Canadian Journal of Regional Science
Revue canadienne des sciences régionales

Données massives et sciences du territoire

Mathieu Charron[1], Richard Shearmur[2], et Geneviève Beauchemin[3]

[1]Département des Sciences Sociales, Université du Québec en Outaouais, Gatineau QC; [2]School of Urban Planning, McGill University, Montréal QC; [3]Assistante de recherche, étudiante à la maîtrise, Département des Sciences Sociales, Université du Québec en Outaouais, Gatineau QC. Adressez vos commentaires à mathieu.charron@uqo.ca

Soumis le 8 août 2017. Accepté le 25 juin 2018.

Charron, M, Shearmur, R, et Beauchemin, G. 2018. Données massives et sciences du territoire. *Canadian Journal of Regional Science / Revue canadienne des sciences régionales 41(1/3)*, 15-27.

Les Big Data, traduit ici par données massives, attirent beaucoup d'attention depuis quelques années. Elles suscitent en effet beaucoup d'espoir de toutes parts, notamment auprès des technologues, des hommes d'affaires et des politiciens. En sciences sociales, toutefois, elles semblent susciter plus de critiques et d'inquiétudes que d'attentes. En fait, il nous semble, comme d'autres, que les sciences sociales boudent un phénomène qui pourtant façonne de plus en plus leurs objets d'études. Les données massives ouvrent cependant de nouvelles voies pour la compréhension des phénomènes sociaux.

L'objectif de cet article est de résumer les principaux enjeux soulevés dans la littérature sur les données massives. Il s'adresse d'abord aux chercheurs en sciences sociales et plus particulièrement aux chercheurs des sciences régionales et territoriales. En effet, nous faisons un effort particulier pour présenter les impacts des données massives sur la recherche et les objets d'étude de ces disciplines. Nous présentons aussi quelques possibilités concrètes de données nouvelles qui, à notre avis, peuvent éclairer notre compréhension des territoires. Ce faisant, nous espérons commencer à éclairer ce qui nous semble être un angle mort de la littérature en sciences sociales de langue française.

Lire sur les données massives, c'est un peu comme regarder un ciel étoilé : on se sent si petit! En effet, cette littérature est un monde de superlatifs où le giga fait figure de micro. On évoque des quantités littéralement plus grandes que nature comme les pétabits (10^{15} bits). Ces valeurs incommensurables ne manquent pas de nous étourdir. Elles ne sont passées du rêve à la réalité que récemment, grâce aux nouvelles possibilités offertes par les technologies informatiques. Suivant les « lois de Moore », les supports technologiques (puissance et capacité de stockage) sont en éruption continue depuis les années 1960. Dans ce contexte, les métaphores cataclysmiques sont de mises pour décrire la déferlante de données. On a d'abord évoqué l'avalanche (Miller 2010) mais le déluge est peut-être plus approprié (Hu et al 2014). On peut facilement se noyer dans cette abondance où les arches et les îles se font rares. Ces métaphores illustrent autant la fascination que l'angoisse suscitées par ce qu'on appelle communément le *big data*.

Cette appellation, traduite ici par « données massives », a mis un certain temps à s'imposer. C'est que les changements sont autant massifs que récents. Ils surviennent si rapidement qu'ils ne peuvent être assimilés qu'après un certain délai. Selon Google Trends, *big data* ne faisait l'objet que de rares recherches avant 2010. Si le terme s'est depuis imposé, c'est qu'il représente désormais un phénomène incontournable.

Médias sociaux, génétique, communication mobile, achats en ligne, cote de crédit, surveillance électronique, stations climatiques, cartes de fidélité, wiki, cartographie interactive, agriculture connectée, ville intelligente… Les données massives s'accumulent partout. Mais elles ne fascinent pas que les fervents de technologie. Elles suscitent évidemment l'intérêt des entrepreneurs en quête de marchés, des financiers en quête de profits, des politiciens en quête d'électeurs. Elles intéressent aussi les décideurs et les développeurs, soucieux de prendre des décisions éclairées. Les données massives sont alors valorisées, dans tous les sens du terme. On les compare parfois au pétrole et on évoque une économie basée sur les données, le nouveau carburant. Apple, Google et Facebook seraient en voie de déclasser Shell, Exxon et Toyota. Elles intéressent enfin les chercheurs de divers horizons, à des degrés et, nous le verrons, pour des raisons qui diffèrent.

Dans cet article, nous nous intéressons à l'impact qu'ont et que risquent d'avoir les données massives sur l'étude des territoires. Comme chercheurs en sciences sociales (géographes en l'occurrence), nous avons l'habitude d'appuyer nos travaux sur des méthodologies quantitatives, sur l'analyse de base de données à l'aide d'outils statistiques. Nos préférences méthodologiques nous ont inévitablement amené à nous interroger sur les possibilités offertes par cette nouvelle réalité. Si les données quantitatives auxquelles nous avons accès nous permettent d'éclairer nos réflexions, elles en posent aussi les limites. En ce sens, les données massives ouvrent de nouvelles perspectives très prometteuses. Ceci dit, nous abordons les données massives avec beaucoup de prudence, d'autant plus que leur collecte est rapide et pourraient être mieux encadrée.

Comme nous le verrons, notre ambivalence se reflète dans la littérature et les pratiques. Pour les uns, les données massives sont une panacée, une ressource qui permettra d'éclairer les décisions, d'optimiser les pratiques, de régler les problèmes. Pour

les autres, c'est une boîte de pandore qui ouvre la voie aux fraudes, à la surveillance et aux abus de pouvoir.

Cet article est le fruit de nos recherches et de nos réflexions. Il s'agit à la fois d'une revue critique de la littérature et d'une réflexion sur la portée des données massives sur la compréhension des territoires. Il ne s'agit donc pas d'un article scientifique traditionnel dans la mesure où aucune hypothèse n'est vérifiée. C'est une description d'une phénomène émergent. En fait, il s'agit de notre lecture personnalisée. Nous avons été surpris de constater que les sciences sociales n'ont générées que peu d'écrits sur ce phénomène pourtant très actuel, surtout en français. Cet article se veut donc un effort pour répondre à un besoin, celui d'introduire les chercheurs aux principaux enjeux que posent les données massives aux sciences du territoire. Ce faisant, nous présentons aussi quelques sources de données massives et discutons de leurs portées et limites.

Définir les données massives

Big data et données massives sont des appellations relativement nouvelles. Elles réfèrent à des changements récents et non à un objet précis. Les (tentatives de) définitions sont donc nombreuses.[1] Néanmoins, certains éléments de définition ressortent plus souvent que d'autres.

Des données

On se rabat souvent sur les mots qui composent « données massives ». L'expression renvoie alors à des données, beaucoup de données. En fait, le caractère massif repose sur l'augmentation récente de la capacité à capter, à stocker et à gérer des informations, et aux immenses banques de données qui en résultent. Les plus vastes exploitent le plein potentiel des supports technologiques et se situent aux limites de leurs capacités. Ce faisant, la collecte, la gestion et l'analyse de ces données demandent de la créativité et des compétences nouvelles. L'augmentation des capacités technologiques correspondrait alors à un saut qualitatif impliquant de nouvelles pratiques.

Si le volume (ou plutôt la masse) est à la base de l'appellation, ce critère serait moins pertinent aujourd'hui, et plus particulièrement pour les sciences sociales. En effet, la progression continue des capacités fait en sorte qu'il est désormais possible de gérer les volumes de données nécessaires à beaucoup d'analyses dans ce champ à partir d'un ordinateur portable. On pourrait arguer que nous en sommes au point où les conditions technologiques ne comptent plus et que seuls les développements conceptuels et méthodologiques sont désormais requis.

Les données massives se distinguent aussi des données traditionnelles sur d'autres critères : la variété, la rapidité (*velocity*), la valeur, la précision (*resolution*), la flexibilité (Hu et al 2014; Kitchin 2014; Boullier 2015). Nous ne nous attarderons pas à chacun de ces critères mais leur nombre montre que les données massives ne sont pas que massives. Elles sont aussi précises, souvent géolocalisées, et régulièrement remises à jour. On pourrait alors même avancer que certaines bases de données peu volumineuses peuvent être considérées comme des données massives. En fait, en sciences sociales, *les données massives renvoient aux bases de données non traditionnelles qui reposent sur des possibilités ouvertes par des technologies récentes* (internet, capteurs et émetteurs, géolocalisation continue, etc.) et qui se distinguent des grandes enquêtes menées par les instituts statistiques.

La nouveauté des dispositifs de collecte fait en sorte que cette dernière a été développée rapidement, sans la réflexion soutenue dont ont pu bénéficier les dispositifs de collecte traditionnels, sans souci de représentativité plus large ou de stabilité dans la définition des concepts ou des mesures (Shearmur 2015). On parle alors de données « crues » (*raw data*), qui comportent d'importantes lacunes et requièrent de grands efforts de structuration même pour les analyses les plus simples. Elles sont souvent collectées de façon ciblée pour des besoins opérationnels (Feinleib 2014) – la gestion de stock, la téléphonie mobile, ou la gestion de réseaux de transport, par exemple. Les données massives peuvent aussi être considérées comme des données secondaires dans le sens où elles n'ont pas été collectées pour répondre à des questions de recherche identifiées à l'avance (Kitchin 2014). Au contraire, elles sont souvent collectées pour des raisons opportunistes (« on va les enregistrer au cas où ») ou involontaires (des « traces » ont été enregistrées par erreur ou par automatisme). Ces « données d'échappement » (*data exhaust*) ont (ou pourront avoir) une valeur marchande et font l'objet de modèles d'affaires. Elles pourront être réutilisées plusieurs fois, pour des raisons différentes et complètement détachées de leur essence d'origine (Mayer-Schönberger & Cukier 2014).

Évidemment, toutes les sources de données massives ne se conforment pas à cette description. Certaines données pouvant être qualifiées de massives font en effet l'objet de réflexion soutenues et d'un suivi méticuleux. Mais étant donné la variété des possibles, ici et tout au long du texte, nous limitons notre description aux caractères qui nous semblent les plus répandus et les plus distinctifs.

Un phénomène

Pour Boyd & Crawford (2012), le *big data* est bien plus qu'une nouvelle forme de données, il s'agit d'un véritable « phénomène culturel, technologique et savant. » (p. 663, notre traduction) Pour elles, le phénomène repose sur une croyance répandue selon laquelle les données massives permettent des conceptions novatrices, auparavant impossibles, dont émane une impression de vérité, d'objectivité et de précision. Miller (2010) évoque l'idée que des structures intéressantes sont cachées dans les données massives et que les méthodes traditionnelles ne permettent pas de les révéler. Il y aurait des trésors à trouver, mais aucune carte pour nous y diriger.

Comme d'autres croyances, celles-ci s'appuient autant sur des faits nouveaux que sur les fantasmes qu'ils stimulent. Faits ou fantasmes, les lec-

tures du phénomène social sont souvent aussi grandioses que les progrès techniques qui le soutiennent. Pour Boullier (2015, 19), « le changement d'échelle entraîne un nouveau cadre de pensée [...] invente un monde. » Kitchin (2014) parle d'une révolution épistémologique, d'un nouveau paradigme, inductif et orienté données (*data driven*). Dans un éditorial qui a suscité beaucoup de réactions, Anderson (2008) laisse même entendre que les données massives vont (enfin!) nous permettre de nous débarrasser des théories et de la méthode scientifique.

Nous reviendrons sur ces enjeux mais, pour l'instant, retenons que les données massives, comme phénomène, ont des conséquences qui les dépassent largement. Révolutionnaires ou pas, l'apparition de sources de données stimule le développement de nouvelles méthodes de structuration, de visualisation, de modélisation, d'analyse et de rapport au réel.

Enjeux

Boyd & Crawford (2012) font l'analyse du phénomène des données massives à partir du cadre des « phénomènes sociotechniques » et mentionnent que, comme les autres phénomènes de cette catégorie, celui du *big data* déclenche des rhétoriques à la fois utopiques et dystopiques mais que les espoirs et craintes masquent des évolutions plus subtiles. Les prochaines pages sont consacrées à quelques-uns des principaux enjeux soulevés pas les données massives. En exposant ces enjeux, nous tâcherons de faire ressortir autant les limites que les ouvertures, les dangers que les potentiels. Aussi, nous ramènerons régulièrement l'attention aux conséquences pour le développement territorial. Ces enjeux sont présentés en trois temps, qui correspondent aux principales étapes liées aux données massives : collecte, partage et analyse.

Collecte

Une collecte inégale : Le phénomène des données massives semble se développer dans un univers étrange, vaporeux, détaché des contraintes du monde matériel. Confortée par l'analogie du nuage, selon laquelle les données seraient partout et nulle part, cette représentation contraste avec la nature résolument concrète des données massives, qui dépendent d'infrastructures bien réelles pour leur production, leur partage et leur analyse (ordinateurs, capteurs, disques durs, fibre optique, etc.).

Cette matérialité n'est pas qu'une anecdote, c'est une condition d'existence. Or, les infrastructures qui supportent les données massives ne sont pas (encore) uniformément réparties entre les territoires et les groupes sociaux. Évoquant la compétitivité et la prospérité économique, le Conseil de la radiodiffusion et des télécommunications canadiennes (CRTC) considère désormais le service internet à large bande comme un service de base, qui devrait être disponible pour tous les Canadiens. Or, 18% des Canadiens, les plus éloignés des grands centres, n'y ont toujours pas accès (CRTC 2016).

On parle alors de fracture numérique (*digital divide*), une réalité qui est très documentée par les organismes internationaux en ce qui concerne les pays en développement. La fracture ne concerne pas uniquement la connexion mais aussi la collecte des données massives. Les informations colligées sont bien plus nombreuses au centre-ville de Montréal que dans un village de la Gaspésie. En fait, la fracture numérique suivrait trois « étapes » : accès à la technologie et aux infrastructures; utilisation effective des technologies; et intégration sociale (Hilbert 2014). La fracture numérique dépasserait donc largement les inégalités en termes d'infrastructure de collecte et de communication. Certains territoires laissent moins de traces et leurs résidents n'ont pas les mêmes possibilités d'exploitation des données massives. De plus, nous y reviendrons, la lecture des données massives demande des compétences spécifiques (notamment en statistique et en informatique), fortement concentrées dans les grandes agglomérations. Enfin, l'attitude générale de la population face aux données massives est tributaire du niveau de pénétration du phénomène. Pour toutes ces raisons, la fracture numérique est un enjeu de développement territorial, malgré la couverture toujours plus complète des réseaux de fibre optique.

De nouveaux objets : Outre la couverture spatiale, les données massives sont caractérisées par une grande présence des informations spatiales, collectées en grandes quantités grâce aux développements concomitants du *Global Positioning System* (GPS). Ainsi, les données massives sont souvent bien adaptées à l'analyse des phénomènes spatiotemporels (Miller 2010). Ces circonstances ont fait en sorte que la géographie a été confrontée plus rapidement au phénomène que les autres sciences sociales, ce qui s'est concrétisé par le développement de la branche des Systèmes d'Information Géographique (SIG).

D'autres prétentions sont moins clairement démontrées. Par exemple, Webber, Butler, & Phillips (2015) avancent l'idée que les données massives ont l'avantage d'être composées de données « factuelles » liées à des comportements réels, plutôt qu'à des déclarations de répondants comme dans les données d'enquêtes. Elles relèveraient alors plus de l'observation non participante que de l'entretien directif. Cette idée laisse entendre que les données massives présenteraient aussi des avantages qualitatifs sur les données traditionnelles : un point de vue plus neutre, moins biaisé par les sensibilités des chercheurs et des répondants. Ce genre de représentations est très répandu chez les utilisateurs de données massives. Sans les contredire catégoriquement, nous croyons qu'elles sont partielles et partiales, qu'elles manquent de nuances, notamment en ce qui concerne d'autres sources de biais qui propres aux données de ce type (par exemple les biais spatiaux déjà évoqués, la supposition que les capteurs et autres technologies sont eux-mêmes infaillibles, et que l'hypothèse que la transformation de pulsions électriques en informations se fait par le moyen d'algorithmes neutres).

Une autre de ces prétentions concerne l'exhaustivité de la collecte, qui serait enfin atteinte (*n = all*). Plus besoin de spéculer sur la représentativité des échantillons, nous aurions aujourd'hui accès à tout l'univers des données (Mayer-Schönberger & Cukier 2014), le volume et la variété se substituant à l'exhaustivité et à la représentativité (Boullier 2015). Ces interprétations sont clairement naïves et réductrices (Boyd & Crawford 2012), et pour plusieurs raisons : la presque totalité n'est pas la totalité, la fracture numérique invalide la couverture totale, etc. Ceci dit, elles témoignent d'opportunités et de cadres nouveaux. Il faut en effet se questionner sur ces univers étranges, qui se déploient en « temps réels » et sont composés de clics, de « j'aimes » (*likes*), de gazouillis (*tweets*) et de *memes*. Et il faut aussi reconnaître que les technologies permettent de rejoindre plus facilement plus de répondants.

Des constructions sociales : Pour le chercheur en sciences sociales qui s'y intéresse, les données massives sont (presque) toujours des données secondaires, c'est-à-dire qu'elles sont collectées avec d'autres objectifs en tête (Mayer-Schönberger & Cukier 2014). Ce contexte contribue à la distinction avec les données traditionnelles, celles des grandes enquêtes, dont la collecte était mûrement réfléchie pour convenir aux besoins en connaissances. Le recensement se trouve à l'intersection de ces deux types de données : pour les chercheurs ce sont des données secondaires, mais la définition des variables et le choix des méthodes de collecte ont été socialement construits et ont fait l'objet de débats afin d'assurer leur pertinence sociale (Shearmur 2010). À la différence des données massives, généralement « opportunistes », le recensement prend soin de représenter le plus fidèlement possible la population.

En effet, la collecte de données massives est différente. Les intérêts qui y prévalent n'ont souvent rien à voir avec la meilleure compréhension des phénomènes sociaux mais répondent à des besoins opérationnels précis et techniques. Les choix de collecte sont alors réfléchis et orientés par ces besoins. Or, très souvent, il s'agit de la recherche de profits ou de la gestion de réseaux. Dans ces cas, les données massives sont « produites » selon les opportunités d'affaires ou les besoins de gestion. Elles sont donc modelées par un certain état d'esprit.

Une fois déclenchée, la collecte s'apparente davantage à la pêche industrielle qu'à la pêche sportive : en collectant les informations souhaitées, on se trouve à ramasser beaucoup d'autres choses. Ces données d'échappement (*data exhaust*, *data fumes*) peuvent alors être étudiées, notamment par les chercheurs en sciences sociales. Mais ces « poussières » ne se déposent pas naturellement sur des réceptacles neutres. Elles sont compilées dans des boîtes, arbitrairement définies par un programmeur, probablement un homme, ayant étudié en informatique, habitant une grande ville, travaillant souvent pour une entreprise transnationale (Finn 2017). Sans y réfléchir spécifiquement, ce programmeur fait des choix qui sont souvent socialement situés et qui ont un impact sur la nature des données collectées.

Ainsi, consciemment ou non, les grands collecteurs orientent la représentation du monde et balisent la formulation des questions, et plus particulière celles qui sont orientées données (*data driven*). Qu'elles proviennent d'une stratégie de collecte ciblée ou de ses échappements, les données massives sont donc des construits sociaux qui, très souvent, sont marqués par des intérêts commerciaux. Ce faisant, la collecte introduit des biais qui déforment les objets mesurés. Le chercheur doit prendre conscience de ces biais s'il veut être en mesure de lire correctement les phénomènes sociaux qui y sont consignés (Shaw 2015).

La construction sociale des données massive soulève aussi la question de la performativité en ce qu'elle introduit de nouvelles catégories qui en viennent à influencer les comportements. Par ailleurs, la collecte et la diffusion rapide des données font en sorte que les rétroactions peuvent être presque immédiates : ces données ne décrivent alors plus le monde, mais le façonnent (Shearmur 2015).

Collecter pour mieux vendre : Cet aspect est d'autant plus inquiétant que l'un des objectifs des grands collecteurs de données massives est d'influencer les comportements de consommation (Bernasek & Mongan 2015). On reproche alors aux données massives d'être un outil développé par et pour le consumérisme. Elles offrent aux détenteurs de données la possibilité de connaître les préférences des consommateurs, de les manipuler et de vendre davantage et plus cher en subdivisant et en ciblant les marchés (ibid.). Ainsi, les cartes de fidélité et autres programmes de récompense serviraient davantage à amasser de l'information qu'à fidéliser ou récompenser.

Ces pratiques suscitent plusieurs inquiétudes. Il est entre autre suggéré que les données des consommateurs vulnérables (problèmes de dette ou de santé mentale) ont une grande valeur marchande (Bernansek & Morgan 2015). De plus, malgré une motivation initialement commerciale, la collecte de données personnelles permet le développement de banques de données très sensibles qui pourraient être utilisées à des fins malveillantes. Finalement, l'avantage compétitif des données massives déstabilise les marchés traditionnels, plus ancrés dans les territoires, et favorise des entreprises apatrides n'ayant plus de comptes à rendre aux territoires qu'elles desservent. Le respect des réglementations locales (droit du travail, droit des consommateurs, optimisation fiscale) devient un problème central associé entre autres à Uber, Airbnb, et Amazon.

Consentement contraint et mal informé : Ces collectes se justifient généralement par l'amélioration des services offerts : plus les entreprises en savent sur nous, mieux elles sauront nous diriger là où nous voulons aller. C'est en effet ce que l'on retrouve régulièrement sur les politiques de confidentialité rendues publiques par les sites collecteurs de données. Généralement bien camouflées, on y explique aussi, quoique vaguement, quelles informa-

tions sont collectées et comment elles seront utilisées par l'entreprise. Très souvent, il y est mentionné, au travers de longs textes en petits caractères, que les données pourraient être partagées avec des tiers et que la politique pourrait changer sans préavis (Bernansek & Morgan 2015). Ces déclarations sont davantage des instruments de relation publique que des protections contre d'éventuelles poursuites qui seraient de toute façon grandement compliquées par l'opacité des utilisations réelles, un cadre légal qui accuse un grand retard sur les technologies, et la difficulté à identifier la juridiction territoriale pertinente (Struijs. Braaksma, & Daas 2014, Bernasek & Mongan 2015, Palfrey & Gasser 2016). Malgré ces enjeux, la grande majorité des consommateurs contribuent aux bases de données, volontairement ou inconsciemment. Ils veulent bénéficier des services offerts par Google, du réseautage de Facebook ou des bas prix d'Amazon. Ces comportements font désormais partie du quotidien et sont difficilement contournables.

Ces conditions sont évidemment très éloignées de l'idéal du consentement libre et éclairé qui est la norme en sciences sociales. Ce décalage implique une réflexion rigoureuse sur l'utilisation, par des chercheurs en sciences sociales, de données dont la collecte ne respecte pas les normes éthiques auxquelles ils sont assujettis (Boyd & Crawford 2012). À ce titre, les instituts statistiques nationaux pourraient jouer un rôle dans la régulation de la collecte de données en mettant à profit leur expertise (fondée sur la qualité, la transparence, l'impartialité, le bien commun, la confiance et le respect des lois et de la vie privée) pour faire la promotion des bonnes pratiques de collecte et établir des certifications pour les bases de données massives (Struijs, Braaksma, & Daas 2014).

Ironiquement, alors que des banques de données massives sont collectées sans surveillance et dans l'ombre, les institutions statistiques voient leurs budgets amputés et doivent composer avec des taux de réponse en chute libre, entre autres alimentés par les inquiétudes d'une partie du public envers un gouvernement qui serait trop contrôlant (Charron 2015). C'est pourquoi, dans leurs contacts avec les données massives, les instituts statistiques et les chercheurs universitaires doivent veiller à maintenir leur rigueur, leur sens de l'éthique et, ainsi, leur crédibilité.

Le bien commun : Bien qu'il motive une large part de leur collecte, le potentiel des données massives ne se limite pas aux intérêts marchands. Elles constituent en effet des sources d'informations inestimables pour la gestion de certains pays, qui ne bénéficient pas de collectes de données nationales efficaces (Kshetri 2014).

Parfois, la collecte de certaines données massives est directement motivée par le bien commun. C'est le cas, par exemple, de ces citoyens de la Pennsylvanie qui documentent les effets néfastes de la fracturation hydraulique sur leur territoire (Gabrys, Pritchard, & Barratt 2016). C'est aussi l'esprit général de la production participative (*crowdsourcing*) et de tous les wiki, suivant lesquels la collecte se fait dans l'intérêt général, de façon ouverte et collaborative.

Les nombreux enjeux discutés à cette section amènent Dalton et al (2016) à proposer le développement des *critical data studies* pour confronter le fait que les données massives sont essentiellement produites par et pour les intérêts commerciaux. Andrejevic (2014) évoque quant à lui un *Big Data Divide* entre les collecteurs et les cibles des collectes. Les données massives créeraient une asymétrie de pouvoir au bénéfice des premiers, qui seraient en mesure d'imposer leurs représentations. De cette asymétrie de pouvoir découle inévitablement une asymétrie d'accès, thème de la section suivante.

Partage

Accessibilité théorique : Jusqu'à tout récemment, la moindre collecte de données impliquait la mobilisation de ressources importantes (Miller 2010). Les recensements de population en représentent un bel exemple. Au 19ème siècle, la collecte se faisait au porte-à-porte et les données étaient « enregistrées » sur du papier. Ces feuillets devaient être entreposés dans des salles immenses et leur interrogation demandait énormément de temps et de ressources.

Les premières « machines » ont permis, au tournant du 20ème siècle, d'automatiser une partie des traitements. Le développement de l'informatique a par la suite progressivement facilité la structuration, le partage et l'analyse des données. Ce n'est qu'à partir de 2001 qu'une partie du recensement canadien a été directement collecté sur support numérique. Cette évolution fait en sorte qu'il est désormais possible de télécharger gratuitement, d'un simple clic, des informations très détaillées sur la population de petites communautés. Nous avons accès à des pétabits de données, de sources très variées.

Mais si l'accessibilité technique et monétaire aux données est plus importante, elle n'est pas totale facilité, et loin de là. Nous avons tenté d'obtenir des informations qui nous semblent pertinentes pour la compréhension du développement territorial auprès de certains grands collecteurs : Vidéotron, Bell, AirBnb, Twitter, Equifax. Ce n'est probablement pas très surprenant mais nos courriels n'ont pas reçu d'accusé de réception et nos appels ne se sont jamais rendus à une personne compétente. Peut-être avons-nous cogné aux mauvaises portes. Peut-être nous sommes-nous perdus dans un monde immense et mal structuré. Mais peut-être, aussi, que ce que nous cherchions est jalousement gardé. Dans tous les cas, l'accès aux données massives est tout sauf facilité.

Sans surprise, les sites internet des différentes entreprises contactées ne sont pas adaptés pour le partage de leurs données massives. Il n'existe aucune section « données ». Le seul élément qui s'en approche concerne les politiques de confidentialité. Ces dernières ont pour fonction de répondre aux inquiétudes du public et d'offrir une protection légale. Elles ne concernent pas le partage des données sauf, parfois, pour mentionner l'éventuel

partage avec des tiers. Mais si ce partage est évoqué, il n'est jamais décrit. On ne trouve pas, par exemple, les coordonnées des responsables des données.

Encadrer le partage : Les intérêts privés dirigent l'essentiel de la collecte, du partage et de l'analyse des données massives. Ainsi, les candides tentatives des chercheurs en sciences sociales, non-profitables au sens du marché, restent lettre morte. En fait, les données massives à la source de la plupart des anecdotes sur leur portée (comme, par exemple, *google flu trend*), celles qui ont probablement le plus de valeurs, sont hautement protégées par des intérêts privés, notamment par un nombre limité de détenteurs de données massives (*big data holders*). La valeur marchande (Webber, Butler, & Phillips 2015) et les sensibilités liées à la protection de la vie privée (Lazer et al 2009) compliquent le partage des données massives. Ce qui nous amène à un paradoxe : une large part des données massives, qui seraient amenées à changer notre monde, ne sont accessibles qu'à un nombre restreint de privilégiés.

Le partage des données massives doit faire l'objet de protocoles robustes pour éviter les nombreux écueils potentiels (Lazer et al 2009). Ces protocoles devront trouver un équilibre entre l'intérêt public, la protection des données personnelles et les intérêts commerciaux. De plus, ils devront permettre la recherche libre, indépendante des contraintes que pourraient vouloir imposer les détenteurs de données (Boyd & Crawford 2012). Mais ces protocoles éventuels ne faciliteront pas pour autant l'accès des chercheurs aux données massives. Il semble évident que les détenteurs de données ne considéreront que les demandes crédibles, celles qui sauront démontrer leur pertinence sociale et leur caractère inoffensif vis-à-vis des intérêts des détenteurs. Or, cette crédibilité sera difficile à bâtir sans un accès minimal aux métadonnées massives, sans une connaissance minimale des variables collectées et de la qualité des données. À notre sens, il y a ici un blocage important. Les chercheurs en sciences sociales auront beaucoup de difficultés à développer des projets pertinents s'ils ne connaissent pas le potentiel et les limites des informations collectées. Cette difficulté nous semble d'autant plus importante que les premières formes d'interrogation des données massives devraient suivre une démarche inductive permettant de révéler leur potentiel explicatif.

Cette remarque vaut aussi pour les demandes d'accès à l'information. Les chercheurs en sciences sociales peuvent en effet se voir octroyer l'accès à des données sans le consentement les détenteurs de données. La Commission d'accès à l'information encadre ces démarches qui concernent aussi les organisations privées. Toutefois, il est évident que, pour aboutir, ces demandes doivent être extrêmement solides et que la méconnaissance des données collectées mine grandement le développement d'arguments bien appuyés.

Le contrôle des données : Il semble légalement permis (et généralement admis) que les collecteurs de données « possèdent » l'ensemble des données. En effet, étant donné la divulgation de politiques de confidentialité, les « répondants » ont techniquement accepté les conditions d'utilisation. Ainsi, les détenteurs de données ont beaucoup de contrôle sur les données et leurs utilisations. Ils s'efforcent de ne pas faire trop de bruits pour éviter une crise de l'opinion publique. Et ils profitent de l'effet de nouveauté, du fait que les possibilités techniques ont une très longue avance sur leur encadrement légal. L'exemple de Google est éloquent. Les informations qu'il collige permettent de connaître intimement nos questionnements et nos habitudes, nos désirs et nos angoisses. Il est toutefois difficile de savoir tout ce que Google sait de nous, et de connaître les intentions de la compagnie quant à la mise en valeur de nos données (Bernasek & Mongan 2015).

Il y a donc une grande asymétrie de connaissance et de pouvoir entre les détenteurs de données et les individus qui y sont consignés. Il y a ceux qui possèdent les données massives et il y a les autres. On compare parfois les détenteurs de données (y compris les services secrets gouvernementaux) à un Dieu omnipotent, à Big Brother ou au panoptique de Bentham. Ces comparaisons sont fortes mais elles constituent néanmoins des mises en garde pertinentes en regard du développement très rapide de nouvelles réalités technologiques et sociales.

Analyse

Créer du sens : Les données massives amèneraient un saut qualitatif dans l'appréciation du réel. L'ampleur de leur couverture donne le vertige et ce, dans toutes les dimensions imaginables : volume, couverture spatiale, groupes sociaux, thématiques abordées, temps réel, etc. Elles offrent un nouveau regard extrêmement large et détaillé sur les phénomènes sociaux qu'elles contribuent d'ailleurs à construire. Mais elles viennent aussi avec de nouvelles limites que nous connaissons encore mal.

Étant donné l'ampleur et la nouveauté des changements, il faut se questionner sur la validité des cadres de pensée développés à l'époque « primitive » des grandes enquêtes (Lazer et al 2009). Le défi est de créer du sens de cet océan de données. On peut imaginer le désarroi des journalistes d'enquête qui se sont attaquées aux dossiers de la fuite des Panama papers. Selon la Wikipedia, elle concerne 11,5 millions de documents. Combien y a-t-il d'aiguilles dans ces énormes bottes de foin? Comment procède-t-on à la fouille? Par où commencer?!?

Pour reprendre l'analogie de Bowker (2005), la cuisson des données crues (*raw data*) est un art qu'il faut maîtriser si on veut harmoniser les saveurs sans trop perdre d'éléments nutritifs. De nouvelles techniques de visualisation se développent, comme les représentations de réseaux sociaux (*social graphs*) et les nuages de mots (*word clouds*). Mais la plupart des développements sont moins visibles et concernent les algorithmes, le forage de données (*data mining*) et l'intelligence artificielle (Finn 2017).

Comme nous l'avons déjà mentionné, la fracture numérique ne con-

cerne pas uniquement la collecte et l'accès aux données, traités aux sections précédentes. C'est aussi une question de compétences et, à cet effet, certains groupes et certains territoires sont mieux pourvus que d'autres. Étant donné le pouvoir potentiel des données massives, ces inégalités ne sont certainement pas banales. Ceux qui pourront les valoriser auront un avantage comparatif.

Méfiance, indifférence et éloignement : Aux inégalités d'accès et de compétences, il faut ajouter des différences d'attitudes. Certains groupes sociaux rechignent à reconnaître la portée des données massives. Nous insistons ici sur les chercheurs en sciences sociales qui s'en méfient pour de bonnes et de mauvaises raisons (Miller 2010; Burrows & Savage 2014; Webber, Butler, & Phillips 2015). Cette attitude laisse les coudées franches à ceux qui s'investissent dans ce champ et, conséquemment, le définissent : ingénieurs, programmeurs, informaticiens, physiciens et biologistes (Lazer et al 2009; Hu et al 2014; Mayer-Schönberger & Cukier 2014). L'absence des sciences sociales ouvre la voie à toutes sortes de prétentions parfois candides et parfois arrogantes, parfois rafraîchissantes et parfois inquiétantes.

Nos lectures ont montré que les chercheurs en sciences sociales qui investissent le champ le font surtout pour le critiquer. Quelques autres travaux visent à décrire le phénomène et discuter des principales avancées, mais il est très rare d'identifier des travaux qui, tout en émanant des sciences sociales, revendiquent l'utilisation de données massives. Certains auteurs expliquent cette timidité par l'inertie des sciences sociales, qui seraient conformistes et conservatrices, ses chercheurs dédaignant déroger de leurs habitudes confortables et maîtrisées pour oser manipuler de nouveaux matériaux (Burrows & Savage 2014; Webber, Butler, & Phillips 2015).

Nous sommes portés à penser que le manque d'implication des sciences sociales s'expliquent aussi par le système de financement de la recherche (qui tend à supporter des projets qui reposent sur des bases solides au détriment des projets exploratoires) et par les stratégies des détenteurs de données (qui partagent données et ressources avec les chercheurs qui les aideront à résoudre des problèmes techniques et non à comprendre des phénomènes sociaux). Toujours est-il que certains chercheurs en sciences sociales sont inquiets. Les sciences sociales pourraient perdre leur place, rendues caduques par les sciences dures qui, elles, seraient désormais mieux positionnées pour rendre compte des faits sociaux (Savage & Burrows 2009; Kitchin 2013; Boullier 2015).

Au même moment, là où ça se passe : En général, ceux qui ont l'opportunité de manipuler les données massives n'ont pas (ou peu) de formation et de contacts avec les sciences sociales. Ils s'avancent néanmoins vers ce terrain, probablement parce que les données les y amènent. Le contact avec les données se fait souvent avec un esprit d'optimisation, défini par les intérêts économiques et organisationnels qui orientent les projets (Kitchin 2014). On ne cherche pas d'abord à comprendre mais à gérer, à vendre ou à diminuer les coûts. On offre alors des recettes miracles et des solutions clés-en-main. Les guides de données massives s'étalent en grands nombre dans les sections « affaires » des libraires (ex. Feinlieb 2014). Les «big data solutions » abondent sur internet. Il est alors question de villes intelligentes, d'efficacité des réseaux, des organisations et des communications.

Forts de ces expériences, certains chercheurs s'aventurent plus loin dans le terrain des sciences sociales et vont jusqu'à se revendiquer comme les successeurs d'Auguste Comte et de sa « physique sociale » (Pentland 2014). Il nous semble que ces travaux ont un intérêt dans la mesure où ils font des propositions novatrices et originales, en grand décalage avec les cadres imposés en sciences sociales. Mais ce faisant, inévitablement, on réinvente aussi la roue, on répète des mêmes vieilles erreurs…

On peut aussi voir une certaine arrogance dans ces propositions, qui prétendent expliquer des phénomènes sociaux sans aucune référence aux travaux des sciences sociales, à l'exception de quelques essais bien diffusés, ceux qui se vendent dans les aéroports. C'est aussi de la « *land-rover research* », qui évoque ces hauts fonctionnaires qui, en quinze minutes, prennent connaissance de la réalité d'un bidonville avant de retourner dans le confort de leur véhicule et hôtel climatisés (Dalton et al 2016). Les données massives permettent en effet un découplage total entre le centre de recherche et la réalité de terrain. Ce phénomène n'est pas nouveau mais les données massives, qui à elles seules demandent beaucoup de temps et de compétences pour être décodées, risquent de l'amplifier. Ce découplage constitue un enjeu particulièrement pertinent pour le champ du développement territorial.

La fin de la théorie : Candides ou arrogants, certains utilisateurs de données massives estiment pouvoir se passer de théories, d'hypothèses et de méthode scientifique (Anderson 2008). Cette prétention s'explique probablement par la nature inductive de leur travail, qui leur laisse croire qu'ils n'ont qu'à « écouter » les données (Calude et Longo 2016). La qualité d'une tendance sera alors déterminée par sa capacité à prédire, quels que soient les mécanismes sous-jacents. Ce raisonnement concorde bien avec la logique marchande suivant laquelle tout ce qui importe est la bonne gestion ou le profit. Ainsi, la causalité n'a que peu d'importance. Qu'elle existe ou non, que les relations soient récursives, contingentes ou indirectes importe peu. Ce qui compte, c'est la corrélation et sa capacité prédictive – corrélations qui sont souvent, comme le soulignent Calude & Longo (2016), fallacieuses. Quand elles disparaissent, on passe simplement à la prochaine tendance. Ces démarches seraient en voie de retirer aux chercheurs en sciences sociales leur statut « d'interprètes privilégiés » des phénomènes sociaux, du moins auprès d'un nombre grandissant de décideurs, autant chefs d'entreprises qu'élus (Mayer-

Schönberger & Cukier 2014; Boullier 2015).

Bien qu'ils soient exagérés, ces discours ont l'intérêt de redorer le blason de l'induction dans la démarche quantitative, qui serait devenue taboue en sciences sociales (Goldberg 2015). L'attitude inductive, qui permet de se laisser surprendre, a pourtant le potentiel de limiter certains biais liés à des cadres d'interprétation rigides et intériorisés. On évoque aussi des conséquences épistémologiques liées au saut qualitatif des données massives (Boyd & Crawford 2012), ouvrant la voie à une troisième génération des sciences sociales (Boullier 2015) ou à un quatrième paradigme des sciences (Kitchin 2014). Dans tous les cas, les données massives demandent la mise à jour de certains cadres… et le développement de nouveaux.

Les voies de la recherche « massive » :
Le modèle de production de connaissance à partir des données massives diffère de celui promu par la recherche universitaire (Lazer et al 2009). Les chercheurs universitaires et gouvernementaux ont longtemps eu un accès privilégié aux données, aux instruments et aux compétences pour interroger le monde social. Leurs résultats, reproductibles et critiquables, sont publiés dans des revues dont l'accès est, en principe, public. La principale motivation est l'avancement des connaissances pour le bien commun. Aujourd'hui, des entreprises privées (comme Google, Facebook ou Apple) collectent des données massives et fines, les analysent et développent des modèles qu'elles gardent secrets, pour exploiter au maximum leur avantage comparatif.

À ce modèle corporatiste, on peut ajouter, de plus en plus, celui de diverses formes de partenariats public-privé. Les centres de recherche universitaires sur les données massives, comme l'Institut de valorisation des données (IVADO) et le Centre de recherche en données massives de l'Université Laval (CRDM_UL), largement financés par le secteur privé, peuvent être inclus dans cette catégorie. Ce modèle repose sur la complémentarité des secteurs et sur le principe que la recherche fondamentale est un bien commun. Il soulève néanmoins d'importantes questions sur l'indépendance de la recherche et le contrôle des programmes par les intérêts marchands.

Quelques chercheurs en sciences sociales choisissent d'investir eux-mêmes le champ ouvert par les données massives (Miller 2010; Savage & Burrows 2014; Webber, Butler, & Phillips 2015). Ils insistent sur la contribution potentielle de décennies de réflexions en sciences sociales au développement des données massives et à leur valorisation pour le bien commun. D'autres – mais ils sont encore rares – tentent simplement d'effectuer des analyses empiriques en employant et en évaluant ces nouveaux types de données (Razin & Charney 2015; van Meeteren & Porthuis 2017). Ces chercheurs visent des analyses « massives » mais éclairées.

Enfin, pour d'autres encore, les données massives doivent faire l'objet de contre-lectures militantes qui permettraient de consolider un discours critique et de révéler « d'autres possibles » (Kitchin 2014; Bruno 2015). Les sciences sociales jouent ici un double rôle d'utilisation et de valorisation des données massives, d'une part, de critique et de sonneur d'alarme, d'autre part. Certains appellent à la constitution des *critical data studies*, qui auraient pour objet les enjeux sociétaux et épistémologiques liés aux données massives (Dalton et al 2016). Cette discipline se trouve dans une position délicate, prise entre l'arbre et l'écorce, où elle critique le phénomène tout en s'y immergeant.

Inventaire de données massives

Cette section est consacrée à la description de quelques unes des sources de données massives qui nous semblent pertinentes pour l'analyse des phénomènes sociaux et plus particulièrement du développement territorial. Cet inventaire n'est évidemment pas exhaustif et nos connaissances de chacune des sources décrites sont loin d'être complètes. Par exemple, nous n'abordons pas les données environnementales ou la production participative (crowdsourcing). Nous espérons néanmoins que le portrait incomplet que nous en dressons pourra être utile aux chercheurs, et qu'il pourra offrir une image plus précise des potentiels ouverts par ces données nouvelles.

Initiatives gouvernementales
Plusieurs gouvernements, de divers pays et paliers, considèrent que les données qu'ils collectent constituent un bien commun et s'efforcent de les partager par soucis de transparence mais aussi dans l'espoir qu'elles pourront être valorisées par les citoyens. Ces démarches sont plus ou moins abouties, selon la durée des programmes et la volonté politique qui les supportent. Dans certains cas, elles permettent la collecte et le partage de données d'une grande qualité, assurée par des organisations spécialisées et dédiées, comme les instituts statistiques nationaux. Dans d'autres, les données sont rendues accessibles par principe mais peu d'efforts sont consacrés pour faciliter leur utilisation; elles sont offertes telles qu'elles sont compilées, selon les besoins opérationnels des fonctionnaires.

Nous ne ferons pas le survol des données partagées par les instituts statistiques nationaux, généralement bien documentées. Nous nous contenterons ici de présenter certaines initiatives de notre connaissance. Par exemple, la Ville de Gatineau[2] s'efforce de rendre accessibles plusieurs de ses données et organise même, comme plusieurs autres municipalités, des hackathons, sollicitant ainsi l'aide du public pour valoriser ses données. Le site de l'Atlas de Gatineau[3] permet de visualiser plusieurs sources de données municipales, comme les fiches de taxations ou le zonage. Bien qu'il ne soit pas possible de télécharger des bases de données directement, ces sites offrent aux chercheurs la possibilité de mieux comprendre la structure des données de la ville et, ainsi, de faire des demandes de données éclairées.

Les municipalités recueillent une panoplie d'informations qui pourraient être utiles aux chercheurs en développement territorial. Ces informations

sont pour la plupart numérisées et les municipalités sont de plus en plus ouvertes à les partager pour des projets qui relèvent de la compréhension des phénomènes sociaux. Pensons, par exemple, aux requêtes et aux plaintes compilées par les centres d'appels non urgents (lignes 3-1-1) ou aux passages enregistrés par les cartes à puce des réseaux de transport en commun. La Ville de Beaconsfield a récemment mis sur pied un programme de collecte intelligente des ordures[4]. Des étiquettes électroniques permettent de mesurer le poids des matières résiduelles déposées à chaque adresse et de distinguer les poids de matières recyclables, de matières compostables et de déchets destinées aux sites d'enfouissement. En plus de leur valeur opérationnelle pour optimiser la collecte, ces informations pourraient être utiles pour mieux comprendre les liens entre les caractéristiques des ménages et leurs productions de matières résiduelles.

Les gouvernements des paliers supérieurs organisent aussi des sites internet destinés à « l'ouverture » de leurs données. C'est le cas de Données Québec[5] qui en plus de rendre disponibles des données, présente une liste des applications qui les mettent en valeur. Notons enfin que différents ministères et organisation gouvernementales rendent disponibles certaines des données compilées, comme le Registraire de entreprises[6] ou le ministère de l'Éducation et Enseignement supérieur[7]. La collecte et le partage de données gouvernementales devraient continuer de se développer dans les années à venir, notamment dans le secteur de la santé.

Données de mobilité

Les données de géolocalisation qui sont collectées pour les utilisateurs de téléphones mobiles ont un intérêt certain pour les géographes et les analystes des transports. En effet, elles offrent un potentiel immense lorsqu'il s'agit de documenter les comportements spatiaux avec un niveau de détail très précis. Comparativement aux enquêtes origines-destinations, elles ont l'avantage de couvrir l'ensemble du territoire (hors régions métropolitaines) et de reposer sur un échantillon bien plus vaste (mais non aléatoire). Ce faisant, elles contiennent des informations pertinentes pour la compréhension des déplacements quotidiens mais aussi pour des déplacements plus rares, de villégiature ou de tourisme. Ces dernières informations sont importantes si l'on considère les enjeux liés à l'économie résidentielle et à la mobilité des capitaux.

Ces données sont toutefois d'une grande sensibilité, puisqu'elle donne la possibilité de carrément « pister » des individus au cours de tous leurs déplacements. C'est la principale raison pour laquelle les compagnies de téléphonie mobile ne partagent pas ces données avec les chercheurs, à quelques très rares exceptions près. Par exemple, en 2013, le Groupe Orange, entreprise française de télécommunications, participait à un projet visant à créer une base de données commune pour éclairer le développement de la Côte d'Ivoire (projet D4D). Les données étaient anonymisées et le partage était contrôlé. Plusieurs universités ont été impliquées dans ce vaste chantier qui a permis, notamment, de développer des méthodes d'identification des zones de pauvreté à partir de l'usage différencié des téléphones et, donc, sans le recours à des données statistiques nationales lacunaires. Depuis, une initiative similaire est lancée au Sénégal[8].

Le fait que ces initiatives n'ont lieu que dans les pays en développement laisse entendre que le potentiel de développement y est supérieur, notamment en l'absence d'autres données de qualité. Mais aussi, la sensibilité du public et des clients est probablement plus grande dans les pays occidentaux. Les chercheurs ont alors recours à des collectes spécifiques à partir d'un petit nombre d'utilisateurs volontaires, fournissant parfois eux-mêmes des téléphones intelligents développés spécifiquement pour la collecte de données (Pentland 2014). Ce type de recherche demande des moyens logistiques et financiers importants et ne met pas en valeur les vastes bases de données présentement collectées par les fournisseurs de téléphonie mobile.

D'autres recherches misent sur les traces géolocalisées laissées par les utilisateurs de médias sociaux. Le *Livehoods Project*[9], par exemple, utilise les gazouillis (*Twitter*) et les check-ins (*Foursquare*) pour cartographier les comportements spatiaux et, ainsi, définir des quartiers ou des aires sociales spécifiques et évolutives. Ces aires seraient plus précises, actuelles et flexibles que celles qui peuvent être développées à partir des données de recensement.

Google

Google est devenu une référence incontournable lorsqu'il est question de données massives. En fait, Google s'est donné comme mission « d'organiser l'information à l'échelle mondiale et de la rendre universellement accessible et utile »[10]. L'entreprise s'est d'abord fait connaître par son moteur de recherche qui s'est rapidement démarqué par sa vitesse d'exécution et la pertinence de ses résultats. Il repose sur un algorithme (*PageRank*) qui permet de trier les pages web selon leur popularité à partir des nombreux liens entre les pages web.

C'est d'ailleurs à partir des informations compilées à partir de son moteur de recherche que Google a développé une des applications les plus emblématiques de l'époque du Big Data : *Google Flu Trends*. L'idée est d'identifier et de localiser les requêtes qui évoquent les symptômes de la grippe et de se servir de cette information pour suivre et prévoir l'évolution des épidémies de cette maladie. À ses début, l'application s'est avérée être très efficace, bien plus précise que les prédictions officielles. Toutefois, les estimations ont été moins bonnes dans les années qui ont suivies. Les requêtes sur le moteur de recherche sont aussi à la source de *Google Trends*, un outil qui permet de connaître la fréquence selon laquelle un terme précis est fait l'objet d'une requête. Il est ainsi possible de connaître la popularité de « Big Data » selon la date, la région et la langue.

Aujourd'hui, Google dépasse largement son moteur de recherche. La

compagnie diffuse d'innombrables informations sur des sujets variés, notamment Google Maps et Google Streetview. Google arrive notamment à documenter en temps réel le niveau de congestion sur les principaux axes routiers sur Google Maps, grâce aux informations obtenues à partir des nombreux utilisateurs de son système d'exploitation pour téléphones mobiles : Android.

Google développe quantités d'autres applications mettant en valeur des données massives comme, par exemple, le Google Car, une voiture sans conducteur qui se déplace en fonction des informations qu'elle capte directement. Les applications sont nombreuses et couvrent de plus en plus de sphères. Elles soulèvent toutefois l'inquiétude quant à la quantité des informations personnelles recueillies et à la possibilité d'apparier ces informations pour offrir un portrait particulièrement intime. Des informations nombreuses et parfois très personnelles sont en effet recueillies à partir du moteur de recherche, de Google Chrome, Google Watch, Android, Gmail ou du Google Cloud et des nombreux marqueurs (cookies) qui leurs sont associés.

La compagnie offre plusieurs services qui permettent notamment de visualiser une partie de ces informations mais l'essentiel des données demeurent confidentielles et inaccessibles aux chercheurs externes. Par exemple, il est possible de « se promener » sur Streetview ou de rechercher une entrée sur Trends mais il n'est pas possible de télécharger une base de données qui rassemblerait les nids de poule photographiés ou les requêtes d'un mot clé selon divers territoires.

Réseaux sociaux

Avec Google, Facebook est l'entreprise la plus emblématique de la nouvelle économie numérique. Elle repose d'abord sur la digitalisation des réseaux sociaux et, de ce fait, elle a très rapidement attiré l'attention des chercheurs qui s'intéressent au capital social, notamment. S'il est clair qu'une amitié Facebook ne vaut pas une amitié réelle, il n'en demeure pas moins que les réseaux complexes qui relient les individus les uns aux autres constituent un matériel riche pour les chercheurs.

Mais Facebook, comme d'autres médias sociaux, ne se limite pas aux réseaux et concernent aussi le contenu des échanges. Ce sont d'ailleurs ces informations qui ont été mobilisées par des chercheurs pour étudier l'influence des émotions sur la participation au réseau[11]. Cette étude illustre le potentiel des données recueillies par Facebook pour la recherche. Elle a cependant suscité la controverse à cause de sa démarche intrusive et de son potentiel manipulateur.

Toujours est-il que les utilisateurs de réseaux sociaux y expriment leurs préférences, leurs humeurs, leurs valeurs et leurs opinions. Ces sites sont donc riches en informations sur les représentations et les comportements, selon de multiples contextes et perspectives. Parmi les réseaux sociaux numériques, c'est Twitter qui est le plus mobilisé par les chercheurs. En effet, il est relativement aisé de télécharger de grandes quantités de gazouillis (tweets), des commentaires de 140 caractères qui permettent d'exprimer opinions et émotions. Ces derniers peuvent être géoréférencés et permettent donc de cartographier la prégnance des idées selon les territoires. On a par exemple représenté la géographie de la haine aux États-Unis à partir de commentaires codés comme homophobes ou racistes[12].

Si les données de Twitter offrent un grand potentiel, il faut aussi souligner leurs limites. La géolocalisation pose le premier problème : les gazouillis peuvent être localisés selon la position exacte de l'appareil duquel ils sont envoyés, le lieu de résidence déclaré par l'utilisateur ou les lieux évoqués dans le message. La seconde difficulté concerne la codification des messages. Par exemple, il peut être très difficile de distinguer automatiquement une opinion assumée de l'ironie. Malgré tout, plusieurs options s'offrent au chercheur pour acquérir et analyser les données de Twitter[13].

Données de consommation

Les comportements de navigation sur internet, les préférences exprimées aux travers des réseaux sociaux et les paradonnées sont désormais utilisées pour le profilage marchand des internautes. À partir de ces informations, les gestionnaires de sites peuvent cibler le placement publicitaire et faire des propositions susceptibles de mener à un achat. C'est d'ailleurs le modèle d'affaire d'Amazon qui, principalement à partir des comportements de ses propres utilisateurs, créé des listes et des suggestions de produits. Après des années de compilation et de développement d'algorithmes, ce modèle est finalement rentable.

Ces informations ont une grande valeur marchande et sont presque toujours gardées secrètes par les entreprises qui les valorisent. Ceci dit, parfois, ces entreprises ont des difficultés à créer du sens de ces vastes bases et sollicitent l'aide d'experts et de consultants. Les partenariats qui en découlent financent des centres de recherche universitaires sur les données massives, comme le Consumer Data Research Centre au Royaume-Uni ou l'Institut de valorisation des données à Montréal. En échange de leur expertise, des chercheurs accèdent à des données d'une grande portée pour la compréhension des phénomènes socioéconomiques. Ils sont cependant contraints de respecter les conditions des détenteurs de données et pourvoyeurs de fonds de recherche.

Certaines entreprises vont parfois solliciter une aide plus globale en rendant certaines données ouvertes. C'était le principe du Netflix Prize, qui offrait un million de dollars au programmeur qui développerait l'algorithme le plus efficace. Pour ce faire, Netflix a rendu public une base de données de plus de 100 millions d'évaluations dénomilanisées sur 18 000 films. Le projet n'a cependant pas été renouvelé après que deux chercheurs ont été en mesure d'identifier des utilisateurs en appariant les données à celles de l'Internet Movie Database (IMDB).

Les données de consommation servent d'abord la recherche de profit.

Toutefois, elles sont aussi riches d'enseignements pour la création de typologies sociales menant à une meilleure compréhension des désirs et des besoins, des préférences culturelles et des comportements économiques. On peut aussi penser que les valeurs des nombreux achats documentés pourraient permettre de développer des indices de coût de la vie plus précis et ciblés, selon les champs de dépense, les groupes sociaux ou les territoires. Cette avenue est d'autant plus prometteuse que les transactions sont aujourd'hui presque toutes numérisées et de plus en plus attachées à un consommateur (par une carte de crédit, une carte de fidélité ou un dossier d'utilisateur). Aussi, dans plusieurs domaines, de nouveaux modèles d'affaires sont explicitement fondés sur la collecte et l'analyse des données pour l'optimisation des ventes : c'est le cas d'Amazon et de Netflix mais aussi de Airbnb, Uber, Expedia.

Données financières

Les informations financières sont parmi les plus sensibles qui soient collectées. Elles se rattachent au statut social mais elles sont surtout sujettes à des revendications diverses, liées à l'impôt et aux cotisations sociales mais aussi à la rémunération et aux habitudes de consommation. Elles sont aussi sujettes à la fraude et aux détournements de fonds. Pour ces raisons, les données financières sont sous haute protection et leur accès est fortement contraint.

Mais elles existent malgré tout. En effet, les informations financières doivent être bien documentées pour, justement, assurer la propriété des capitaux. Elles laissent donc des traces, pour la plupart gérées par des institutions financières et bancaires. Elles permettent de connaître la capacité à payer et les habitudes de consommation, ce qui leur confère une très grande valeur. Les entreprises de crédit et les compagnies d'assurances les mettent ainsi à profit pour mettre sur pied des produits qui optimisent les profits.

Les agences de crédit, comme Equifax et Transunion, rassemblent des données financières massives sur les consommateurs depuis des décennies. Elles vendent ensuite ces informations, notamment aux assureurs. Les particuliers peuvent aussi acheter des données, pour évaluer la solvabilité d'un client ou d'un locataire, par exemple.

La qualité des informations est imparfaite et, parfois, certains individus se sont vu refuser des prêts sur la base de fausses informations. Pour cette raison, les agences de crédits sont tenues de fournir les données concernant un individu gratuitement, une fois par année. Les autres demandes sont dispendieuses et, à notre connaissance, ne font pas l'objet de demandes de chercheurs en sciences sociales. Ces données contiennent toutefois des informations très riches en ce qui concerne des phénomènes sociaux fondamentaux, comme l'endettement, l'épargne et les habitudes de consommation. Il ne fait aucun doute que ces thématiques sont certainement très pertinentes lorsqu'il est question de développement territorial.

Nouvelles collectes académiques

Les technologies informatiques ouvrent la voie à des nouvelles formes de collecte qui permettent aux chercheurs de développer des bases de données bien plus vastes que ce qui était possible auparavant. Les applications de sondage en ligne, comme LimeSurvey ou SurveyMonkey, permettent la construction de plateformes d'enquête très versatiles, que les chercheurs peuvent adapter à leurs besoins et soumettre à des milliers de répondants potentiels. Si ces démarches ont l'avantage d'être flexibles et de faciliter la collecte, cette dernière demeure un défi dans la mesure où l'on vise un échantillon suffisamment vaste et représentatif.

Burrows & Savage (2014) décrivent comment ils ont pu mettre à profit la grande visibilité offerte par le BBC Lab UK pour la collecte massive du Great British Class Survey. Un échantillon de 325 000 répondants a été développé, fournissant des informations sur divers champs, liés aux concepts bourdieusiens de capitaux culturel, social et économique. Ces auteurs déplorent toutefois la réaction générale du milieu universitaire, qui focalisait essentiellement sur les limites de la collecte sans s'ouvrir aux potentiels du nouvel outil ou à la richesse des résultats obtenus. Ils insistent alors sur le fait que toutes les données ont leurs limites et que le chercheur doit les reconnaître et s'y adapter. Il s'agit aussi d'une collecte qui permet la participation des citoyens et la versatilité de l'outil de collecte permet des adaptations fréquentes et rapides, en fonction des problèmes identifiés ou de l'évolution du contexte et des questions de recherche.

Ce projet rappelle aussi celui de la Boussole Électorale, développé par des chercheurs universitaires et offerts en ligne sur le site de Radio-Canada. Cet outil se voulait d'abord éducatif et visait à aider les électeurs à se positionner par rapport aux programmes des principaux partis politiques. Ce faisant, la Boussole Électorale collectait plusieurs informations sur les caractéristiques des électeurs, leurs préférences et leurs opinions politiques. Le projet a été reproduit pour plusieurs élections provinciales et fédérales depuis 2011 et a depuis essaimé vers d'autres pays. Il est aujourd'hui géré par l'entreprise Vox Pop Labs, qui développe des logiciels qui collectent des informations sur les préférences des citoyens.

Données massives et territoires

Nous avons vu que les données massives constituent un véritable phénomène social. Caractérisé par le volume, la technologie et la nouveauté, ce phénomène se développe si rapidement que l'on peine à l'encadrer de balises sécuritaires et réfléchies sur les plans légal, méthodologique et théorique.

La force du phénomène et son manque d'encadrement suscitent alors fascinations et inquiétudes. Ces dernières reposent sur les retombées inégales des données massives, liées à la fracture numérique des territoires et des groupes sociaux. À cet effet, les données massives sont en grandes

parties contrôlées par des intérêts privés et marchands. Elles participent à l'émergence d'une nouvelle économie qui, en plus d'étouffer l'économie locale, échappe à la gouvernance territoriale. Le pouvoir se dissocie de plus en plus de ses emprises territoriales. On s'inquiète alors des conséquences, qui risquent de favoriser davantage le consumérisme et les inégalités au détriment du bien commun et de la démocratie.

Mais les données massives suscitent aussi beaucoup d'espoir. Elles ouvrent de nouvelles possibilités de documentation des phénomènes socioéconomiques qui devraient offrir un éclairage plus vaste et précis aux décideurs, notamment en termes d'information spatiale. Pour être utiles, les données massives doivent cependant être organisées, interprétées et vulgarisées. Or, ces démarches demandent des compétences spécifiques, peu présentent dans les organismes communautaires, gouvernements locaux et autres groupes de citoyens.

Les chercheurs des sciences régionales pourraient ici jouer un rôle central. Pourtant, nous sommes quasiment absents des débats et des projets de mise en valeur, particulièrement dans les régions périphériques. Les collectes massives se font selon les termes définis par les intérêts commerciaux mais nous regardons ailleurs. Nous croyons que les sciences sociales en général, et les sciences du territoire en particulier, peuvent permettre de canaliser le flux des données massives vers l'épanouissement des communautés et des territoires. Nous croyons qu'elles peuvent contribuer à contrôler certains dérapages.

Pour ce faire, il faut d'abord se montrer curieux envers un phénomène aussi fort. Il faut éviter d'en être exclus et promouvoir les échanges d'idées et de données. Par ailleurs, il s'agit d'un beau prétexte pour faire dialoguer qualitatif et quantitatif, théories et pratiques. Les sciences sociales sont fortes d'une longue tradition de pensée critique, notamment au sujet de la collecte et de l'analyse des données, qu'elle devrait partager davantage.

Nous croyons qu'il faut reconnaître les potentiels des données massives. Il faut y réfléchir de manière critique. Nous croyons que les sciences sociales peuvent aider à surpasser les limites des données massives, en identifiant et en décrivant les erreurs de collecte, les obstacles au partage et la construction sociale de l'analyse.

Références

Anderson, C. 2008. The end of theory : the data deluge makes the scientific method obsolete. *Wired*, http

Andrejevic, M. 2014. Big data, big questions : the big data divide. *International Journal of Communication 8*, 1673-1689.

Bernasek, A, & Mongan, DT. 2015. *All you can pay: How Companies Use Our Data to Empty Our Wallets*. New York: Nation Books.

Boullier, D. 2015. Vie et mort des sciences sociales avec le big data. *Socio* 4, 19-37

Bowker, G. C. 2005. *Memory Practices in the Sciences*. Cambridge: MIT Press.

Boyd, D. & Crawford, K. 2012. Critical questions for big data. *Information, Communication & Society 15(5)*, 662-679.

Bruno, I. 2015. Défaire l'arbitraire des faits. De l'art de gouverner (et de résister) par les 'données probantes'. *Revue Française de Socio-Économie 2015/2*, 213-227.

Burrows, R, & Savage, M. 2014. After the crisis? Big data and the methodological challenges of empirical sociology. *Big Data & Society*, April-June : 1-6.

Calude, C, & Longo, G. 2016. The deluge of spurious correlation in big data. *Foundations of Science*, en-ligne, DOI 10.1007/s10699-016-9489-4.

Charron, M. 2015. Le recensement relève du bien commun. *Options Politiques*, mai-juin : 43-45.

CRTC. 2016. http

Dalton, CM, et al. 2016. Critical data studies: a dialog on data and space. *Big Data & Society*, January-June : 1-9.

Feinlieb, D. 2014. *Big Data Boot Camp*. New York: Apress.

Finn, E. 2017. *What Algorithms Want.* Cambridge, MA: MIT Press.

Gabrys, J, Pritchard, H, & Barratt, B. 2016. Just enough data: figuring data citizenships through air pollution sensing and data stories. *Big Data & Society*, July-December : 1-14.

Goldberg, A. 2015. In defense of forensic social science. *Big Data & Society* : July-December, 1-3.

Hilbert, M. 2014. Technological information inequality as an incessantly moving target: the redistribution of information and communication capacities between 1986 and 2010. *Journal of the Association for Information Science and Technology* 65 (4), 821-835.

Hu, H, et al. 2014. Toward scalable systems for big data analytics: a technology tutorial. *IEEE Access* 2, 652-687.

Kitchin, R. 2013. Big data and Human Geography: opportunities, challenges and risks. *Dialogues in Human Geography* 3 (3), 262-267.

Kitchin, R. 2014. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, April-June : 1-12.

Kshetri, N. 2014. The emerging role of big data in key development issues: opportunities, challenges and concerns. *Big Data & Society*, July-December : 1-20.

Lazer, D, et al. 2009. Computational social science. *Science* 323 : 721-723.

Mayer-Schönberger, V, & Cukier, K. 2014. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. New York: Mariner Books.

Miller, HJ. 2010. The data avalanche is here. shouldn't we be digging? . *Journal of Regional Science* 50, 181-201.

Palfrey, J, & Gasser, U. 2016. *Born Digital: How Children Grow Up in a Digital Age*. New York: Basic Books.

Pentland, A. 2014. *Social Physics: How Social Networks Can Make Us Smarter*. New York: Penguin Books.

Razin, E, & Charney, I. 2015. Metropolitan dynamics in Israel: an emerging "metropolitan island state"? *Urban Geography* 36(8), 1131-1148.

Savage, M, & Burrows, R. 2009. Some further reflections on the coming crisis of empirical sociology. *Sociology* 41(5), 765-775.

Shaw, R. 2015. Adoption of geodemographic and ethno-cultural taxonomies for analysing big data. *Big Data & Society*, January-June, 1-16.

Shearmur, R. 2010. A world without data? The unintended consequences of fashion in Geography. *Urban Geography* 31(8), 1009-1017.

Shearmur, R. 2015. Dazzled by data: big data, the census and urban geography. *Urban Geography* 36(7), 965-968.

Struijs, P, Braaksma, B, & Daas, PJH. 2014. Official statistics and big data. *Big Data & Society*, April-June : 1-6.

van Meeteren, M, & Poorthuis, A. 2017. Christaller and "big data": recalibrating central place theory via the geoweb, *Urban Geography*, doi: 10.1080/02723638.2017.1298017

Webber, RJ, Butler, T, & Phillips, T. 2015. Big data and reality. *Big Data & Society*, January-June : 1-4.

[1] Certains efforts de définition renvoient à 3 V (volume, variété, vélocité) alors que d'autres en ajoute deux supplémentaires (véracité et valeur). Si la contrainte liée à la lettre V permet de circonscrire la définition, elle demeure une contrainte inutilement arbitraire. Nous préférons donc ne pas inscrire notre contribution dans ce cadre et proposons alors les éléments de définition qui nous semblent les plus pertinents pour penser les données massives du point de vue des sciences du territoire.

[2] http

[3] http

[4] http

[5] http

[6] http

[7] http

[8] http

[9] http

[10] http

[11] http

[12] http

[13] Pour un exemple : http