

Saporta G. (2006), *Probabilités, analyses de données et statistiques*, Paris, Technip, 2e édition, p. xxxii.

Solla Price D. (de) (1986), *Little Science, Big Science... and Beyond*, New York, Columbia University Press.

Sorokin P. (1956), *Fads and Foibles in Modern Sociology and Related Sciences*, Chicago, Henry Regnery.

Sweeney L. (2002), « k-anonymity: A Model for Protecting Privacy », *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no 5, pp. 557-570.

Van der Maaten L. & Hinton. G. (2008), « Visualizing High-Dimensional Data Using t-SNE », *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605.

Williams N. *et al.* (2013), « Measurement of Human Mobility Using Cell Phone Data: Developing Big Data for Demographic Science », *Working paper No.137*, Center for Statistics and the Social Sciences, University of Washington.