Livre publié par Edition Mimésis, Avril 2019

Sabina Leonelli

La recherche scientifique à l'ère des Big Data

Cinq façons dont les données massive nuisent à la science, et comment la sauver

Table des matières

À Χρυσούλα Σφαιρίδου, Luciana e Luisa Leonelli,
les esprits libres devant moi

Introduction

Nous vivons à l'*ère de la post-vérité*. Dans le monde politique comme dans le monde social, en raison de la facilité avec laquelle les informations sont diffusées par les technologies digitales et les réseaux sociaux, il devient plus difficile que jamais de comprendre quelles sources d'informations sont fiables, et sur quels fondements. Pour chaque type d'affirmation – de la véracité du changement climatique au type de régime alimentaire le plus indiqué pour les diabétiques – il suffit d'effectuer une brève recherche sur Google pour trouver aussi bien des avis qui la valident que des avis qui la rejettent. Il y a tellement de données sur internet, générées et mises en ligne par des chercheurs universitaires mais aussi par mille autres sources d'information – des services sociaux aux services commerciaux, des structures sanitaires locales aux institutions publiques, des supermarchés aux réseaux sociaux –, que cet océan de données se transforme inévitablement en cacophonie d'interprétations dissonantes. Nous trouvons des données qui « prouvent » que boire du vin régulièrement est mauvais pour la santé, mais également d'autres qui « prouvent » que les malades cardiaques devraient boire un verre de vin par jour. Des données qui confirment l'effet négatif du plastique sur l'écosystème marin et des données qui le démentent. Des données qui indiquent l'effet négatif de la pollution sur la santé environnementale et des données qui indiquent le contraire. Et – chose peut-être encore plus déconcertante –, nous trouvons des individus qui utilisent exactement les mêmes données pour en tirer des conclusions opposées, et avec des moyens bien difficiles à évaluer pour qui n'a pas les compétences spécifiques. Dans ces cas-là, nous devons souvent nous interroger sur la légitimité et les compétences de ceux qui se présentent comme les interprètes de ces données, et notre jugement sur *quoi* croire se résume à un jugement sur *qui* croire. Ainsi, dans notre monde hyperconnecté et multimédia, nous finissons par nous éloigner toujours plus des décisions fondées sur des données factuelles, en nous fondant au contraire sur les opinions de ceux que nous jugeons dignes de confiance. Le *statut* de la recherche scientifique comme source fiable de vérité s'est en conséquence affaibli, jusqu'à être vu par les politiques, les journalistes et les entrepreneurs sans scrupules comme équivalent à n'importe quelle opinion et donc sans légitimité.

Étant donnée une telle situation, il peut sembler paradoxal que cette décennie soit encore appelée *l'ère des données* : un moment révolutionnaire pour l'innovation technologique et les mécanismes de recherche, et un triomphe du fondement empirique de la connaissance sur la pure spéculation. Grâce aux technologies digitales et aux systèmes de recherche et de communication toujours plus mondialisés, nous avons à notre disposition d'énormes quantités de données – une montagne de faits qui attendent d'être étudiés et interprétés, et dont l'analyse par des algorithmes pour l'apprentissage automatisé est un facteur fondamental dans le développement de l'intelligence artificielle. Ce qu'on appelle les Big Data, ou mégadonnées, contient la promesse d'un changement radical de la façon dont se fait la recherche et se crée la connaissance, que ce soit dans le monde académique ou en dehors. L'analyse des Big Data permet de planifier, d'orienter et de diffuser la recherche de manière innovante et – comme on nous le répète souvent – plus efficace qu'auparavant. Surtout, elle change la façon dont nous pouvons assembler et intégrer les données provenant de sources très variées, et il devient bien plus facile de créer rapidement et à peu de frais des méthodes pour analyser une vaste quantité de données de différents types. La disponibilité des données en grande quantité favorise la création de systèmes informatiques toujours plus puissants pour pouvoir les analyser, et la création de ces systèmes favorise à son tour l'investissement dans l'accumulation des données.

L'accès et l'analyse des données devient donc le moteur de la recherche : un modèle d'innovation que nous appellerons *data-centré*, c'est-à-dire centré sur les données. La disponibilité des Big Data, et la facilité avec laquelle elles sont produites, est une opportunité fantastique d'extrapoler de nouvelles découvertes et de perfectionner les méthodes de calcul informatiques toujours plus autonomes et sophistiquées, en se fondant sur les plus vastes archives de faits jamais enregistrées dans l'histoire de l'humanité.

Comment est-il possible, dans ce monde de Big Data aussi facilement accessibles, que nous ayons tant de difficultés à discerner ce qui constitue une connaissance fiable ? Comment sommes-nous arrivés à douter de toute vérité, au sein d'une réalité débordante de faits ? Et quelles sont les conséquences de cette situation sur le développement de technologies comme l'intelligence artificielle, qui continueront à transformer radicalement notre société dans les années à venir ? Pour répondre à cette question, il nous faut comprendre deux choses fondamentales. La première est le lien entre la production des données et la production de connaissance, et la façon dont les données doivent être gérées afin de pouvoir confirmer ou démentir une affirmation. La seconde repose sur l'énorme difficulté et les immenses ressources nécessaires au traitement et à l'analyse des données au point qu'elles puissent être utilisées pour créer des interprétations fiables et faire l'objet d'une évaluation critique. L'objectif de ce livre est de clarifier ces deux aspects, afin d'illustrer comment les Big Data doivent être préparées et manipulées dans le but de faciliter les analyses et les interprétations, et de réfléchir sur le lien profond entre présupposés théoriques, méthodes et technologies utilisées pour analyser les données et la fiabilité de la connaissance qui en est tirée. Ce livre propose de montrer comment l'adoption effrénée des Big Data, et des moyens automatisés pour les interpréter, peut avoir des conséquences désastreuses pour la crédibilité et la qualité du savoir produit – et comment cette perspective peut et doit être évitée pour le bien de l'humanité et de l'ensemble de la planète.

Mes observations se fondent sur quinze années passées à suivre et analyser les processus avec lesquels les chercheurs produisent, gèrent et interprètent les données comme sources de connaissance. Dans mon travail en philosophie des sciences, je m'intéresse aux systèmes créés au cours de l'histoire pour concevoir des descriptions et des explications du fonctionnement du monde. Ceci est lié d'une certaine façon à l'épistémologie, à savoir la branche de la philosophie qui étudie la manière dont on obtient de la connaissance. Je suis surtout fascinée par la capacité humaine à dépasser nos limites intellectuelles, physiques et sociales pour développer des idées ingénieuses et des technologies extrêmement sophistiquées, qui ont un impact considérable sur l'environnement social et écologique. Pour cette raison, j'ai concentré mes recherches sur l'étude des pratiques et stratégies utilisées par les scientifiques pour produire, diffuser et analyser les données. J'ai examiné les façons dont les données viennent à circuler à travers des contextes différents, et celles dont les chercheurs – particulièrement ceux qui travaillent dans les institutions publiques, dans le domaine biologique et biomédical – gèrent et analysent leurs données : comment ils les agencent et les archivent, comment ils en parlent, comment ils justifient leurs propres actions et surtout comment ils les transforment en source de connaissance. J'ai interrogé des centaines de scientifiques dans le monde entier, notamment de nombreux pays européens, des États-Unis, d'Afrique du Sud, du Nigeria, de Chine et d'Inde ; et j'ai moi-même participé à la création et au développement de grosses infrastructures pour la gestion des données, notamment les règlementations et les institutions introduites récemment par la Commission Européenne pour faciliter l'usage des données de recherche afin de stimuler

l'innovation et atteindre le bien commun[1]. Et évidemment, en tant que chercheuse à la tête de divers projets, je produis et je gère des données de tous types (photographies, vidéos, documents historiques et milliers de pages de transcriptions de mes entretiens avec des chercheurs) ; cette dernière expérience participe également à construire la vision des données que je propose ici.

D'emblée, cette approche montre comment la vie des données est extrêmement complexe, particulièrement quand celles-ci sont conservées, standardisées, partagées et agrégées au sein de banques de données et autres plateformes digitales. L'attention au rôle des données dans la recherche scientifique souligne comment l'utilisation des Big Data présente autant de risques que d'opportunités, aussi bien pour les chercheurs que pour la société en général. Beaucoup des problèmes révélés par ce livre comme étant partie intégrante de la gestion des données vont bien au-delà du monde de la recherche, et ils se manifestent à chaque fois que nous cherchons à juger les fondements empiriques de la connaissance à laquelle nous nous fions – que ce soit quand nous lisons un article étonnant, décidons d'utiliser un médicament particulier ou recherchons des informations sur Google. La recherche scientifique offre un microcosme dans lequel les questions méthodologiques et épistémologiques sur la manière dont les données génèrent de la connaissance, peuvent être abordées et étendues à d'autres situations de recherche (par exemple dans le journalisme, la politique, dans le secteur privé et dans les services publics).

L'étude de la façon dont les données circulent à travers différents contextes nous apprend qu'il n'y a pas moyen de séparer clairement les données scientifiquement pertinentes de celles qui ne le sont pas. Tout dépend de la situation dans laquelle les données sont utilisées. Les données personnelles comme la taille, le poids et la situation familiale par exemple, peuvent présenter un intérêt médical si elles sont utilisées par un médecin à des fins de diagnostic ; un intérêt scientifique si elles sont utilisées par un épidémiologiste pour étudier la santé d'une population ; un intérêt affectif si elles sont collectées pour tracer un arbre généalogique ; ou un intérêt commercial si elles sont utilisées par une chaîne de supermarchés pour identifier les préférences d'achat. Les données acquièrent une valeur différente selon les mains dans lesquelles elles tombent, et sont toujours pertinentes dans plusieurs domaines. Par ailleurs, les données peuvent être valorisées de beaucoup de manières et pour de nombreuses raisons différentes au même moment. C'est justement cette multiplicité qui les rend intéressantes comme objets d'analyse : d'une part, les données promettent de documenter certains aspects de la réalité de manière fidèle et précise, de façon à en faciliter l'étude ; d'autre part, la valeur attribuée aux données chaque fois qu'elles sont utilisées a un effet déterminant sur la manière dont elles sont gérées et interprétées.

Pour comprendre le rôle des données dans la société contemporaine, il est inévitable et essentiel d'admettre que *tous* les types de données (qu'elles soient ou non produites et utilisées par les chercheurs, et qu'elles soient ou non reconnues comme sources légitimes de connaissance) ont une valeur commerciale potentielle, en particulier lorsqu'elles sont agrégées pour analyser et prédire les comportements de masse. Il suffit de jeter un œil à la liste des industries ayant le plus grand et le plus rapide succès au niveau national et international pour comprendre que les entreprises et les *start-up* qui s'occupent de l'analyse des données ont eu une croissance exponentielle au cours de la dernière décennie, et que leurs services sont désormais acceptés

[1] De 2016 à aujourd'hui, j'ai travaillé comme experte en Science Ouverte et représentante de la Global Young Academy pour la Commission Européenne.

comme une part essentielle de n'importe quel secteur – de l'organisation d'une campagne électorale au lancement d'un nouveau produit. Collectionner, mobiliser et analyser des données n'est pas une activité limitée au monde de la recherche, mais plutôt une expression fondamentale du développement économique de type capitaliste qui caractérise le libre-échange mondial. Ce n'est pas par hasard que la croissance de Google, Apple, Facebook et Amazon ait été d'une rapidité vertigineuse, jusqu'à ce que ces entreprises soient parmi les plus riches et les plus puissantes au monde, et que des *start-up* comme Prophese, GreenFlex, Linkfluence et EnergyWay, soient considérées comme les plus prometteuses pour notre développement économique aussi bien à l'échelle nationale qu'à l'étranger. La croissance économique se fonde toujours un peu plus sur la création de services personnalisés et optimisés pour respecter les exigences de clients spécifiques, comme par exemple le calcul de la consommation énergétique d'un foyer ou les différentes applications de nos téléphones qui promettent de mesurer quotidiennement notre activité physique et nos conditions de santé. Ces services sont rendus possibles par le développement d'algorithmes sophistiqués et de stratégies pour analyser les comportements des consommateurs, algorithmes qui fonctionnent uniquement s'ils sont capables de puiser au sein de vastes sources de données sur les individus et sur leurs conditions de vie (résidence, environnement, transports, etc.).

Ces considérations illustrent comment l'épistémologie, l'éthique et l'économie politique sont des aspects complémentaires et essentiels pour la compréhension du fonctionnement des Big Data. C'est seulement en ayant une vision complexe du rôle social, culturel, économique et politique des données que nous pouvons comprendre l'effet des Big Data sur le monde scientifique et ce que cela signifie pour la société. Mon analyse dans cet ouvrage se fonde sur l'hypothèse essentielle que la science des Big Data ne se discerne ni ne se sépare facilement du monde extérieur à la recherche : comme nous le verrons dans les prochains chapitres, intérêts et valeur commerciale, politique, affective et économique sont inévitablement liés à l'éventuelle valeur scientifique des données comme sources de connaissance.

Chapitre 1. Que sont les Big Data ?

Les Big Data : entre quantité et qualité

Il y a tant de façons de qualifier les Big Data[2]. Un point de départ accepté par beaucoup est la *quantité*. Les technologies digitales développées au cours des trente dernières années fournissent une puissante capacité de production, de conservation, et d'analyse d'un nombre croissant de données. Ce n'est pas par hasard que les deux caractéristiques les plus souvent associées aux Big Data sont le volume et la rapidité. Le *volume* se réfère à la dimension des *fichiers* utilisés pour archiver et diffuser les données et qui, grâce au pouvoir croissant des processeurs électroniques, augmente vertigineusement et d'une manière impossible à percevoir clairement pour le système cognitif humain (qui, parmi nous, comprend vraiment la différence entre un trilliard et un quadrilliard, nombres qui, pour les gens qui travaillent avec les Big Data, sont relativement normaux ?). La *rapidité* se réfère au rythme effréné et toujours plus soutenu avec lequel les données sont générées par des technologies, comme par exemple le séquencement du génome.

En mettant l'accent sur le nombre de données et le format digital, cette définition ne tient pourtant pas compte des quatre facteurs qui concernent la *qualité* des Big Data et qui sont fondamentaux pour leur utilisation :

1) La variété des types de données en usage, comprenant des données en format non-digital (comme par exemple les données imprimées sur papier) et des données qui, bien qu'étant en format digital, ne sont pas facilement analysables par des algorithmes (les photographies par exemple) ;
2) Le fait que ce qui est perçu comme de grandes quantités de données et une grande vitesse dépend complètement des technologies pour produire les données, les archiver et les analyser, et donc change continuellement d'une année à l'autre. Par exemple, alors qu'au début du millénaire, les Big Data étaient des données trop nombreuses pour être annotées avec un *spreadsheet*[3] normal de Microsoft Excel, désormais, on peut concevoir des trillions de données obtenues par l'usage des *social media*[4] comme Facebook ; au contraire, il y a trois siècles, on avait les collections de milliers d'observations faites par les métrologues, cartographes, et astronomes du monde entier, extrêmement difficiles à analyser et à intégrer en cartes géographiques sans accès à un ordinateur[5] ;
3) La dépendance de l'analyse des données du contexte dans lequel elles sont évaluées et utilisées, qui peut immensément varier selon la situation et les demandes formulées par les analystes – un facteur fondamental pour mon analyse, sur lequel je reviendrai au chapitre trois ;

---

[2] Kitchin et McArdle (2016) ont identifié vingt-six moyens de décrire les Big Data qui sont utilisés dans la littérature scientifique. Le site suivant en contient encore plus : https://datascience.berkeley.edu/what-is-big-data/

[3] « Tableur », en anglais dans le texte. [NDT]

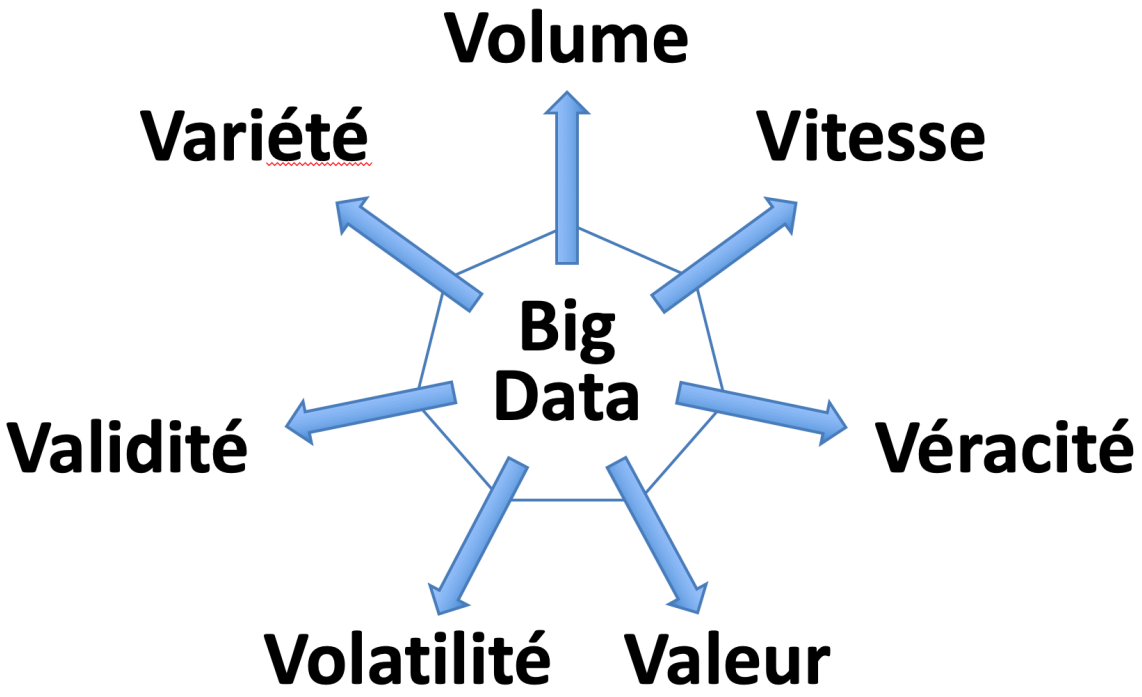
[4] « réseaux sociaux », en anglais dans le texte. [NDT]

[5] Comme l'ont montré de nombreuses études, ces défis ne sont pas nouveaux dans l'histoire de la science. Chercheurs en astronomie, météorologie et taxinomie s'occupent de gérer d'énormes quantités de données depuis des centaines d'années. Pour les études historiques de ce phénomène, on peut se référer aux éditions spéciales *Data-Driven Research in Biology and Biomedicine* dans la revue « Studies in the History and the Philosophy of the Biological and Biomedical Sciences », particulièrement à la contribution de Müller-Wille et Charmantier (2012) ; et *Historicizing Big Data* dans la revue « Osiris » (Aronova *et al*. 2018).

4) Le fait qu'il soit impossible d'analyser les Big Data sans avoir accès à ce qu'on appelle les *métadonnées*, c'est-à-dire les informations sur leur provenance (comment elles ont été générées, par rapport à quoi et dans quelles circonstances) qui permettent aux analystes d'évaluer si les données sont fiables, et quelles interprétations sont plausibles.

Pour tenir compte de ces aspects, d'autres caractéristiques ont été associées aux Big Data ces dernières années (figure 1)[6]. Au-delà de la variété des formats, on parle également de la *variété* des phénomènes auxquels les données peuvent se référer, et des approches utilisées pour les analyser ; de la *véridicité* dans l'interprétation des données et dans le mode au sein duquel elles représentent la réalité[7] ; de la *validité* des données par rapport aux modes dans lesquelles elles sont analysées ; de la *volatilité* dans le temps, c'est-à-dire la capacité des données à rester fiables et lisibles malgré l'évolution des nouvelles technologies de stockage ; et de la *valeur* qui leur est assignée par divers secteurs de la société, y compris celle, incroyablement variable, qui dépend de la période historique ou de la localité.

Figure 1. Les sept « v » des Big Data (réalisation Michel Durinx)



Les Big Data ne sont donc pas seulement de « nombreuses données ». Ce qui les caractérise vraiment sont les différents modes dans lesquels elles sont produites et diffusées à travers divers

[6] Pour une étude des « v » utilisées pour caractériser les Big Data (qui varient de trois à dix secondes selon qui les lit), on peut consulter par exemple Nordmandeau 2013 ; Ward et Backer 2013 ; Mayer-Schönberger et Cukier 2013 ; Marr 2015 ; Kitchin 2014 ; Borgman 2015 ; Sætnan *et al.*2018.
[7] Cai et Zhu 2015 ; Floridi et Illari 2014.

secteurs sociaux. C'est en cela que consistent le pouvoir et la véritable promesse des Big Data : *permettre d'instaurer des connexions entre des secteurs et des approches avec lesquels il était auparavant difficile – que ce soit par les barrières sociales ou pour des raisons techniques – de dialoguer directement*. Au lieu de chercher à définir ce que sont les Big Data en termes de caractéristiques physiques et de quantité, je propose donc de les caractériser en vertu de la façon dont elles sont utilisées. Les Big Data sont des données de type et de provenance différents, qui sont mises en relation les unes avec les autres, souvent sous forme digitale et d'une manière qui se prête à l'apprentissage automatisé, afin de produire de nouvelles formes d'analyse et de connaissance. Comme deux éminents sociologues des données, Boyd et Crawford, l'ont énoncé, l'expression « Big Data » indique « la capacité d'explorer, d'agréger et de relier de vastes ensembles de données[8] ». Pour comprendre comment fonctionnent les Big Data, nous devons donc nous intéresser aux structures, institutions et compétences/professions qui rendent possible cette capacité.

Révolution ou exagération ? Les Open Data et l'approche data-centrée

Les Big Data ont été accueillies et décrites comme une révolution de la façon d'acquérir de la connaissance – d'après Mayer-Schönberger et Cukier, c'est « une révolution de la façon dont nous travaillons, nous vivons et nous pensons[9]. À la suite de ce type de déclarations triomphales, les opportunités inhérentes à l'analyse des Big Data grâce à des algorithmes toujours plus intelligents ont généré d'autres attentes de la part des gouvernements, de l'industrie et des chercheurs du monde entier. D'une part, la disponibilité des Big Data promet de rendre plus précis les modèles utilisés pour calculer et prévoir de futurs scénarios – que cela concerne les prévisions météorologiques ou la probabilité qu'un individu déterminé tombe malade du cancer. D'autre part, la possibilité de relier entre elles les données d'origines différentes promet d'ouvrir de nouvelles voies pour la recherche, et de révéler des corrélations et des liens jusqu'alors invisibles aux chercheurs, mais facilement identifiables par les ordinateurs. L'utilisation des Big Data est donc strictement associée à une accélération, non seulement dans la production d'un nouveau savoir scientifique, mais aussi dans la répercussion de ce savoir en innovations et produits pour l'utilisation de tous les jours – certains d'entre eux, comme par exemple la surveillance de la propagation des maladies contagieuses et la capacité à prévenir des désastres environnementaux, peuvent servir à la résolution des grands défis sociaux de notre époque, comme le changement climatique, les pandémies et la pollution.

Les Big Data sont également liées à une révolution (toujours en cours) de la manière dont on communique les résultats de recherche, et que l'on appelle « Science Ouverte ». C'est l'idée – qui n'est certainement pas nouvelle pour les chercheurs, mais toujours plus problématique à réaliser en pratique, étant données la compétitivité du monde scientifique et la privatisation croissante des découvertes, sous forme de permis et de droits d'auteur – que les textes, le *software*[10], ou les données produites au cours de la recherche doivent devenir accessibles facilement, et sans coût, à quiconque veut les utiliser. La dynamique en faveur de la Science Ouverte est particulièrement forte dans les confrontations de projets sponsorisés par des organismes publics. Différents gouvernements, européens ou non, ont récemment avancé que leurs citoyens, puisqu'ils soutiennent la recherche public par le paiement des impôts, ont le

[8] Boyd et Crawford, 2012, p. 663.
[9] Mayer-Schönberger et Cukier (2017) proposent une approche triomphaliste du pouvoir des Big Data que j'ai fortement critiquée par le passé (Leonelli, 2014).
[10] « logiciel », en anglais dans le texte. [NDT]

droit d'accéder à tous les résultats produits par les scientifiques grâce aux financements étatiques, y compris leurs données[11]. Cette notion d'Open Data, ou données ouvertes, est particulièrement attractive pour ceux qui travaillent sur les Big Data, avec l'idée que plus les données sont disponibles pour être analysées et mises en relation librement entre elles, plus les opportunités de faire de nouvelles découvertes augmentent[12]. Pour cette raison, beaucoup de grandes corporations, comme par exemple l'industrie pharmaceutique GlaxoSmithKline et le géant biotechnologique Monsanto (bientôt Bayern), sont en train d'« ouvrir» certaines de leurs données, dans l'espoir que cela facilite les collaborations avec le secteur public et ouvre la voie à des analyses plus fructueuses et mieux informées que celles que ces compagnies peuvent développer *in-house*[13].

Le lien entre les Big Data et les Open Data est la raison même pour laquelle l'avènement des Big Data a une portée révolutionnaire. En soi, tirer parti des nouvelles formes d'agrégation de données grâce aux nouvelles technologies n'est pas nouveau pour le monde de la recherche. La diffusion de la presse au XVII[e] siècle a transformé les moyens dont les biologistes communiquent et vérifient la découverte de nouvelles espèces, et a déterminé le développement des systèmes modernes pour quantifier la biodiversité (c'est-à-dire le numéro et le type d'espèce qui se trouve dans chaque écosystème, qui sont fondamentaux pour identifier les cas d'extinction ou l'évolution de nouvelles espèces). L'invention de systèmes de stockage comme les *punch cards*[14] a complètement modifié le travail des démographes et des épidémiologistes au XIX[e] siècle, en les aidant à discerner les comportements de populations entières au niveau national. Les techniques statistiques et le développement d'ordinateurs toujours plus puissants dans les premières années du XX[e] siècle ont apporté un niveau d'innovation similaire. Et il est évident que l'histoire entière de la médecine est faite de tentatives pour mettre ensemble des types de données complètement différentes (le régime alimentaire et l'anatomie, ou les antécédents familiaux et l'exposition climatique), afin de trouver des combinaisons de facteurs qui s'associent de manière régulière lors de l'apparition d'une maladie. Ces efforts ont, à leur tour, généré de nouvelles approches de stockage et de visualisation des données, comme les nomenclatures utilisées pour harmoniser le vocabulaire médical, de manière à ce que les médecins se comprennent entre eux et puissent échanger leurs observations ; les techniques utilisées pour s'assurer que les données recueillies sur une longue période demeurent comparables ; et les registres qui conservent de manière fiable et confidentielle les données personnelles pertinentes pour la recherche épidémiologique, ce qui permet aux scientifiques de les analyser sans nécessairement trahir la *privacy*[15] des patients[16].

Ce n'est donc pas seulement l'opportunité d'analyser une vaste quantité de données avec de nouvelles technologies qui distingue ce moment historique des époques précédentes dans l'histoire de la science. Ce qui le rend extraordinaire est plutôt *le statut acquis par les données*

[11] Pour une analyse détaillée du concept de Science Ouverte on peut consulter, le rapport de la Royal Society (Boulton *et al.*, 2012). Pour une analyse de ses implications pour la façon dont la recherche est évaluée et valorisée, on peut voir le rapport que j'ai écrit à ce propos pour la Commission Européenne (European Commission 2018).
[12] Pour une analyse du rôle des Open Data dans le monde des Big Data, consulter le dernier rapport de *Science International* (2015). Les principes associés à l'utilisation des Open Data sont développés par Mauthner et Parry (2013).
[13] « en interne », en anglais dans le texte. [NDT]
[14] « cartes perforées », en anglais dans le texte. [NDT]
[15] « vie privée », en anglais dans le texte. [NDT]
[16] On peut consulter par exemple l'analyse des techniques d'inférence épidémiologique de Broadbent (2013), et la remise en cause de l'harmonisation du discours médical proposée par Ankeny (2014).

elles-mêmes comme composante et résultat fondamental de la recherche scientifique, dont la production implique une obligation de partage, afin que les données soient réutilisées par le plus de personnes possibles (Open Data), et donc servent de passerelles grâce auxquelles différents groupes sociaux peuvent communiquer et travailler ensemble (Big Data). Jusqu'à l'avènement de ce qui est généralement reconnu comme le premier journal scientifique en Europe, le *Philosophical Transactions* de la Royal Society à Londres (fondé en 1665), les données ont été conceptualisées et manipulées comme des objets privés, propriété des scientifiques qui les produisent et qui seuls ont la capacité de les interpréter correctement. Encore maintenant, une très grande majorité des publications scientifiques ne publie qu'une infime partie des données produites dans le cadre d'un projet, partie sélectionnée pour prouver et convaincre de la véridicité de l'interprétation proposée par les auteurs. Toutes les données qui ne fonctionnent pas bien comme preuve dans ce sens, ou qui pourraient même être interprétées de manière très différente, sont éliminées et, d'une manière générale, ne sont pas soumises à des analyses ultérieures par les autres scientifiques. Dans cette approche, que nous appellerons *théorie-centrée*, l'utilité des données réside dans le fonctionnement comme preuve de la plausibilité d'une hypothèse déjà donnée, confirmant ainsi des affirmations que les chercheurs interprètent comme de nouvelles découvertes. La connaissance théorique est vue comme un guide et une référence fondamentale, non seulement pour savoir comment mener des recherches scientifiques, mais aussi pour savoir comment en concevoir l'objectif final. Dans la vision théorie-centrée, la chose la plus importante que les scientifiques produisent sont les nouvelles théories, généralement exprimées à travers des formules ou des textes, et publiées au sein d'articles et de livres qui remplissent, sans hasard, les étagères de chaque bibliothèque universitaire. Tous les autres composants de la recherche, des données aux modèles, des techniques aux instruments, sont vus comme secondaires à la création de théories.

Cette approche est désormais en train de laisser la place à un moyen bien différent de penser la recherche. Dans la vision data-centrée de la science, l'objectif de la recherche comprend la production de différents types de résultats, qui incluent aussi bien les théories, que les modèles, les techniques d'investigation, les méthodes expérimentales et les données elles-mêmes. Les données ne sont donc plus simplement un pas vers la création de nouvelles théories. Dans l'approche data-centrée, les données sont vues comme des entités publiques qui ont une valeur scientifique indépendante de leur rôle de preuve pour une hypothèse déterminée, et qui peuvent être interprétées de façons différentes selon l'habileté et les intérêts des chercheurs qui les analysent. Nous sommes donc en train d'assister à une *réévaluation radicale de la potentialité des données à générer de la connaissance* – une révolution dans laquelle les efforts et les technologies dévolus au partage, à la mobilisation, à la visualisation et à l'intégration des données sont vus non seulement comme des instruments de découverte, mais comme d'importantes modalités de découverte en soi[17]. Cette révolution dans la communication des données s'est répercutée également sur le secteur privé, où la valeur économique potentielle des données ne facilite pas toujours leur diffusion de manière « ouverte », mais encourage certainement la vente et l'achat de données, ainsi que leur utilisation comme moyen de transaction, ou bien comme monnaie dans les échanges commerciaux. Il y a de nombreuses façons « d'ouvrir » les données, la plupart d'entre elles ne prévoyant pas une ouverture totale

[17] Hey *et al.*(2009), un ouvrage produit par des scientifiques de Microsoft et librement accessible sur internet, propose divers exemples de cette tendance. Une défense détaillée de cette interprétation du data-centrisme se trouve dans l'ouvrage *Data-Centric Biology*, (Leonelli, 2016a).

et gratuite à quiconque la réclame, mais plutôt la possibilité de les acquérir selon des conditions déterminées[18].

Big Data en mouvement : le pouvoir des infrastructures

Le développement du data-centrisme et la combinaison des Big et des Open Data ont d'énormes implications pour la façon dont la recherche scientifique – et en particulier l'organisation et l'utilisation des données – est menée, organisée, dirigée et évaluée. La capacité des données de servir de sources de connaissance repose sur leur *mobilité* : c'est-à-dire leur capacité de circuler au sein de différentes situations d'analyse et de réutilisation, et d'être liées avec tous les types de données différentes possibles. Sans mobilité, il n'y aurait pas de Big Data, parce que les données ne réussiraient jamais à sortir des situations spécifiques dans lesquelles elles sont générées, et il serait impossible d'agréger et de confronter entre elles les données récoltées dans des circonstances différentes.

L'importance de la mobilité explique comment des données produites au format standard et par des technologies hautement diffusées et conventionnelles, comme par exemple les données géographiques GPS fondées sur les mesures satellites, peuvent être d'une grande valeur scientifique et commerciale : grâce au niveau d'harmonisation et de digitalisation de ces données, il est facile de les partager et de les mettre en relation avec d'autres types de données (comme par exemple la localisation des restaurants ou l'intensité du trafic), ce qui permet ensuite de générer des cartes et des indications routières en temps réel, comme celles de Google-Maps. La majorité des données pertinentes pour l'analyse scientifique, et particulièrement dans le domaine biologique et biomédical, ne sont pas pour autant harmonisées de cette façon. Au contraire : ces données sont généralement très différentes, de par les techniques utilisées pour les produire, leur format, ou par le type d'objet auquel elles se réfèrent. Des observations recueillies sur la forme des feuilles d'une espèce de plante spécifique, par exemple, peuvent varier de par la façon dont les chercheurs mesurent la superficie de la feuille, dans les critères utilisés pour choisir les feuilles à mesurer, dans le type d'instruments utilisés pour relever la rugosité et la couleur des feuilles, dans la fréquence avec laquelle les mesures sont prises, et enfin dans les noms utilisés pour désigner la plante en question et les différentes parties de la feuille. Ces variations dépendent de la tradition scientifique du groupe et du type d'usage que les chercheurs font des données ainsi recueillies : des chercheurs qui espèrent enquêter sur les corrélations entre la forme et le profil génétique de la plante utilisent souvent des mesures, des noms et des critères différents de ceux retenus par des chercheurs intéressés par la relation entre la taille des feuilles et leur vitesse de croissance[19].

Il est important de noter comment la pluralité des approches ne se vérifie pas par hasard ou par manque de coordination entre les chercheurs. Elle a au contraire une fonction épistémique très précise. Ces domaines scientifiques, ainsi que la recherche sur l'environnement, le climat et la

[18] Les enthousiastes des Open Data dans le milieu scientifique voudraient voir toutes les données pertinentes pour la recherche considérées comme des biens publics disponibles sans conditions à quiconque les réclame. Cette position n'est pourtant pas tenable, étant donnés les coûts soutenus par le secteur privé pour créer des données, et la possibilité d'abus liée à la circulation de certains types de données (les données personnelles par exemple). Le panorama de modalité d'ouverture est donc très varié, avec différentes options et possibilités de diffusion. Les principes FAIR (pour *Findable*, *Accessible*, *Interoperable* and *Reusable*), par exemple, ont été proposés comme alternatives aux modèles d'ouverture radicale (Wilkinson *et al.*, 2016).

[19] J'ai développé cet exemple en détails dans Boumans et Leonelli (2019). Pour une étude détaillée de cas équivalents dans le domaine biomédical, voir Leonelli (2017a).

géologie, sont dédiés à l'étude de phénomènes grandement variables dans le temps et l'espace. Expliquer les caractéristiques et les comportements d'espèces, d'individus, d'organes ou d'écosystèmes particuliers implique inévitablement d'étudier la spécificité de ces systèmes, de façon à pouvoir interagir avec eux de manière efficace (on invente ainsi des stratégies pour augmenter le rythme de croissance des plantes économiquement importantes, comme le blé, ou des interventions utiles au soin de maladies particulières, comme la terrible épidémie causée par le champignon *Fusarium*, qui menace de faire disparaître définitivement la banane du marché occidental). Au cours des siècles, les scientifiques ont élaboré des méthodologies hautement spécialisées pour pouvoir étudier et comprendre les propriétés uniques de ces phénomènes. La variété et la potentielle incompatibilité des données qui résultent de ces études ne doivent donc pas surprendre : elles sont plutôt une conséquence inévitable de la nécessité de produire des systèmes de connaissance qui s'adaptent le plus possible à la nature des divers objets et opérations d'intérêt scientifique – une situation que les philosophes nomment « pluralisme[20] ».

Le pluralisme scientifique génère d'énormes problèmes lors de la mobilisation des données. Avant tout, il faut trouver des moyens de les stocker de manière sûre et accessible à quiconque souhaite tenter de les analyser. Les archives ne peuvent cependant pas être de simples contenants (« *data dumps* ») au sein desquels les données sont jetées à peine ont-elles été produites. L'archive fonctionne seulement si elle est accompagnée de systèmes pour chercher et trouver les données de manière systémique et pertinente pour leur réutilisation. Comme de nombreux philosophes du XX^e^ siècle l'ont déjà signalé, parmi lesquels Michel Foucault et Jacques Derrida, la clé pour la gestion d'une archive est la façon dont elle est organisée[21]. La structure de l'archive, les mots-clés utilisés pour catégoriser et retrouver les données, les modèles et algorithmes utilisés pour les visualiser – tous ces éléments ont un effet décisif sur la façon dont les données sont interprétées et réutilisées, particulièrement dans les cas où l'interprétation est générée par des personnes qui n'ont absolument pas participé à la création des données elles-mêmes, et qui ne connaissent donc pas les circonstances de production. En même temps, la nécessité de mettre en ordre les données transforme leur gestion en un problème épistémologique et de gestion complexe, surtout puisque personne ne sait clairement quel type de classement est le plus ou le moins utile à l'interprétation des données et à leur transformation en nouvelles connaissances – ni même si un tel classement « idéal » existe, étant donné que l'organisation des données est conditionnée par le contexte et l'objectif pour lequel les données sont analysées.

Comme nous l'avons déjà vu, la capacité à lier les données obtenues de sources diverses, et donc pas immédiatement comparables, est particulièrement importante dans le contexte des Big Data. Cela a des conséquences non seulement pour le classement, mais aussi pour le formatage des données – la forme qui leur est donnée pour les faire circuler (par exemple le type de *fichier* utilisé pour les digitaliser). Cela implique de faire en sorte que le format des données soit tel qu'il permette de les visualiser toutes ensemble, ce qui transforme ainsi des groupes différents de données en une unique source de connaissance. Le format est donc partie intégrante des conditions qui déterminent comment les données sont interprétées, et le choix du format à utiliser est loin d'être un problème solvable de manière partiellement ou intégralement

[20] Les ouvrages Dupré (1983), Kellert *et al.* (2006) et Chang (2012) sont des références classiques pour une introduction au pluralisme.
[21] Par exemple Foucault (1967) et Derrida (1995).

automatisée, ou fondée sur des solutions universellement applicables. De manière générale, le formatage des données nécessite de prendre des décisions réfléchies, fondées sur une bonne connaissance théorique des formats techniquement possibles, mais aussi sur une familiarité avec le domaine et les phénomènes en question. Ceux qui gèrent les bases de données (curateurs, informaticiens, archivistes, scientifiques des données) ont souvent pour mission de prendre ces décisions, qui visent à évaluer les conséquences scientifiques des formats possibles, et à vérifier l'incidence d'éventuels changements sur les moyens dont les données peuvent être agrégées et analysées.

L'évaluation et le choix de classement et de format concernant les données sont des opérations fondamentales, que ce soit pour le fonctionnement interne de chaque banque de données, que pour la façon dont chacune d'entre elles se lie aux myriades d'autres infrastructures chargées de données potentiellement utiles pour la comparaison et la mise en relation avec nos données originales. Les infrastructures utilisées pour gérer et mobiliser les données doivent être le plus facile possible à coordonner et à relier entre elles (une propriété un peu vague que les chercheurs appellent l'*interopérabilité*[22]), afin de permettre aux données d'entrer et de faire partie du *network*[23] global qu'est l'univers des Big Data. Une telle ambition est en réalité un énorme défi logistique et scientifique, étant donnée l'écologie complexe de ces infrastructures, qui évoluent énormément selon l'objectif, le fonctionnement, les financements, les perspectives futures et la position géographique.

Les archives – qui dans le cas des Big et des Open Data prennent souvent la forme d'une banque de données digitales rendue disponible à la consultation par un vaste public *via* des publications sur internet – sont un lieu de grand pouvoir dans l'épistémologie des Big Data. La façon dont les banques de données sont structurées détermine qui peut les utiliser, où, et à quelles fins ; quelles données sont accessibles et lesquelles ne le sont pas ; et quel type d'interprétation il est possible de leur donner. Ce pouvoir, que les sociologues de la science décrivent comme le *pouvoir des infrastructures*[24], est exercé par les experts impliqués dans le développement et la manutention des banques de données, qui à leur tour se divisent en différentes catégories selon les types de données et le domaine en question. Du point de vue informatique, nous trouvons immanquablement des experts en technologie de l'information, en informatique, et eningénierie électronique, qui développent le *software* et le *hardware* nécessaire au fonctionnement des banques de données. Mais il y a également des experts en techniques de stockage et de catalogage ; des experts dans la façon dont les données gérées par une banque de données doivent être formatées, afin d'être compatibles avec les données gérées par une autre banque de données au niveau local et international ; et des experts dans le domaine d'application des données elles-mêmes, capables de gérer, de catégoriser et de visualiser les données, de manière à ce qu'elles soient intelligibles et exploitables pour des utilisateurs qui travaillent dans des domaines spécifiques et différents entre eux (le médical et l'environnemental, par exemple[25]).

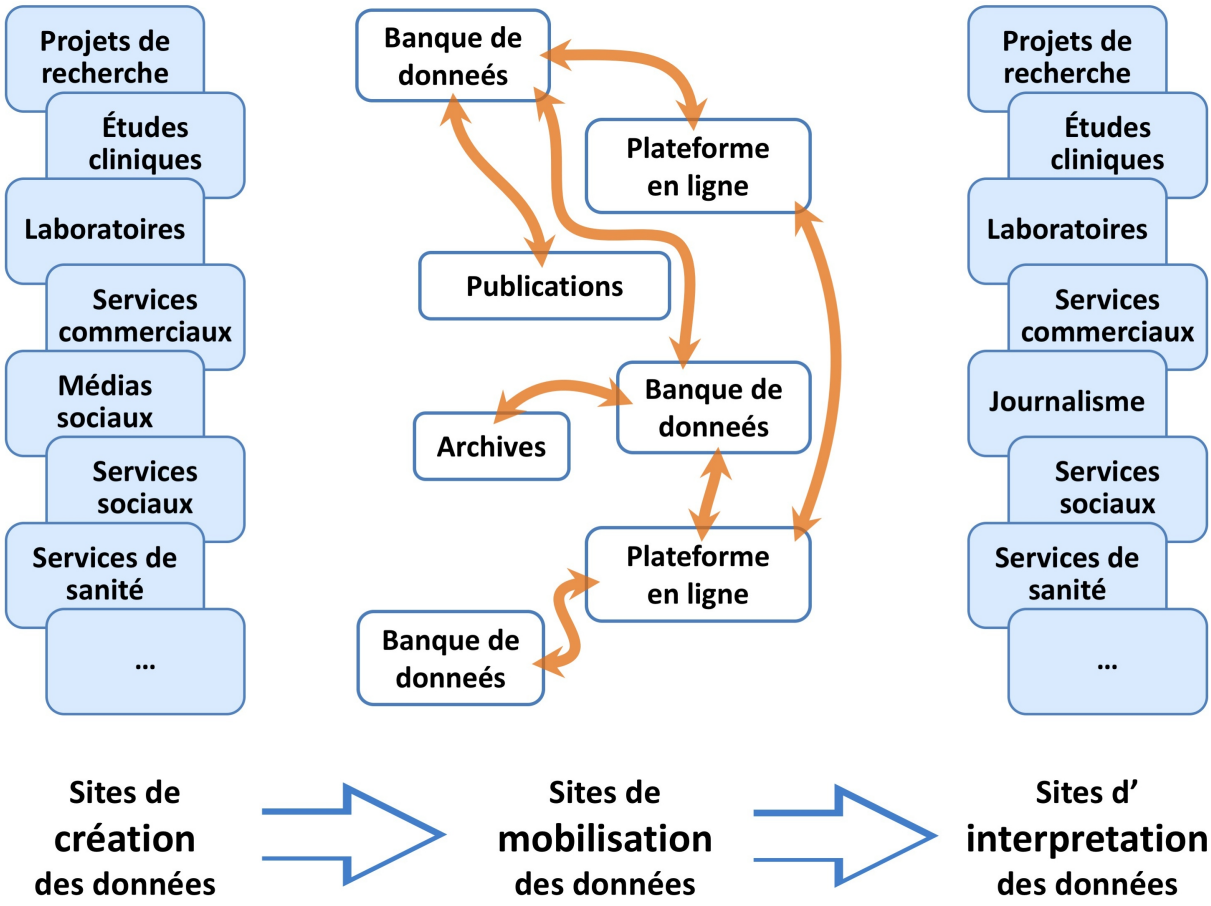
[22] Voir par exemple Sansone *et al.* (2012).
[23] « réseau », en anglais dans le texte. [NDT]
[24] Par exemple Bowker (1994), Hine (2006), Wouters *et al.* (2013), Mongili et Pellegrino (2014), Ribes et Polk (2015).
[25] Sur les tensions qui caractérisent la communication entre ces groupes, voir Edwards *et al.* (2011).

Suivre la circulation des données

La façon dont les données sont transportées et rendues réutilisables est donc fortement répartie. Pour commencer, il est difficile de trouver, pour chaque banque de données, des situations où un unique individu comprend tous les aspects et les types d'*expertise* impliqués dans la mobilisation des données. Beaucoup plus souvent, la compréhension des choix et des techniques à travers lesquels les données sont formatées et mobilisées est répartie entre différents individus, chacun ayant des obligations et des perspectives différentes, et pas toujours nécessairement en contact les uns avec les autres ni même capables de se comprendre mutuellement. Les choses se compliquent ultérieurement une fois que les données circulent d'une banque à une autre, et forment ainsi des liens entre des infrastructures aux origines et aux objectifs différents, qui utilisent chacune des présupposés et des critères très hétérogènes pour sélectionner, classer, formater et visualiser les données. Dans la circulation d'un site à l'autre, d'une situation de recherche à l'autre, et d'une archive à l'autre, les données elles-mêmes se transforment, soit dans leur forme, soit dans le contenu – un fait inévitable et nécessaire au transport des données d'un contexte à l'autre et à leur réutilisation dans le cadre des Big Data, mais souvent oublié par ceux qui considèrent les données comme les représentations immuables et fidèles de la réalité (je reviendrai sur ce point, très important du point de vue épistémologique, au chapitre trois).

La complexité et l'importance épistémique et scientifique de cette situation m'ont menée à passer de nombreuses années à littéralement « suivre les données » : c'est-à-dire à tenter d'étudier la façon dont les données circulent exactement, dans quelles conditions et avec quels résultats et conséquences pour les chercheurs impliqués, la connaissance produite et les secteurs sociaux qui sont influencés par cette connaissance (figure 2). Cette enquête est rendue nécessaire par le manque de traces du passage des données, dû à l'habitude des chercheurs de ne pas citer correctement les banques de données qu'ils utilisent pour leurs travaux. Cela rend très difficile de suivre la façon dont les données produites sur un site spécifique sont ensuite absorbées par une ou plusieurs banques de données et autres formes de mobilisation, et de là sont ensuite trouvées par les chercheurs et les analystes intéressés par leur interprétation.

Figure 2. Représentation graphique des circulations des données, afin d'illustrer deux types de mouvements : des sites de production aux sites de mobilisation, et de là aux sites d'interprétation. Il est important de noter comment les sites impliqués dans les trois stades de circulation (dont sont mentionnés seulement quelques exemples dans la figure) peuvent varier énormément dans l'espace ou le temps : les données scientifiques circulent régulièrement d'un continent à l'autre, d'une manière marquée par des interruptions et des arrêts prévus, ou non (les données archivées depuis longtemps peuvent par exemple être redécouvertes par hasard, ou bien demeurer enfouies dans un magazine sans jamais revoir la lumière du jour). (copyright Sabina Leonelli, réalisation Michel Durinx).

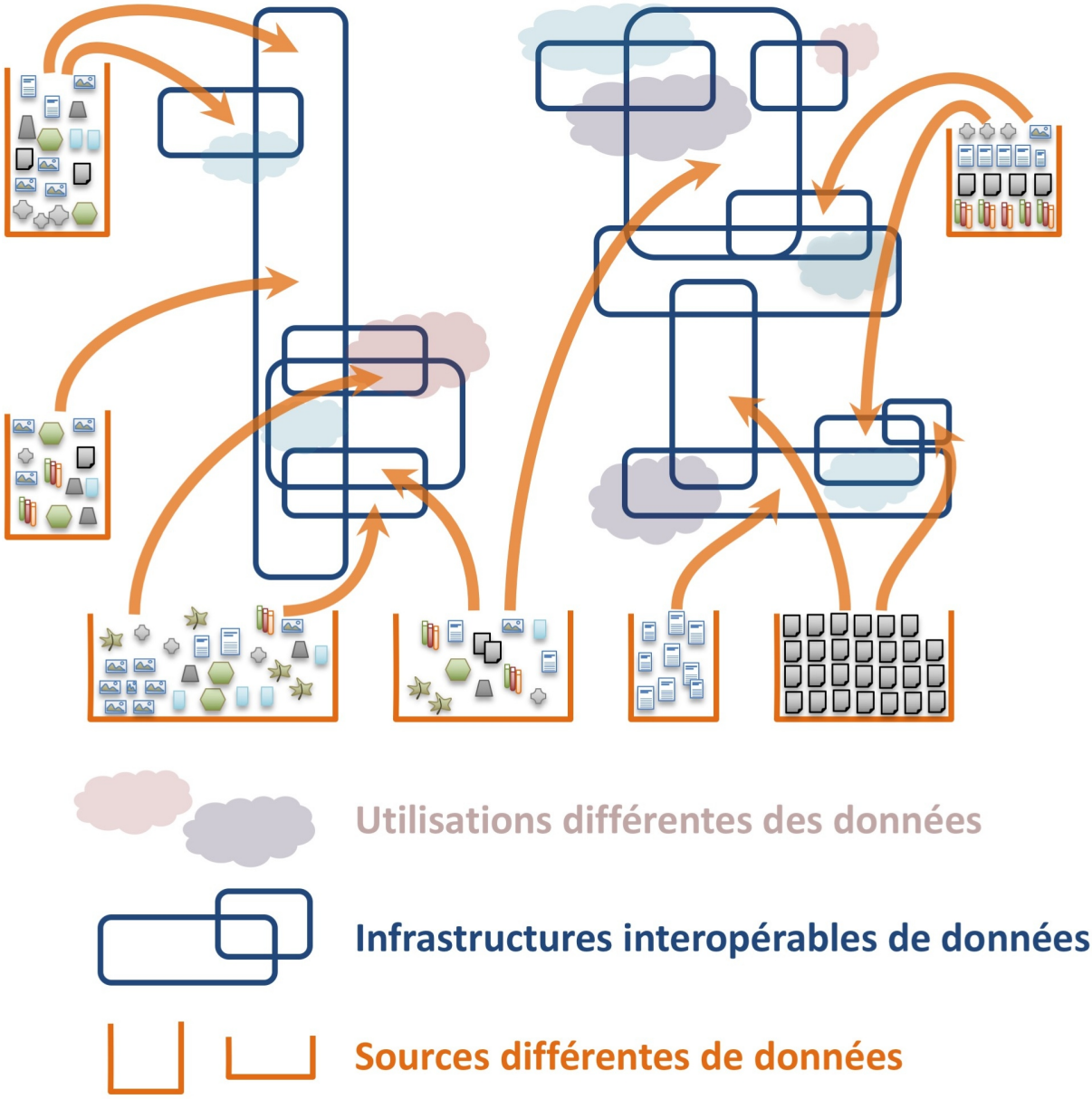


Mes études débutent généralement par une analyse de l'histoire, la structure et l'organisation de grosses banques de données utilisées pour conserver et mobiliser les données dans le monde de la recherche. La reconstruction des raisons ayant mené aux choix effectués pour chaque banque de données, et des moyens avec lesquels les données sont gérées, est toujours très difficile, étant donné que ces choix sont effectués à des moments différents par des personnes différentes, et répondent souvent à des situations politiques, scientifiques et économiques hétérogènes qui ne sont pas documentées de manière systématique. Les meilleures banques de données sont celles qui sont structurées, avec une vision claire de l'incidence qu'elles auront sur la future interprétation des données, et de la façon dont la perception de cette incidence dépend de conditions/critères et d'intérêts spécifiques. Toutes les banques de données ne se fondent pas sur une vision claire de ses objectifs et stratégies de gestion des données. Beaucoup ont été mises en place simplement pour avoir un lieu où conserver les données produites par les projets ou les services, sans une motivation précise ou une rationalisation des choix effectués pour l'organisation des données. Cette situation génère d'énormes problèmes quand le volume des données croit, et devient toujours difficile à organiser pour qu'elles demeurent facilement accessibles.

Cependant, partir d'une idée bien définie n'est pas non plus suffisant pour créer une banque de données efficace et respectée par ses utilisateurs. Les types de public auxquels la banque de données se réfère changent souvent au cours des ans, ainsi que leurs besoins et la façon dont ils espèrent pouvoir consulter les données – sans parler des continuels changements de technologies et de *software* utilisés pour maintenir les archives digitales. Le personnel des

banques de données change également, souvent sans grande continuité ni conscience de la façon dont les choix effectués en amont conditionnent les décisions prises dix ou vingt ans plus tard. En conséquence, les moyens dont les données sont extraites et analysées par la même banque de données tendent à se diversifier toujours plus au fur et à mesure que passe le temps – une situation que j'ai documentée dans divers cas où j'ai suivi les données, à travers la façon dont elles sont gérées dans les banques de données, ou bien celle dont elles sont exploitées par les chercheurs qui utilisent ces infrastructures. Établir le lien entre la structure d'une archive et la façon dont les données sont utilisées n'est pas une opération qui peut se résoudre une fois pour toutes : ce lien doit au contraire être problématisé et régulièrement mis à jour, selon la façon dont le monde autour des archives continue à évoluer.

Figure 3. Représentation graphique des circulations de données, afin d'illustrer la complexité des interdépendances entre les différentes sources de données, les infrastructures impliquées dans leur transport et les portions de données limitées qui sont alors utilisées pour faciliter des découvertes déterminées. Bien que compliquée, cette représentation statique et abstraite ne tient pas compte de la complexité ultérieure de mettre à jour et de corréler ces ressources au fil du temps (copyright Sabina Leonelli, réalisation Michel Durinx).



Reproduire les circulations des données à travers des systèmes de gestion et des analyses aussi complexes, caractérisés par de nombreuses relations de dépendance aux réalités sociales, biologiques et technologiques en perpétuel changement, est une façon de souligner la nature technique et peu transparente de ces processus, et l'impact considérable et dramatique qu'ils ont sur la recherche data-centrée et ses résultats. Les difficultés inhérentes à la reproduction des circulations de données expliquent au moins en partie le fait que le rôle fondamental des archives est souvent sous-estimé par ceux qui considèrent les Big Data comme une opportunité d'accélérer le progrès scientifique, tout en le rendant moins coûteux – ce qui inclue de nombreux gouvernements européens, qui voient l'avènement des Big Data comme un moyen de réduire les dépenses dédiées à la recherche sans pour autant nuire à l'innovation. Mes recherches ont montré que cela est en réalité une illusion dangereuse, aux conséquences potentiellement dramatiques, que ce soit pour le monde de la recherche ou pour la société dans son ensemble. La manutention des banques de données qui permettent la mobilisation et l'interprétation des données nécessite des investissements considérables, l'implantation de

stratégies de participation *sociale* et de contrôles et règlementations adaptés, ainsi que le recrutement de nouveaux types d'experts spécialisés dans les divers aspects de la gestion des données. Par manque de ces investissements, l'utilisation des Big et des Open Data risque de se transformer en un désastre aux conséquences sévères pour la fiabilité et l'impact social de la connaissance ainsi produite. Le prochain chapitre explore cinq types de risque concrètement associés à la mauvaise gestion des Big Data.

Chapitre 2. Signaux d'alarme : cinq façons dont les données nuisent à la recherche

Conservatisme : le problème des données anciennes

Bien que l'utilisation des Big Data soit souvent vue comme une source d'innovation, il y a de bonnes raisons de penser que se fier aux Big Data renforce au contraire le conservatisme des processus et des résultats de recherche. C'est avant tout la conséquence de la complexité croissante des banques de données et des choix respectifs de *standard* et d'algorithmes utilisés pour sélectionner, classifier et visualiser les données. Cette complexité rend toujours plus difficile et plus coûteux le maintien des banques de données de manière à ce qu'elles reflètent les développements informatiques et scientifiques les plus récents, et crée une incitation considérable à maintenir des catégories prédéterminées – une attitude dogmatique qui ne convient pas à la recherche.

Comme on l'a déjà vu au premier chapitre, la façon dont les Big Data sont classées et formatées a un effet déterminant pour la façon dont elles sont ensuite analysées et interprétées. Une des explications à cet effet est que les catégories utilisées pour classer les données ont inévitablement une valeur sémantique. En d'autres termes, ces catégories reflètent les suppositions conceptuelles et méthodologiques des créateurs, gestionnaires et/ou utilisateurs des données elles-mêmes, qui à leur tour résultent de leur manière de concevoir le monde. Au cours de mes travaux philosophiques, j'ai analysé la valeur de ce cadre conceptuel, argumentant qu'il devait être perçu comme une forme de théorie – que j'appelle *théorie classificatrice*. Un des meilleurs exemples de ce type de théorie est ce qu'on appelle les « bio-ontologies » : un système d'organisation des données extrêmement populaire dans les sciences de la vie, qui se fonde sur l'identification et la définition précise des phénomènes auxquels les données peuvent se référer, et sur les relations entre eux[26]. Des théories comme les bio-ontologies sont très utiles – beaucoup diraient indispensables – pour l'organisation des données, mais extrêmement difficiles à modifier et à actualiser.

Dans une étude publiée en 2011, avec quelques-uns des acteurs de la bio-ontologie la plus utilisée dans la biologie contemporaine (la Gene Ontology, inventée pour organiser les données génétiques), j'ai étudié divers cas où ce système était adapté à de nouveaux développements scientifiques[27]. Un exemple est la réélaboration liée à la découverte que le cytosquelette[28], traditionnellement considéré comme un organe externe au noyau cellulaire, contient en réalité des fragments de noyau. Cette découverte a contraint les curateurs de la Gene Ontology à changer non seulement leur définition du cytosquelette, mais aussi les façons dont ce terme était lié à d'autres termes qui décrivent l'anatomie de la cellule. Ces changements ont eu un impact significatif sur les relations entre les données associées à ces catégories ; et le résultat est qu'un biologiste qui cherche quelles données sont associées au cytosquelette, aurait trouvé en 2013 des résultats différents de ceux de 2011. Le cas le plus révélateur est la catégorie même de « gêne », qui est notoirement ambiguë et qui désigne des éléments différents selon la tradition intellectuelle et la perspective théorique adoptée par celui qui l'utilise. Il y a ceux qui pensent aux gênes comme au code de la vie, partie fondamentale de la reproduction ; et ceux qui le

[26] Leonelli 2012 ; 2016a.
[27] Leonelli *et al.*
[28] Ensemble de polymères biologiques qui confèrent à la cellule son architecture et sa mécanique [NDT].

voient en revanche comme l'un des nombreux composants (notamment les stimuli environnementaux et la structure de l'intégralité du tissu cellulaire) qui interagissent dans le développement de l'organisme[29]. Selon le point de vue, l'importance donnée aux données génétiques par rapport aux données cellulaires change considérablement, et la théorie classificatrice utilisée pour classer ces données reflète fidèlement ces préférences conceptuelles.

Donc s'il y a une chose que le data-centrisme n'est justement pas, c'est la « mort de la théorie[30] ». Non seulement les données sont toujours produites en lien avec des attentes précises et des conjectures conceptuelles, mais les mêmes stratégies choisies pour mobiliser les données ont toujours un sens précis du point de vue théorique, dont les fondements empiriques et les implications doivent être explicites et réévaluées régulièrement pour contrôler qu'elles sont encore adaptées – une tâche rendue encore plus urgente à cause de la continuelle accélération de la production de savoir scientifique liée au nombre croissant d'investissements dans la recherche au niveau mondial. Malgré cela, à l'intérieur de chaque banque de données – particulièrement celles qui fonctionnent depuis plus de dix ans – il est commun de trouver de la confusion et de l'ignorance sur les raisons et les critères utilisés au fil des ans pour choisir, adapter et actualiser les modalités de gestion des données. D'une part, il y a souvent trop de gens concernés pour systématiquement parvenir à suivre qui a fait quels changements, pourquoi et sur quels critères. D'autre part, le rythme de travail et les pressions qui caractérisent la gestion des données n'encourage pas celui qui travaille dans ce domaine à prendre le temps de justifier ses propres choix de manière détaillée. La plupart des infrastructures utilisées pour mobiliser les données se nourrissent de financements à court terme et répondent à l'obligation de démontrer leur utilité à leurs financeurs le plus tôt possible, de manière à avoir l'opportunité de recevoir des fonds ultérieurs. Celui qui gère les données est donc soumis à la forte pression de développer une ressource qui devient fonctionnelle au plus vite. Le délai est court pour évaluer avec calme les avantages et les inconvénients des différents moyens de classer les données, et la sémantique des catégories qui sont adoptées dans ce but, et encore plus pour prendre en compte les variétés d'actions entreprises pour créer et maintenir l'archive. L'employé qui décide de changer de travail, après avoir passé plusieurs années à configurer une banque de données, laisse donc rarement un guide écrit sur la façon et les raisons utilisées jusque-là pour classer et formater les données.

Le résultat est ce que Bruno Latour a appelé une « boîte noire » : une technologie construite sur le fondement de plusieurs suppositions – qui concernent par exemple les caractéristiques des objets sur lesquels les données ont été produites, ou le meilleur format pour mobiliser les données en question – qui deviennent des composantes invisibles et donc incontestables de l'infrastructure[31]. Les philosophes William Wimsatt et James Griesemer, en travaillant en collaboration avec la psychologue Linda Caporeal, ont plus tard affiné l'idée de boîte noire dans leurs travaux sur les « échafaudages »[32] utilisés au cours de l'évolution culturelle, qui est facilement applicable au cas des Big Data. Les échafaudages sont les suppositions conceptuelles, sociales et matérielles nécessaires à la construction de théories, technologies ou

[29] Barnes et Dupré 2008 ; Müller-Wille et Rheinberger 2012.
[30] L'idée que l'analyse des Big Data implique la mort de la théorie, c'est-à-dire l'idée que les critères théoriques utilisés pour produire et gérée les données ne sont pas pertinents pour leur interprétation, a surtout fait beaucoup de bruit suite à un éditorial de Chris Anderson justement appelé « The End of theory », publié dans la revue *Wired* en 2008 (Anderson 2008).
[31] Latour, 1987.
[32] « Scaffolds » en anglais.

infrastructures. Une fois que le processus de construction est établi, les échafaudages finissent habituellement par être enlevés, mais laissent– exactement comme les structures utilisées pour construire des bâtiments – cependant une empreinte importante sur ce qui a été construit. Pour celui qui souhaite intervenir sur la construction ou en vérifier la solidité, il est indispensable de comprendre quelles structures ont été utilisées au cours de la construction. Mais cela signifie retracer pas à pas les processus et les stratégies utilisées pour mettre en place la construction, ce qui est difficile une fois que le produit en question est fini et que la mémoire de ce qui a été fait est progressivement perdue[33].

Dans une situation où les choix effectués lors de la configuration des données deviennent toujours plus difficiles à retracer au fur et à mesure que le temps passe, il devient plus difficile également d'isoler les parties de l'infrastructure qui nécessitent des mises à jour ou qui sont devenues obsolètes en raison de nouveaux développements scientifiques. La situation s'aggrave particulièrement lorsque nous prenons en compte la multitude de banques de données qui peuplent chaque branche de la recherche scientifique, chacune desquelles contenant des suppositions qui influencent la circulation et l'interopérabilité des données, et qui ne sont pourtant pas souvent mises à jour de manière fiable et régulière. Simplement pour donner une idée concrète des chiffres, la prestigieuse revue scientifique *Nucleic Acids Research* publie chaque année une édition spéciale sur les nouvelles banques de données pertinentes pour la biologie moléculaire : les nouvelles infrastructures référencées dans l'édition spéciale étaient 56 en 2015, 62 en 2016, 54 en 2017 et 82 en 2018. Ces dernières ne sont qu'une petite part des centaines de banques de données créées chaque année en lien avec les sciences de la vie, qui elles-mêmes sont liées à la quantité encore plus grande de banques de données utilisées en médecine, sciences de l'environnement et agronomie. Le fait que beaucoup de ces infrastructures ont des fonds à court terme génère un pourcentage croissant de ressources, qui demeurent consultables *online*, mais qui sont mortes depuis longtemps, en ce sens que personne n'est plus impliqué dans leur prise en charge – une situation pas toujours visible par les utilisateurs, qui se fient souvent au contenu des banques de données sans vérifier si elles sont gérées et alimentées de manière attentive. À quel moment ces infrastructures deviennent-elles obsolètes ? Et quels sont les risques inhérents au tissage d'un réseau toujours plus étendu d'infrastructures qui dépendent les unes des autres, étant données les disparités entre les façons dont elles sont gérées et la difficulté à identifier et comparer leurs prérequis et les théories et échafaudages utilisés pour les construire ?

Un de ces risques est justement le conservatisme généralisé qui provient de la tendance au recyclage de données anciennes, dont les caractéristiques et modalités de gestion deviennent toujours plus opaques avec le temps, au lieu d'encourager la production de nouvelles données, dont les caractéristiques répondent de manière spécifique aux exigences de celui qui les utilise et aux circonstances. Dans des domaines comme la biologie et la médecine, qui étudient des organismes vivants, et donc par définition en développement et évolution perpétuels, la confiance croissante dans des données anciennes dont se nourrit l'analyse des Big Data est particulièrement alarmante. Il n'est pas du tout évident, par exemple, que les données récoltées sur des bactéries ou des champignons prélevés en forêt et ramenés en laboratoires il y a dix, vingt ou cent ans sont une source fiable pour expliquer le comportement de la même espèce de bactéries aujourd'hui ou dans le futur – et en réalité, les biologistes qui étudient l'évolution de pathogènes néfastes à l'homme ont d'énormes problèmes lorsqu'ils utilisent les différentes

[33] Caporael *et al.* 2014.

sources de données disponibles actuellement afin de retracer la diffusion géographique d'infections et pathologies au cours des dernières décennies[34]. Cela vaut aussi pour les données recueillies sur des cultures cellulaires *in vitro*, qui sont sujettes à de stricts contrôles, justement pour s'assurer qu'elles restent les mêmes d'une génération à l'autre, et qui, malgré cela, ont pourtant tendance – comme le fait la vie elle-même, selon ce que nous avons tous appris grâce à Jeff Goldblum dans *Jurassic Park* – à échapper aux contrôles et à muter de manière imprévisible[35].

Encourager les chercheurs à retourner continuellement vers une analyse et une réanalyse des données anciennes a sûrement l'avantage de réduire les investissements nécessaires à la création de nouvelles données, un facteur qui est tentant pour beaucoup d'agences qui sponsorisent l'utilisation des Big et des Open Data – particulièrement avec les coûts astronomiques souvent associés à l'acquisition des technologies et des matériaux utilisés pour produire les données, où même une simple expérience biologique peut atteindre des centaines de milliers d'euros. Ce choix peut cependant générer un autre type de coût, bien plus difficile à quantifier, et à l'impact potentiellement désastreux pour la science : le développement de la recherche conservatrice et peu créative, vouée à maximiser ce que nous savons déjà sur le monde naturel, au lieu d'en explorer la perpétuelle évolution de manière toujours renouvelée et de plus en plus sophistiquée.

Manque de fiabilité : le problème des données douteuses

Un autre risque important, qui devient évident lorsqu'on considère les conditions de mobilisation des données, est la difficulté à en évaluer la qualité. Un des problèmes principaux lorsqu'on se fie à une banque de données abandonnée est le manque de garantie que les données qu'on y trouve sont encore crédibles. Encore plus préoccupant est le fait que – abandonnées ou non – les banques de données actuellement en circulation divergent énormément dans leur approche du contrôle de la qualité des données. Certaines infrastructures n'ont aucun mécanisme pour filtrer les données fiables de celles qui ne le sont pas, et elles justifient ce choix par l'idée que leurs utilisateurs ont des intérêts et des critères d'utilisation différents : ce n'est donc pas le rôle de la banque de données de décider quelles données sont acceptables et dans quel but. D'autres infrastructures voient ce contrôle comme une responsabilité fondamentale des banques de données envers leurs utilisateurs. Cette position est fréquente, spécialement dans les domaines où les utilisateurs n'ont ni le temps ni les connaissances nécessaires pour vérifier les données trouvées dans les banques, et qui finissent donc par considérer comme acquis que les données mises sur internet sont systématiquement fiables.

La diversité des approches entre les banques de données est déjà un obstacle considérable à la circulation des données, surtout que les utilisateurs ne sont pas toujours informés et conscients des contrastes entre les critères d'inclusion utilisés par différentes banques de données. Encore plus problématique est le fait que, même quand les gérants des banques de données se préoccupent d'en garantir la fiabilité, il n'existe pas de critères universels pour déterminer quelles données sont « bonnes » ou non. Il est vrai en effet que les jugements sur la qualité et la fiabilité des données comme source de connaissance varient fortement selon les objectifs de

[34] Leonelli, 2018.
[35] Landecker, 2007.

la recherche et la manière dont les données sont utilisées[36]. Par exemple, de nombreux chercheurs en médecine clinique considèrent les données sur l'expression génétique (ce qu'on appelle les *microarrays*[37]) comme intrinsèquement fiables, parce qu'elles sont produites grâce à des instruments standardisés et sont donc plus facilement comparables entre elles, contrairement aux données physiologiques prélevées sur les patients. J'ai interrogé de nombreux biologistes qui considèrent au contraire que ces données ne sont pas fiables, parce que les échantillons et les instruments utilisés pour les produire sont extrêmement sensibles aux changements thermiques et lumineux des laboratoires dans lesquelles elles sont produites[38]. La conclusion de ce discours est que les *microarrays* sont souvent incorporés à des banques de données biomédicales mais pas à celles de données biologiques – une situation paradoxale étant donné que de nombreux chercheurs, qui travaillent par exemple en oncologie, utilisent ces deux types de banques de données comme sources pour leurs études.

En général, l'adoption de formats et de stratégies de classement uniformisés aide à rendre les choix de gestion des données plus clairs et plus facilement modifiables, ce qui facilite également les contrôles de qualité et de fiabilité des données elles-mêmes. Le pluralisme qui caractérise le travail scientifique, et donc la diversité des méthodes utilisées par les chercheurs pour étudier la nature, rendent cependant impossible d'adopter des *standards* universels – et même quand des *standards* existent, de grands efforts sont dédiés à les adapter aux conditions spécifiques de la recherche en question[39]. L'applicabilité des *standards* dépend de la capacité de ceux qui les utilisent à mettre les données en lien avec des situations de recherche déterminées. Il existe certaines méthodes statistiques pour contrôler divers aspects des données, de la façon dont elles sont agrégées à leur homogénéité. La statistique n'aide pourtant pas à vérifier si les méthodes utilisées ont du sens, selon les propriétés des objets d'étude en question. Par exemple, un projet sur le rythme circadien des plantes – les différentes façons dont leur métabolisme fonctionne durant les vingt-quatre heures d'une journée – peut produire des données parfaites du point de vue technique, mais qui peuvent être problématiques par rapport aux gênes qui sont analysés, à la fréquence à laquelle les données sont générées, et aux conditions dans lesquelles les plantes elles-mêmes ont grandi.

L'évaluation de la qualité des données dépend non seulement de l'utilisation, mais aussi de la connaissance que les chercheurs ont des conditions dans lesquelles les données ont été produites. Celui qui possède une expérience des instruments et des matériaux utilisés pour générer les données dispose d'une connaissance intime des contrôles et des méthodes utilisés, ce qui joue un rôle significatif pour juger la fiabilité des données. Une fois que les données circulent dans des contextes différents de ceux dans lesquels elles ont été produites, cette connaissance intime vient souvent à manquer, ce qui crée des obstacles ultérieurs dans l'évaluation de la qualité des données et de la crédibilité des interprétations particulières. En outre, comme toutes les activités humaines, la production de données peut également être mal faite, ou avec des motivations qui ont peu de rapport avec la recherche de la vérité. Sans critères clairs avec lesquels évaluer les données, il est impossible d'identifier les données qui ne sont pas fiables ou les produits conçus pour créer une vision déformée de la réalité. Il suffit de penser

[36] L'ouvrage sur la qualité de l'information publié par Phyllis Illari et Luciano Floridi (Floridi et Illari, 2014) contient de nombreux exemples et réflexions sur les implications de ce phénomène ; on peut consulter également Edwards (2010) et Leonelli (2017b).
[37] « Puces à ADN », en anglais dans le texte [NDT].
[38] Leonelli, 2012.
[39] Bowker et Satr, 1999 ; Timmermans et Epstein, 2010.

aux efforts réalisés par l'industrie du tabac au cours des cinquante dernières années, bien étudiés par les collègues Naomi Oreskes et Brian Conway, pour produire des données qui prouveraient que fumer est bénéfique pour la santé[40]. Le risque à collectionner les Big Data est de mettre ensemble de grandes quantités de données sans prêter attention aux différents degrés éventuels de fiabilité comme source de connaissance, et donc de construire un énorme château de cartes.

Ce problème est clairement reconnu et largement discuté par les gérants des banques de données et les chercheurs qui les utilisent. Tous s'accordent sur la solution. D'une part, les banques de données doivent faire leur possible pour vérifier les objectifs et la rigueur des méthodes employées pour produire les données, et acquérir autant d'informations que possible sur les circonstances de production des données, de manière à pouvoir fournir à ses utilisateurs la représentation la plus complète possible de l'histoire des données. Les banques de données sont donc responsables de la *décontextualisation* des données : le choix de quelles données inclure, et surtout de quelles informations donner aux utilisateurs sur leur provenance (métadonnées) sont deux facteurs fondamentaux pour la mobilisation des données à l'extérieur de leurs sites de provenance. D'autre part, étant donné que les critères de qualité varient selon la situation d'utilisation, chaque utilisateur qui s'apprête à créer de nouvelles interprétations des données doit prendre la responsabilité d'en étudier la fiabilité par rapport à son environnement de recherche. C'est cela la *recontextualisation* des données : les utilisateurs exercent leur propre jugement, fondé sur leurs connaissances et leurs capacités, pour formuler leur propre interprétation de la signification des données et de leur crédibilité[41]. Les banques de données fournissent un support indispensable au processus de recontextualisation grâce à la sélection de métadonnées, qui décrivent de manière appropriée les choix effectués dans la production et l'organisation des données, ce qui donne donc aux usagers la possibilité d'évaluer eux-mêmes la signification et l'impact de celles-ci[42].

La différence entre les connaissances et les intérêts de ceux qui produisent les données et de ceux qui les réinterprètent pour de nouveaux objectifs est à l'origine du problème de la qualité des données, puisque ces deux groupes de chercheurs divergent souvent dans les critères qu'ils utilisent pour leur évaluation. En même temps, cette différence est aussi la raison pour laquelle l'analyse des Big Data peut mener à de nouvelles découvertes : c'est justement le fait que celui qui réutilise les données tend à les évaluer avec un regard neuf – et des compétences et intérêts nouveaux – qui génère la possibilité d'interprétations innovantes et différentes de celles proposées par le passé. Cette observation est fondamentale pour comprendre l'indissoluble lien entre l'analyse des Big Data et le risque qu'il y a à se fier aux données douteuses. Le risque de non-fiabilité fait partie intégrante de la manière dont les Big Data sont mobilisées et analysées, et la science data-centrée ne pourrait exister sans ce risque.

---

[40] Oreskes et Conway, 2010.

[41] Un facteur qu'il faut mentionner, même si la place manque ici pour en parler de manière approfondie, c'est l'importance que l'accès aux matériaux et instruments originellement utilisés pour produire des données puisse avoir dans leur recontextualisation. Souvent l'unique façon pour un biologiste de comprendre la signification potentielle des données trouvées sur internet est d'évaluer à son tour les mêmes échantillons (qui sont des cultures cellulaires ou des variations génétiques particulières de la même espèce) utilisés par ceux qui ont créé les données. Certaines banques de données, comme celles qui sont liées aux organismes modèles, incluent des informations sur les échantillons originels au sein de leurs fonctionnalités, mais cela requiert des ressources considérables inaccessibles à la plupart des infrastructures (Ossorio, 2011, Leonelli 2016a).

[42] Pour un examen approfondi des processus de décontextualisation et recontextualisation, et leur signification scientifique et philosophique, voir Leonelli (2016a).

Mystification : le problème des données partielles

Le risque qu'il y a à se fier aux données anciennes ou douteuses peut déjà sembler un lourd prix à payer pour la consultation des Big Data, mais ce n'est pas, d'après moi, le plus élevé. Le risque le plus significatif provient plutôt du problème des données partielles : c'est-à-dire du fait que les banques de données fournissent des informations très sélectives, qui représentent donc seulement une petite partie de la réalité, sans pour autant toujours fournir les instruments nécessaires pour analyser les conséquences de cette limite.

La nature sélective des banques de données n'est pas un problème en soi. Chaque étude scientifique simplifie nécessairement la réalité de manière à pouvoir en étudier un aspect spécifique, et cette capacité à concentrer et modeler le monde pièce par pièce est à l'origine du succès de la recherche scientifique. Ce qui me préoccupe n'est donc pas le fait même de réduire la réalité, dans son infinie complexité, à un échantillon de données limitées dans leur portée et leur représentativité. Cette réduction est une composante essentielle de chaque processus d'analyse. Le « risque de mystification » provient en revanche de la tendance des utilisateurs des Big Data à oublier que ce qu'ils manipulent n'est pas un échantillon complet ni particulièrement équilibré de la réalité, mais plutôt une sélection effectuée en partie à cause de limites pratiques et en partie pour des raisons conceptuelles. Ignorer la nature sélective des données mises *online*, et les raisons pour lesquelles ces données, et pas d'autres, sont accessibles et analysables par des algorithmes, aide sûrement à accélérer l'interprétation des données, mais en même temps, cela facilite au contraire la production d'interprétations qui mystifient les faits au lieu d'aider à les comprendre.

Quelles sont les principales sources de distorsion dans la sélection des données présentes dans les archives digitales ? Un coup d'œil rapide aux banques de données les plus utilisées au monde à des fins de recherche révèle immédiatement que la majeure partie des initiatives couronnées de succès a porté sur des données facilement traitables du point de vue informatique, comme par exemple les séquences génétiques, qui sont exprimées par des lettres (A,C,G,T) facilement analysables au moyen d'algorithmes[43]. Construire une banque de données qui collectionne des données difficiles à digitaliser et à analyser, comme les photographies ou les dessins faits à la main par exemple, requiert des investissements bien plus consistants que ceux nécessaires pour une infrastructure dédiée aux données numériques ou symboliques. Certes, il existe des initiatives très intéressantes qui permettent la mobilisation et l'analyse informatique d'images complexes : c'est le cas par exemple des résonances magnétiques utilisées pour révéler les structures anatomiques. Mais ce sont des efforts qui demeurent limités face à l'énorme quantité de ressources dédiées à la circulation de données plus facilement traçables – un facteur qui influence et réduit de manière significative la diversité des données accessibles *online*.

Une autre observation, immédiatement évidente quand on considère le panorama des Big Data, est que les infrastructures qui bénéficient de la meilleure réputation sont celles gérées par des institutions bien financées, riches en ressources humaines, situées dans des lieux de pouvoir du monde scientifique (Boston, Singapour, Cambridge, Oxford, Pechino, San Francisco) et où l'anglais est l'unique langue utilisée en toutes circonstances. Pas seulement : ces institutions et leurs banques de données respectives ont tendance à travailler suivant des traditions de recherche particulièrement en vogue et reconnues par le monde politique et industriel – raison

[43] Pour une étude détaillée de la façon dont a été trouvée cette notation et son impact sur la digitalisation de la biologie moléculaire, voir Rheinberger (2011), November (2012) et Stevens (2013).

pour laquelle elles bénéficient de fonds considérables et d'une visibilité notable dans le monde scientifique, ce qui leur permet de faire connaître leurs efforts au niveau international et à encourager les chercheurs du monde entier à accepter leurs critères d'étude, plutôt que perdre du temps à les contester[44]. Cela génère une énorme distorsion dans la provenance des données rendues accessibles par ces infrastructures ; mais surtout, celui qui crée les banques de données tend à les adapter à ses propres préférences méthodologiques et conceptuelles, ce qui génère ainsi des archives qui contiennent en grande partie ses propres données et celles de ses collègues aux préférences similaires. Il est de plus évident que ceux qui assistent à la création des banques de données en comprennent mieux le fonctionnement, et sont donc plus à même d'utiliser ces infrastructures à leur avantage – que ce soit pour mobiliser leurs propres données ou pour analyser celles des autres.

Ces privilèges ont de lourdes implications pour le niveau et la qualité de la participation au sein des banques de données en question. Une étude que mon groupe de recherche a conduite dans divers laboratoires africains entre 2014 et 2016 montre que celui qui travaille dans ces laboratoires est souvent impressionné par celui qui fait de la recherche dans des laboratoires mieux reconnus au niveau international[45]. Cela signifie que, d'une part, de nombreux chercheurs africains n'osent pas contester le travail des banques de données américaines ou européennes, qu'il soit correct et utile de leur point de vue ou non et, d'autre part, que ces chercheurs réussissent rarement à en faire usage de manière optimale. Une des raisons est l'idée, souvent utilisée par les banques de données, que les usagers ont accès aux versions plus récentes du *software* nécessaire à l'analyse, alors que beaucoup de chercheurs qui ne travaillent pas à Stanford ou au MIT disposent de programmes plus anciens, et n'ont pas toujours accès à une connexion internet suffisamment puissante pour leur permettre de charger et de travailler avec les Big Data[46]. Une autre raison est l'hésitation, exprimée par plusieurs chercheurs que nous avons interrogés, à partager ses propres données au sein d'infrastructures internationales, due à la peur que les chercheurs avec plus de visibilité et des moyens plus puissants qu'eux n'en profitent : par exemple pour produire des analyses plus rapidement que ceux qui ont des instruments et des infrastructures moins sophistiqués[47].

Cette situation ne paraîtra pas surprenante aux lecteurs, étant donnée la complexité des processus de mobilisation des données décrits jusqu'à présent, et la variété de compétences, d'instruments et d'investissements nécessaires pour réussir à les analyser. Il est pourtant important de réfléchir au résultat de ce lien puissant entre le pouvoir (économique et culturel) des sites de mobilisation de données et le type de données qui deviennent disponibles, comme les Big Data. Les données rendues disponibles par des infrastructures digitales, qui constituent la source d'une grande partie des études sur les Big Data, sont extrêmement sélectives et privilégient les résultats des groupes de recherche qui ont du succès, qui travaillent dans un milieu anglophone et qui connaissent un certain confort financier (par rapport à d'autres). Les

[44] Avec Rachel Ankeny, nous avons analysé en détail ces configurations de modalités de recherche et ces structures institutionnelles et administratives, que nous appelons les répertoires de recherche (Ankeny et Leonelli, 2016). Les travaux de Pestre (2003) et de Cambrosio *et al.* (2014) sur les « régimes de connaissance », ainsi que ceux de Daston (1995) et Strasser (2011) sur l'économie morale de la communauté scientifique, sont importants pour comprendre l'impact des institutions sur la recherche.
[45] Bezuidenhout *et al.* (2016, 2017).
[46] Vermeir *et al.* (2018).
[47] Ce facteur est particulièrement évident dans les laboratoires moins équipés (Bezuidenhout *et al.* 2017), mais est aussi et paradoxalement cité par des chercheurs qui travaillent dans les meilleurs laboratoires du monde (Fecher *et al.* 2015, Levin *et al.* 2016, Levin et Leonelli 2016).

données produites par des groupes de recherche sur des sites moins visibles mais bien équipés sont difficilement incluses, et les conditions d'inclusion, dans les rares cas où elle advient, sont dictées par l'élite universitaire. Cela génère une énorme disparité soit (1) dans les sources et les types de données qui peuvent être analysées comme des Big Data (que nous pouvons appeler *disparité de représentation*) soit (2) dans la possibilité pour les chercheurs du monde entier – sans parler des citoyens dont la vie est profondément influencée par la façon dont leurs comportements sont datifiés et donc rendus traitables pour être analysés scientifiquement – de critiquer les instruments, les *standards* et les infrastructures utilisés pour mobiliser les données (que nous appellerons *disparité de participation*). Ces deux formes de disparité soulignent et empirent considérablement l'inégalité qui existe déjà entre les traditions de recherche et la gestion du rapport entre science et société au niveau mondial, d'une manière qui génère une distorsion préoccupante et une mystification potentielle de la réalité représentée par les Big Data.

Un exemple flagrant de la disparité des représentations est le fait que le groupe le plus étudié par la majeure partie des données biomédicales en circulation est celui des classes moyennes supérieures de la population de nations fortement développées, et surtout celles d'origine caucasienne et de genre masculin. Ce n'est certes pas nouveau dans l'histoire des sciences, mais ce qui devient particulièrement préoccupant à l'époque des Big Data est la facilité avec laquelle ce problème d'échantillonnage et de représentation est mis de côté. Des disciplines comme la médecine, la sociologie et l'épidémiologie ont passé les deux derniers siècles à développer des méthodes extrêmement sophistiquées pour identifier les sources de distorsion et de discrimination dans leurs données, ainsi que les manières dont la recherche est menée et interprétée afin d'en tenir compte. Le choix du secteur à étudier, et donc du type de données qu'on cherche à obtenir, est un aspect fondamental du travail scientifique et de la façon dont les chercheurs justifient les découvertes qu'ils obtiennent de leurs études. Cette réflexivité et cette capacité à s'ouvrir à la critique, fondamentale pour la crédibilité de la science elle-même, ne sont pas toujours respectées dans le domaine des Big Data. Certains pensent même que l'accès aux Big Data rend inutile toute réflexion sur l'échantillonnage et sur la représentativité des données qui sont analysées : si l'on accepte l'idée que les Big Data fournissent des informations sur *tout*, on accepte aussi l'idée qu'il suffit de les mettre ensemble pour obtenir une plateforme empirique fiable et bien équilibrée – et donc incontestable – pour la recherche future[48].

Au contraire, je soutiens ici l'idée que les Big Data fournissent des informations sur *très peu* de choses et d'une manière qui tend à exclure, ou tout au moins à rendre plus difficile tout type d'opposition constructive. L'idée que les Big Data contiennent une représentation complète de la réalité est une illusion qui détruit l'esprit critique avec lequel les chercheurs affrontent l'analyse et l'interprétation des données empiriques. Le risque réside dans la distorsion ou dans l'occlusion des raisons pour lesquelles les chercheurs sélectionnent des données pertinentes pour leur recherche. À cause de la partialité inhérente aux sources et aux types de données qui sont mobilisées *online*, les chercheurs qui travaillent sur les Big Data se retrouvent souvent à travailler sur des échantillons choisis non pour des raisons scientifiques mais pour des raisons de pure convenance – économique, politique ou culturelle – qui ne sont généralement pas clairement signalées par les banques de données, et dont le potentiel de distorsion n'est donc pas pris en compte dans l'analyse des données elles-mêmes. Cette situation reflète un

[48] Par exemple dans Mayer-Scönberger et Cukier (2013).

phénomène social bien plus vaste, c'est-à-dire le monopole croissant des entreprises aux grandes ressources financières et technologiques – Google en particulier – sur le développement des instruments de gestion et d'analyse des données. La conséquence immédiate est le rôle toujours plus passif joué par le reste de la société pour déterminer quelles données comptent, pour quoi, et comment elles sont utilisées.

Je n'ai pas l'intention de nier ici que l'avènement des Big Data a amené de réels changements dans la manière dont se fait la recherche, mais plutôt de contester l'optimisme avec lequel certains analystes ont décrit l'impact des technologies sur la géographie et le caractère inclusif de la science. Les historiens Bruno Strasser et Paul Edwards par exemple, ont souligné comment les Big Data ont donné l'opportunité à des lieux comme Singapour ou la Chine de défier l'hégémonie scientifique occidentale[49]. Je suis plus pessimiste : je vois les développements en Orient comme le résultat d'énormes investissements plutôt que comme celui de l'adoption de la technologie en soi, et l'utilisation des Big Data comme une opportunité de renforcer ultérieurement le rôle de la richesse économique pour déterminer quel type de recherche est considérée comme significative et fiable. Contrairement à la vision qu'on a des Big Data et des Open Data porteuses de démocratie et instigatrices de participation sociale dans la recherche, la manière dont la science est gouvernée et financée ne semble pas être remise en question par les Big Data, mais au contraire, l'inégalité de pouvoir et de visibilité entre les différentes nations et communautés scientifiques continue de croître. La divergence digitale entre celui qui a accès aux données mais a aussi la capacité de les utiliser, et celui qui ne l'a pas, s'élargit – et l'on passe donc d'une situation de *digital divide* à une situation de « *data divide* », ce qui nous amène au problème suivant.

Corruption : le problème des données malhonnêtes

Le triomphe des critères financiers et opportunistes, sur lesquels les scientifiques font leurs choix des données à mettre *online*, révèle une profonde tension dans le monde de la recherche data-centrée. Il existe des efforts très sophistiqués, particulièrement dans le secteur de la recherche publique, voués à stabiliser les méthodes pour recontextualiser les données, en comprendre la provenance et les modalités de gestion et, donc, les interpréter d'une manière qui reflète les circonstances et les limites sous lesquelles elles ont été produites et mobilisées. Grâce au niveau d'ouverture et de transparence de ces initiatives, qui comprennent notamment les groupes de travail de la Research Data Alliance et le travail associé à la construction en cours de l'European Open Science Cloud, il est justement possible de référencer les milliers de moyens ingénieux avec lesquels celui qui gère les banques de données dans la sphère publique cherche à contrebalancer les risques identifiés jusqu'ici, et, s'il ne peut les éviter, il peut au moins les atténuer. Cependant, l'immense majorité des données produites à des fins de recherche (sans compter les données produites dans d'autres domaines, puis absorbées par des banques de manière à permettre de nouvelles découvertes) est toujours générée au sein d'un secteur d'intérêt commercial et souvent privatisé.

Cette tendance a deux conséquences d'une grande pertinence pour le rôle des données dans la recherche scientifique. Tout d'abord, la production et la possession de données dans chaque secteur ont été transformées en marchandises au sens marxiste du terme. Non seulement les données sont traitées comme des biens commerciaux et sont en tant que tels soumis aux lois du

[49] Strasser et Edwards 2018.

libre-échange, mais la vente et l'achat des données et la croissante prise de conscience de leur rôle fondamental dans l'économie capitaliste en ont énormément stimulé la mobilisation par des plateformes digitales[50].

Une portion croissante des ressources nécessaires pour collecter, conserver et analyser les Big Data est donc sous le contrôle d'entités aux intérêts commerciaux avant tout, dans le secteur public (gouvernement) ou privé (entreprises actives dans le milieu de la recherche), avec toujours moins d'opportunités données à qui a moins de pouvoir économique et social pour participer à la construction d'instruments et de stratégies d'analyse et d'interprétation. En d'autres termes, nous assistons présentement à la construction d'une oligarchie (pour ne pas dire un monopole) sur l'information et la production de la connaissance, où la logique d'exclusion joue un rôle plus fort que la dynamique en faveur de l'inclusion associée à l'idée d'Open Data.

La deuxième conséquence de la privatisation des Big Data est que cela rend plus difficile l'ouverture non seulement des données elles-mêmes mais, surtout, des informations sur la façon dont elles ont été produites et gérées pour en faciliter l'interprétation. La commercialisation des données s'apparente plutôt à une ambiguïté sur leur *statut* de bien public ou privé. Prenons le cas des données personnelles, c'est-à-dire des données qui identifient les caractéristiques d'un individu particulier (comme le nom, l'adresse, le numéro de compte courant par exemple). Il n'est pas rare que des corporations comme Facebook, Google et les centaines d'entreprises nées au cours de la dernière décennie, pour faciliter l'acquisition et la vente de données à des fins commerciales, défendent simultanément deux idées apparemment opposées : celle que les données personnelles sont en grande partie publiées puisqu'elles sont facilement accessibles (comme par exemple notre prénom, nom et adresse) – et qui sont donc réutilisables pour n'importe quel but une fois que les personnes en question ont donné leur accord – et celle que les données personnelles sont, si pas toujours privées, au moins *privatisables*, et donc sujettes à être vendues et achetées comme n'importe quel produit. Malgré la contradiction, l'entier modèle financier de ces compagnies se fonde sur l'acquisition et sur la réutilisation des données personnelles de millions de personnes dans le monde entier, de manière à contourner le plus possible les réglementations en vigueur dans les nations où ces utilisateurs résident. La confusion qui règne sur ce que veut dire « posséder » les données est amplement et régulièrement exploitée pour encourager les citadins à céder les droits sur l'utilisation de ces données à une quantité croissante d'entreprises, souvent sur la base d'un accès à des services qui facilitent leur vie de tous les jours (comme des informations sur le trafic, le cinéma le plus proche et les prévisions météorologiques du week-end), mais sans avoir conscience des façons dont ces données peuvent être réutilisées et vendues comme des marchandises sur le marché libéralisé.

Le fait que ce commerce des données – et particulièrement des données personnelles – puisse avoir de lourdes conséquences pour les individus et les communautés est heureusement devenu de plus en plus évident, même pour celui directement impliqué dans ce type de travail. Un des moyens les plus évidents de diffuser ce message est d'introduire des initiatives vouées à réguler le trafic des données et des usages auxquels elles peuvent être soumises. L'exemple le plus progressiste de ce type de réglementation est la législation pour la protection des données

[50] Pour une analyse de ce phénomène et de la façon dont il a évolué au cours des dernières années, voir par exemple Thrift (2005), Beer (2016) et Srnicek (2017).

personnelles (Règlement général de la protection des données, ou RGPD) lancée en 2018 par la Commission Européenne, dont l'intention est précisément de protéger les citoyens de l'abus de leurs données, en problématiser le commerce et la réutilisation, et de trouver les formes les plus sophistiquées de stockage et de mobilisation des données elles-mêmes[51]. Le RGPD suit le chemin tracé par les nombreux rapports préparés par des institutions comme l'Organisation pour la Coopération et le Développement économique (OCDE) et les Nations unies, qui recommandent depuis longtemps l'introduction de réglementations de ce genre, et de mesures pour favoriser le débat et l'éducation publique sur les conséquences de l'utilisation de technologies et de mesures de surveillance toujours plus sophistiquées de la part d'organismes publiques et privés.

Ce qui est moins évident, et moins discuté dans la sphère publique, est à quel point la privatisation des données, ou la réglementation de leurs circulations, a des implications très graves pour le monde de la recherche et pour la connaissance produite. Avant tout, ces formes de contrôle et de mobilisation se traduiront par un moyen de sélectionner quelles données sont mobilisées de manière ouverte et transparente. Les corporations relâchent habituellement les données qu'ils considèrent de moindre valeur commerciale et dont l'interprétation nécessite une aide du secteur public. Cela introduit une ultérieure distorsion des sources et des types de données accessibles *online*, avec des données plus profitables et plus complexes – mais aussi potentiellement plus intéressantes et plus fructueuses comme fondement empirique pour la connaissance – tenues à l'intérieur des archives de celui qui les produit, à l'abri des regards indiscrets et de la possibilité de réutilisation à d'autres fins. Même les moyens dont les citoyens – y compris les chercheurs – sont encouragés à interagir avec les banques de données, et les sites d'interprétation des données, sont radicalement restreints à des formes de participation qui génèrent une future valeur commerciale, comme par exemple l'évaluation des applications digitales (« *rate your app !*[52] »), qui sont utilisées pour améliorer les produits, et donc le rendement économique des produits développées à partir des données. Plusieurs sociologues ont récemment décrit ce type de participation sociale comme une forme d'exploitation, ou bien de travail non rémunéré ; un phénomène que les économistes décrivent comme étant une part essentielle de la *sharing economy*[53].

Ces modalités d'exploitation des données – et de qui les produit et/ou les fournit – représentent une tendance au renforcement de la valorisation économique des données *au détriment de leur valeur scientifique*. Comme on l'a vu dans l'introduction de ce livre, les données ont toujours de nombreux types de valeur à la fois, qu'elle soit scientifique ou affective, commerciale, politique, ou culturelle. Par exemple, Niccolò Tempini a montré comment les données personnelles extraites des *social media* comme Twitter et extraites pour produire de la connaissance médicale sont nécessairement évaluées comme des éléments scientifiques ou bien comme des produits commerciaux, elles qui sont des informations constitutives du sens de l'identité des individus, et des éléments dont le partage est à l'origine de la formation de groupes sociaux[54]. Ces modes de valorisation des données ne sont pas nécessairement en conflit les uns avec les autres, mais des tensions et des divergences émergent souvent et ont un impact décisif sur la façon dont les données circulent et sont interprétées. Dans le cas de la vente et de l'achat

[51] Pour plus de détails sur le RGPD et comment il fonctionne, voir le dernier livre de Curioni (2017).
[52] « Note ton appli ! », en anglais dans le texte [NDT].
[53] Sur l'exploitation inhérente à l'utilisation des Big Data, voir Prainsack (2017), Prainsack et Buyx (2017) et Srnicek (2017).
[54] Tempini (2017) ; voir aussi Harris *et al.* (2016) et Leonelli (2016a).

de données personnelles entre entreprises d'analyse, la valeur des données en tant que produits commerciaux – qui inclue l'évaluation de la vitesse et de l'efficacité avec lesquelles l'accès à certains types de données peut aider à générer de nouveaux produits – est souvent prioritaire sur les questions scientifiques comme, par exemple, la représentativité, la fiabilité et le conservatisme des données et des méthodes utilisées pour les analyser. Dans de nombreux cas, cela peut dégénérer en décisions scientifiquement problématiques ou bien seulement désintéressées de l'étude des conséquences des hypothèses et des procédures utilisées – un manque d'intérêt qui se traduit facilement en ignorance des discriminations, des inégalités et des erreurs potentielles dans les données qui sont prises en considération. Ce type d'ignorance est hautement stratégique et économiquement productif, puisqu'il permet l'utilisation des données sans avoir de scrupules sur les potentielles implications scientifiques et sociales. Dans ce scénario, le jugement sur la qualité des données se limite à un jugement sur leur utilité pour produire rapidement l'analyse ou la prévision requise par le client. Il n'y a pas de mesures incitatives, dans ce système, qui encouragent la considération des implications de ce type d'analyses sur le long terme.

Le risque est donc que la commercialisation des données s'accompagne d'une séparation croissante des données elles-mêmes de leur contexte, sans aucune possibilité de recontextualisation. L'intérêt pour l'histoire des circulations de données, la pluralité de leur valeur affective ou scientifique et le réexamen de leur provenance disparaît à long terme, et est remplacé par la fossilisation croissante de la valeur économique des données – dans un processus parallèle à celui que Marx a appelé, de manière célèbre, « aliénation ». Il est clair que ce type de valorisation des données ouvre la voie à la production, à la gestion et à l'analyse des données à des fins malhonnêtes et tendancieuses. Et ainsi revenons-nous à la question de la *post-vérité* évoquée dans l'introduction de ce livre. Dans des situations où la valeur commerciale attribuée aux données domine largement l'intérêt autour de leur valeur scientifique, il est parfaitement possible d'abandonner complètement la recherche de données véridiques, correctes et dont la manipulation prévoit une représentation fiable de la réalité. Ainsi, des procédures de fabrication de données prolifèrent, avec pour seul but de fournir de la crédibilité à des positions et des hypothèses préétablies et commodes du point de vue politique, commercial et social. Dans ces cas-là, la production des données ne peut modifier ce en quoi l'on croit déjà, parce que les uniques données qui comptent sont celles qui peuvent être utilisées pour soutenir ou renforcer des opinions déjà existantes, ou pour améliorer des produits déjà conçus, indépendamment de leur valeur scientifique ou sociale.

Le scandale qui a éclaté en 2018 concernant l'acquisition et l'utilisation des données personnelles par Facebook à des fins politiques est un parfait exemple de ce mécanisme à l'œuvre. Dans ce cas, les chercheurs impliqués, de l'institut privé Cambridge Analytica, ont été payés par diverses entités politiques, y compris par la campagne en faveur de la sortie du Royaume-Uni de l'Union Européenne pour le référendum de 2016, afin d'analyser les données personnelles présentes sur Facebook, avec pour objectif de générer des méthodes efficaces de persuasion grâce auxquelles les citoyens identifiés comme « vulnérables » pourraient être bombardés de messages qui les inciteraient à voter d'une certaine manière. La véridicité des données elles-mêmes est ainsi jugée uniquement en fonction de l'efficacité de l'intervention sociale et politique que la compagnie est payée pour faire. Des données jugées non pertinentes ou qui ne s'accordent pas à ces préférences sont éliminées de l'analyse, et une fois identifiés les segments de la population sur lesquels intervenir, il y a peu d'intérêt à explorer des contradictions potentielles ou des problèmes dans les données mêmes (ou à les confronter à

d'autres sources d'information) – ce qui génère ainsi une connaissance partielle, non fiable ou corrompue. Dans un cas comme celui-ci, la tendance à évaluer la valeur des données sur des critères commerciaux est en net contraste avec la fonction principale des données, qui est épistémique. Si cela n'est pas pris en compte lorsqu'elles sont produites, disséminées et analysées, on risque non seulement la commercialisation totale de la recherche, mais surtout la production d'interprétations influencées par les intérêts et les objectifs d'entités spécifiques – qu'il s'agissent de grandes corporations ou de groupes d'individus particuliers – d'une manière impossible à contester[55].

Dommages sociaux : le problème des données sensibles

Le cas de Cambridge Analytica révèle également un autre problème, traité en dernier non seulement parce qu'il recouvre tous les autres problèmes analysés jusqu'ici, mais concerne aussi le rôle social de la recherche, de manière plus générale. C'est le problème des données sensibles, c'est-à-dire des données qui peuvent être utilisées pour représenter les caractéristiques d'individus ou de groupes de personnes. L'analyse toujours plus sophistiquée de ces données, et l'opportunité de les lier entre elles offerte par les Big Data, ouvrent la porte à une compréhension toujours meilleure des exigences réelles des citoyens, et donc à des décisions politiques, sociales et environnementales mieux informées et plus efficaces. Parallèlement, une gestion ratée de ces données, ou l'adoption de méthodes ou de buts de recherche problématiques du point de vue éthique et social, peuvent facilement provoquer d'énormes dommages aux personnes impliquées – en les rendant par exemple vulnérables à de la surveillance et à de la manipulation de la part de personnes mal intentionnées, ou en générant de la connaissance douteuse ou partielle sur eux, qui est ensuite utilisée par les services sociaux, commerciaux, médicaux ou d'assurance pour établir quel type d'assistance donner et sous quelles conditions.

Avec de telles promesses d'innovation sociale et technologique faites en lien avec les Big Data, qui varient des voitures sans pilote aux moyens d'optimiser la consommation d'énergie, il est facile de sous-évaluer la gravité et la diversité des problèmes causés par une gestion immorale des données et peu attentive aux implications sociales de leur utilisation. Des lois comme le RGPD et d'autres réglementations dédiées à la protection des données tendent à se concentrer sur les droits des individus, en signalant comment l'efficacité avec laquelle différents types de données reliées entre elles peut générer des risques considérables pour les particuliers. Cela est très certainement vrai, particulièrement étant donnés l'inégalité, la corruption et le manque de fiabilité inhérents à quelques systèmes utilisés actuellement pour produire, mobiliser et interpréter les Big Data. Mais ce qui est peut-être encore plus préoccupant, et moins débattu dans l'espace public, est le risque posé par l'utilisation des Big Data pour des *groupes* de citadins. Envisager de potentiels dommages à la collectivité plutôt qu'à l'individu est extrêmement important, parce que cela permet d'étendre notablement l'ensemble des données définies comme *sensibles*. Des questions éthiques et sociales émergent non seulement par rapport aux données personnelles, mais aussi par rapport aux données qui documentent les caractéristiques d'une localité particulière, et qui fournissent ainsi des indications qui peuvent être utilisées pour justifier des interventions de différents types, qui à leur tour peuvent avoir des effets positifs ou négatifs sur les habitants. Des données sur la santé, la distribution démographique et l'utilisation des transports faite par les habitants d'un quartier déterminé,

[55] Leonelli (2016a) ; Ebeling (2016) ; Sunder Rajan (2017) ; Murphy (2017).

peuvent par exemple être utilisées pour justifier la construction d'un parc ou l'approbation d'un nouvel ensemble de bâtiments. Dans cette optique, des données sur le climat, l'environnement et la biodiversité présentes dans une certaine zone géographique, même sans identifier un individu précis, peuvent avoir des conséquences significatives pour la communauté humaine qui réside dans cette zone – ce qui en fait des données sensibles. C'est justement cette caractéristique – la pertinence de ces types de données dans la production de connaissance sur les habitudes et les préférences des individus – qui en souligne aussi bien le potentiel que les risques.

Un des projets dans lesquels j'ai été impliquée récemment est une tentative (courante dans le domaine de la recherche sur les Big Data) de relier les données médicales aux données climatiques et aux données extraites de *social media* comme Twitter, pour comprendre dans quelle mesure l'apparition des symptômes de l'asthme et d'autres maladies respiratoires saisonnières, signalée par des individus qui s'en plaignent sur Twitter, est associée à des conditions climatiques particulières. L'utilisation de données extraites sur Twitter (comme par exemple « aujourd'hui je n'arrive pas à respirer correctement », ou « ça aurait été une belle balade à vélo sans ce maudit pollen ! ») est particulièrement productive pour ce type de recherche en raison du manque de données qui documentent les types moins sévères d'asthme – qui est dû à son tour au grand nombre de situations où les patients se sentent mal mais ne vont pas nécessairement voir un médecin, et ne laissent donc pas de trace de leur mal-être dans les archives médicales. Ces données sont utilisées pour générer des projections du lieu et du moment où les personnes commencent à souffrir d'asthme, qui sont ensuite testées et confrontées aux données médicales sur le lieu et le moment où se produisent les cas les plus graves, sur le niveau de dépenses publiques pour chaque hôpital liées aux maladies respiratoires et sur le type de végétation auquel les citoyens de ces régions sont exposés (pour vérifier quel type de pollen ou d'herbe est présent et associé aux différents foyers). Cet ensemble de Big Data permet ainsi de produire des explications de la fréquence et des caractéristiques d'épidémies d'asthme, et aussi de leur association au microclimat et à la flore locale, avec l'objectif d'aider la santé publique à gérer les hôpitaux et à décider combien de ressources investir dans le traitement de l'asthme dans les différentes régions, et quand et comment renforcer ou diminuer ces ressources au cours de l'année.

Le projet est un parfait exemple des grandes opportunités offertes par les Big Data et les Open Data : une connaissance détaillée des conditions dans lesquelles l'asthme apparaît, qui peuvent sûrement aider les études médicales dédiées à la prévention et au traitement des maladies respiratoires, et une base factuelle pour organiser et motiver les interventions étatiques et les choix sur la façon d'équiper les hôpitaux à l'avenir. Elles sont considérées universellement comme des avantages pour le bien public. Et pourtant, même dans ce type de recherche, nous trouvons des risques considérables associés à la sélection, à la gestion et à la potentielle interprétation des données. Le choix d'utiliser Twitter, par exemple, est fortement conditionné par le fait que celui-ci est l'un des rares réseaux sociaux qui permet (de manière limitée) la réutilisation de ses données par la recherche – alors que Facebook et Instagram demandent une grosse somme d'argent pour remettre des données sur leurs utilisateurs, et sont donc plus difficiles à utiliser pour un projet de recherche financé par de modestes fonds publics. Malheureusement Twitter est aussi une plateforme aux utilisateurs d'un type plutôt spécifique : la majorité d'entre eux a entre vingt-cinq et quarante-cinq ans, habite en ville plutôt qu'à la campagne, est de classe moyenne-supérieure et possède un bagage culturel supérieur à la moyenne de la population. Les données extraites sur Twitter tendent donc à mal représenter les

groupes habitants en zones rurales et ayant moins d'accès à la santé publique, alors qu'ils sont pourtant les plus exposés aux changements saisonniers associés à l'asthme (comme la saison des pollens et de la tonte de l'herbe dans les champs).

Mon rôle dans ce projet est justement celui d'analyser les conséquences potentielles de cette partialité, et d'identifier les moyens avec lesquels les chercheurs peuvent les prendre en compte au cours de leurs recherches et dans les formes de connaissance qu'ils finissent par produire. Dans ce cas, l'exigence de tenir compte des implications sociales de la nature des données sensibles utilisées dans le projet a créé l'exigence de faire des recherches supplémentaires, sur les utilisateurs de Twitter et sur leur répartition sur le territoire par rapport au reste de la population – ce qui aide à quantifier les limites de ces données dans la représentation de la population dans son ensemble –, et de s'interroger à chaque étape sur les façons dont l'échantillon spécifique fourni par les données Twitter se lie à des parties de la population qui en sont exclues. Ces considérations prennent du temps et peuvent être interprétées comme un affaiblissement des résultats obtenus, spécialement pour le gouvernement qui préfère recevoir des réponses claires à ses propres interrogations plutôt que des réponses nuancées par des mises en garde sur les limitations de la connaissance obtenue et sur la discrimination potentielle inhérente aux données. Les chercheurs du projet auraient certes plus de succès médiatique et financier s'ils proposaient une solution facile, rapide et non ambiguë pour prévoir l'apparition d'une épidémie grâce à l'analyse automatique des Big Data, sans s'embêter à réfléchir aux possibles exceptions ou au pouvoir discriminant de tel instrument envers ceux qui ne sont pas représentés dans les données qui l'alimentent. L'honnêteté avec laquelle les chercheurs du projet signalent les limites de leurs projections – et les manières dont celles-ci doivent être nuancées et corrigées selon la situation – est en revanche grandement préférable du point du vue scientifique et social, justement parce que la connaissance ainsi produite signale de manière explicite les conditions dans lesquelles elle peut être considérée comme fiable.

Comment utiliser les données sensibles est donc un problème épistémique ou éthique, et il est impossible de distinguer les critères utilisés pour produire de la connaissance fiable de ceux utilisés pour s'assurer que les méthodes utilisées ne renforcent pas les discriminations sociales injustes et arbitraires. Le manque d'une séparation claire entre conduite scientifique correcte et conduite éthiquement correcte est particulièrement important dans le cas des Big Data. La réflexion sur les conséquences sociales de l'utilisation de données anciennes, partielles, non fiable**s** et corrompues est toujours inexorablement liée à une évaluation de la valeur éthique des choix effectués lors de la sélection, la gestion et l'interprétation des données. Dans ce sens, la valeur scientifique et la valeur éthique des données ne sont pas seulement nécessairement en conflit mais sont généralement associées l'une à l'autre. Cela n'est pas toujours reconnu par les scientifiques eux-mêmes, dont l'impatience à obtenir des résultats – due à l'énorme pression exercée par les universités et les sponsors – les rend réticents à prendre leur temps pour évaluer les éventuelles conséquences sociales de l'utilisation de certaines données. Même dans le cas du projet sur l'asthme précédemment décrit, l'attitude des scientifiques impliqués pourrait être considérée comme schizophrénique : d'une part ils sont réellement heureux de travailler avec un philosophe, et très intéressés par l'idée de produire des prévisions exactes et fiables du point du vue scientifique ou social, mais d'autre part, ils sont frustrés par le fait que gérer les données de manière éthique – par exemple en recherchant exactement quelles sources de données ils sont autorisés à utiliser, quel type de discrimination elles peuvent contenir et comment elles peuvent être rendues visibles pour ceux qui voudraient ultérieurement reproduire notre

recherche – réclame du temps et des efforts, et donc prolonge inévitablement le processus de recherche et en rend les résultats moins sensationnels.

Cette schizophrénie est malheureusement familière et bien compréhensible pour n'importe quel chercheur, étant donnée la façon dont la recherche académique est financée et évaluée dans de nombreux pays européens : le Royaume-Uni est le premier d'entre eux, mais également la France.[56] Surtout lorsqu'on considère les attentes créées par les prophètes de la « réutilisation facile » des données, les bailleurs de fonds des recherches faites sur les Big Data cherchent à obtenir des résultats très rapidement et d'un grand impact économique, souvent sans tenir compte du temps et des efforts nécessaires pour traiter et étudier les données de manière à en vérifier la fiabilité, la représentativité et l'incidence sociale[57]. Les mécanismes de fiabilité généralement reconnus dans le monde scientifique n'aident pas dans cette situation, puisque la crédibilité et la visibilité des chercheurs sont souvent associées à la quantité de publications produites et à la nature sensationnelle de leurs déclarations, qui attire l'attention des journaux et des bailleurs de fonds. Le potentiel de corruption et l'incitation à reléguer à l'arrière-plan tout scrupule moral sont donc continuellement présents, même si sous différentes formes, que ce soit pour la recherche menée dans le secteur privé ou pour celle menée dans le secteur public – et il est facile de voir comment cela se heurte à l'ambition de produire des fondements solides, fiables et socialement acceptables pour la connaissance.

L'éthique comme partie intégrante de la science

L'éthique est la situation où, ceux qui étudient ce que le philosophe Luciano Floridi appelle l'infosphère – c'est-à-dire la façon dont l'introduction de technologies digitales est en train de changer le monde – nous avertissent du potentiel destructeur de l'utilisation des Big Data et du besoin urgent de concentrer les efforts de gestion et d'utilisation des données de manière active et réfléchie en direction de l'amélioration des conditions humaines[58]. D'après les mots de Floridi, « les TIC[59] ouvrent de grandes opportunités, lesquelles, cependant, impliquent l'énorme responsabilité intellectuelle de comprendre de telles technologies et de les faire fructifier de la manière la plus appropriée[60] ». On doit ajouter à ces avertissements un élément important, spécialement dans le domaine d'utilisation des données par la recherche : il est essentiel que les *questions éthiques et sociales soient vues comme partie intégrante des exigences techniques et scientifiques associées à la gestion et à l'analyse des données*. La gestion éthique des données ne s'obtient pas seulement par la réglementation du commerce, de la recherche et de la gestion des données privées, qui est pourtant un énorme pas en avant, comme l'a montré le RGPD, ni par l'introduction de contrôles sur le financement de la recherche, bien qu'ils soient très importants. Pour garantir que l'utilisation des Big Data soit le plus scientifiquement et socialement avancée possible, il est nécessaire de dépasser la conception de l'éthique comme quelque chose d'externe et d'étranger à la recherche, qui s'occupe des conditions et des conséquences de la science, mais pas des contenus. Questions et implications éthiques doivent

[56] Le plan national pour la Science Ouverte, lancé en 2018, est une étape prometteuse pour résoudre ce problème (Plan national 2018).
[57] Pour une étude approfondie des incitations à l'œuvre dans le monde académique à faire des recherches autour des Big Data et des Open Data, voir le rapport que j'ai écrit pour l'Union Européenne en 2017 (European Commission 2017).
[58] Floridi (2017). Voir aussi le rapport sur la Data Governance préparé par la Royal Society & British Academy en 2017.
[59] Technologies de l'information et de la communication.
[60] Floridi (2017).

au contraire être soulevées à chaque étape de la recherche faite sur les Big Data, et deviennent donc une composante fondamentale de l'éducation et du travail de ceux qui s'occupent des données et des méthodes utilisées pour les visualiser et les analyser[61]. Jugements et décisions éthiques se dissimulent dans chaque aspect de la gestion des données, y compris ceux qui, à première vue, semblent purement techniques, et donc neutres du point de vue social.

Cela devient particulièrement évident dans des contextes de recherche éloignés des objectifs de valeur publique et collective, qui au contraire suivent des objectifs déterminés par des gouvernements pour des bénéfices politiques à court terme, par des corporations avec de gros intérêts financiers ou par des méthodes d'évaluation scientifique centrées sur la quantité plutôt que sur la qualité des résultats. Mais pas seulement : même la recherche faite au nom du bien public peut être problématique quand elle ne prend pas le temps d'évaluer ce que signifie exactement le « bien public » et pour qui, et ce que sont les conséquences plus globales de l'élaboration de certains types d'analyse des Big Data. Un des problèmes les plus importants dans ce domaine, ainsi que le plus difficile à contrôler et à gérer, et celui qu'on appelle le « double usage » : c'est-à-dire le fait qu'une technologie développée avec de bonnes intentions peut toujours être exploitée de manière problématique du point de vue éthique (comme par exemple dans le cas de l'extraction des données personnelles sur Twitter, qui peuvent être utilisées pour améliorer la vie des groupes et des individus, mais aussi pour en surveiller le comportement et développer des systèmes de surveillance et d'assurance toujours plus prédateurs et instrumentaux[62]). Cela n'est pas un phénomène circoncis aux Big Data : toutes les technologies sont soumises au même problème, puisqu'il n'est jamais possible de contrôler comment un instrument déterminé sera utilisé une fois produit et distribué. Le fait que les technologies, comme les algorithmes et les banques de données, ont toujours un double potentiel est donc une évidence, et cette situation rend les attentes triomphalement positives sur l'utilisation des Big et des Open Data encore plus absurdes.

Il est vrai que les Big Data et les Open Data ont le potentiel de renforcer la participation à la recherche, la progression de la connaissance et l'efficacité des processus d'enquête : mais pour les mêmes raisons, elles ont aussi la possibilité de compromettre – ou même de saboter – la qualité et la fiabilité de la connaissance produite avec des méthodes scientifiques, ce qui abîme ainsi irrémédiablement la perception sociale de la valeur de la science. Il est vital dans ce contexte de trouver des moyens de gérer les données, afin d'encourager le respect des droits et de la dignité humaine, au niveau individuel comme au niveau collectif[63]. L'intégration des Big Data illustre le lien indissoluble entre les problèmes techniques, comme ceux de conserver les données de manière sûre et d'en vérifier la validité, et les problèmes éthiques, comme celui de stabiliser l'incidence potentielle de l'utilisation des Big Data sur les individus et les communautés[64]. La gestion de la confidentialité et de la sécurité des données est déterminante pour l'étude et le traitement des acteurs de la recherche et pour la manière dont les données sont reliées entre elles[65].

[61] Leonelli (2016b).
[62] Rappert et Selgelid (2013).
[63] Vayena et Tasioulas 2016.
[64] Dove *et al.* 2016 ; Mittelstadt et Floridi 2016 ; Leonelli 2016b.
[65] Tempini et Leonelli 2018.

Chapitre 3. Comment éviter le pire : l'approche relationnelle pour l'épistémologie des Big Data

À l'origine de ce livre, il y a l'envie de comprendre ce que signifie aujourd'hui produire de la connaissance aux fondements empiriques, et de quelle manière la science continue à se confronter et à se distinguer d'autres formes de connaissance qui ne se fondent pas sur l'étude empirique de la nature. Immanquablement, cela inclut l'acquisition d'instruments afin de comprendre le rôle que les données jouent pour inspirer, corriger, confirmer ou démentir nos intuitions, et ce que cela implique pour le processus d'extraction et de connaissance des données. Une épistémologie claire des données – Big, Open, ou autres – est donc un élément fondamental pour le développement de stratégies qui optimiseront les processus de recherche data-centrée et les rendront aussi robustes que possibles, pour faire face aux cinq problèmes fondamentaux dont nous avons parlé au chapitre précédent. Ce chapitre offre ainsi une vision de l'épistémologie des données – une représentation philosophique de ce qu'elles sont, de la façon dont elles fournissent des informations et dont elles sont utilisées pour créer de la connaissance – qui aide à identifier et à mieux comprendre les sources de précarité inhérentes au traitement actuel des Big Data. Le prochain chapitre montrera ensuite comment cette base philosophique sert de point de départ à des propositions de modalités d'intervention sur la production, la gestion et l'analyse des Big Data, afin d'atténuer les risques posés par le conservatisme, le manque de fiabilité, la mystification, la corruption et les dommages sociaux décrits au chapitre précédent.

Visions contraires du rôle des données dans les processus de recherche

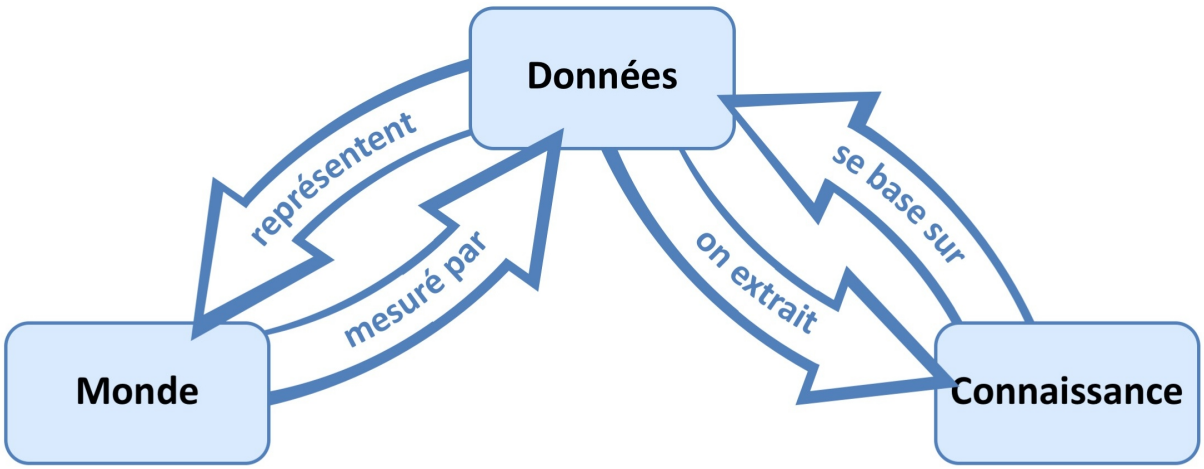
Jusqu'à maintenant, j'ai étudié ce que sont les Big Data, mais je n'ai pas parlé en détail de ce que sont les données au sens plus général : j'ai préféré commencer par discuter des problèmes liés à la gestion et à l'utilisation des Big Data comme sources de connaissance. Maintenant, je veux montrer pourquoi la majeure partie des problèmes observés dans le traitement des Big Data sont dus, au moins en partie, à une certaine conceptualisation des données et la façon dont elles contribuent à la production de connaissance. Cette position philosophique est bien ancrée, mais profondément erronée, et souvent prise de manière inconsciente et implicite : nous l'appelons la « vision représentative » des données.

Selon cette approche, les données sont des représentations fiables de la réalité, produites par l'interaction entre l'homme et le monde. Les interactions par lesquelles les données sont générées peuvent survenir dans n'importe quelle situation sociale, que cela se produise dans un cadre de recherche ou non. Les exemples vont de la biologie qui mesure la circonférence d'une cellule en laboratoire, en notant ensuite le résultat dans un fichier Excel, au maître qui compte le nombre d'élèves de sa classe et l'inscrit sur le registre. Ce qui compte comme une donnée dans ces interactions est l'objet créé au cours du processus de description et de mesure du monde : ces objets, qui peuvent être aussi bien digitaux (les fichiers Excel) que concrets (le registre scolaire), constituent une trace de la réalité qui fournit le point de départ nécessaire pour l'étudier et en tirer de nouvelles formes de connaissance. C'est la raison pour laquelle les données forment un socle pour notre savoir empirique : la production des données est équivalente à la « capture » des caractéristiques du monde afin de les étudier systématiquement. D'après la vision représentative, les données sont donc des objets au contenu fixe et immuable, dont la signification en tant que représentations de la réalité doit être étudiée et progressivement

révélée par le recours correct à des méthodes scientifiques. Les données générées par l'étude de la forme des cellules sont examinées pour améliorer la connaissance de la physiologie et de la structure de l'organisme. Les données créées lorsque l'on compte les élèves d'une classe peuvent être agrégées avec d'autres données similaires récoltées dans d'autres classes et d'autres écoles, ce qui génère ainsi une base empirique pour évaluer la densité des élèves par rapport au territoire et la fréquence à laquelle ils vont à l'école – et ainsi produire une connaissance des caractéristiques et des limites des structures scolaires actuellement utilisées.

Dans la vision représentative, les données sont donc la porte d'entrée pour accéder au monde de manière systématique, contestable et reproductible par d'autres. Cette fonction est souvent présentée en opposition à la connaissance que nous tirons de notre perception sensorielle, fondamentale dans la vie de tous les jours, mais qui peut se révéler illusoire et traîtresse – par exemple quand j'ouvre les yeux dans une pièce sombre et que je ne sais plus déterminer si c'est ma vue qui ne fonctionne plus ou si l'électricité est coupée. Comme d'innombrables philosophes, de tradition analytique (Locke) ou de tradition continentale (Kant), l'ont discuté dans leurs œuvres, la connaissance acquise *via* notre perception n'est pas vérifiable de manière objective, parce qu'elle dépend de notre point de vue et de notre capacité cognitive et sensorielle. Les données sont conceptualisées comme une alternative à ce solipsisme potentiel : c'est-à-dire comme des objets publics qui peuvent être échangés, débattus et critiqués par quiconque, et dont la signification ne dépend pas de la perception d'un unique individu. Dans cette interprétation, les données forment une base objective pour l'acquisition de connaissance, et c'est cette objectivité – la possibilité de tirer du savoir de l'expérience humaine en évitant cependant les limites de la perception subjective – qui rend justement la connaissance *empirique*.

Figure 4. La production de connaissance d'après la vision représentative des données (copyright Sabina Leonelli, réalisation Michel Durinx).



À première vue, cette position peut sembler incontestable. Penser les données comme des représentations objectives de certains aspects de la réalité semble être indispensable à la crédibilité de la science et à l'existence même de l'empirisme comme alternative à la connaissance purement théorique et rationaliste. D'après la vision représentative des données,

les méthodes et les résultats scientifiques sont supérieurs à d'autres formes d'étude de la réalité, car ils sont capables d'étudier le monde de manière fidèle et exacte. Si les experts réussissent à créer des représentations crédibles du réel dans leurs études, en planifiant par exemple des expériences intelligentes, la connaissance qui en découle devient automatiquement fiable. Si l'on suit la même logique, la rhétorique généralement liée aux Big Data est celle de l'accumulation du savoir par induction : accumuler des données obtenues grâce à des méthodes fiables génère une montagne de faits prêts à être analysés, et plus il y a de faits produits et reliés entre eux, plus il y a de connaissance qui peut en découler.

Pourtant, l'analyse des conditions dans lesquelles les Big Data sont concrètement utilisées révèle plusieurs problèmes de la vision représentative du rôle des données dans la recherche[66]. Avant tout, il est clair que tout le monde n'a pas (chercheurs compris) les instruments et les connaissances nécessaires pour tirer du savoir des données, spécialement de celles produites par des laboratoires et des méthodes spécialisées, comme c'est le cas de beaucoup de données scientifiques. Les données ne parlent jamais d'elles-mêmes, et divers types de données impliquent des préparations et des instruments d'analyse différents pour qu'elles puissent être interprétées. La façon dont les données sont interprétées dépend donc, au moins en partie, du type d'habileté et des connaissances de celui qui les étudie[67]. En soi, cela n'est pas forcément un problème pour la vision représentative des données, dont les partisans peuvent simplement répondre qu'il y a des moyens corrects et des moyens incorrects d'interpréter la manière dont les données représentent la réalité, et que celui qui prend la responsabilité de les analyser doit posséder la capacité adéquate pour le faire correctement. Mais qu'est-ce qu'une interprétation « correcte » dans le contexte des Big Data et des Open Data où les données sont mobilisées et réutilisées de mille manières et à des fins toujours différentes ?

Le fait qu'il n'y ait pas qu'une seule interprétation possible d'un ensemble de données est extrêmement difficile à justifier au sein de la vision représentative. Comme on l'a vu au chapitre précédent, c'est justement la possibilité d'utiliser des méthodes, des capacités et des critères interprétatifs différents qui rend la mobilisation et l'agrégation des Big et des Open Data aussi productive et potentiellement révolutionnaire. L'insistance sur l'utilité de recontextualiser les données dans des domaines différents, et donc d'en donner plusieurs interprétations toutes également valables, démontre que l'analyse des données n'est pas un processus purement objectif où la signification des données est progressivement révélé. Les données peuvent être utilisées pour représenter de nombreux aspects différents de la réalité, et la validité de chacune de ces interprétations dépend des circonstances spécifiques du processus d'analyse, y compris de l'habileté à manipuler des hypothèses théoriques qui permettent aux individus et/ou aux algorithmes d'organiser et de visualiser les données afin de conforter une certaine conceptualisation du réel. En d'autres termes, l'interprétation des données est continuellement *filtrée* par le point de vue et par les compétences de celui qui les utilise. La conceptualisation des données comme objet d'une signification immuable et indépendante du contexte ne s'accorde pas avec cette observation fondamentale, et génère même des attentes erronées sur la façon dont les données fournissent des informations sur le monde. À cause de l'approche représentative, beaucoup voient les données comme des faits incontestables et privés d'aspects

[66] Même Floridi a une vision critique de la conceptualisation des données en tant qu'objets au contenu sémantique fixe et indépendant du contexte. Son alternative à la vision représentative, qui est compatible et parallèle à mon approche, mais qui n'est pas focalisée de manière spécifique sur le monde de la recherche, est centrée sur l'étude du concept d'information (Floridi, 2017).
[67] De Regt 2017 ; de Regt *et al.* 2009.

théoriques et subjectifs – une supposition qui ne tient pas compte de l'histoire des données et de leur provenance, ni des circonstances conceptuelles, matérielles et sociales dans lesquelles on peut les interpréter.

Un deuxième problème de la vision représentative des données est d'expliquer comment la gestion des données peut avoir un tel poids sur la façon dont elles sont interprétées. Comme nous l'avons vu, la présentation des données, la manière dont elles sont identifiées, sélectionnées et incluses (ou exclues) dans les banques de données, et les informations fournies aux usagers pour en aider la recontextualisation, sont fondamentales à la production de connaissance et en conditionnent fortement les contenus. De plus, l'analyse de la façon dont les données circulent d'un contexte à l'autre révèle que les attentes et les compétences de celui qui manipule et mobilise les données détermine non seulement la manière dont elles sont traitées, mais détermine aussi ce qui est considéré comme une « donnée » en soi et le format dans lequel elle est disponible – ce qui influence ensuite la possibilité et la manière de les traiter et de les interpréter.

Prenons par exemple le cas des observations botaniques faites par loisir et ensuite utilisées comme données. Faire des photographies et noter des observations sur les plantes rencontrées le long du chemin est le passe-temps préféré de beaucoup de personnes, qui choisissent souvent de conserver les objets ainsi obtenus pour leur unique plaisir personnel – par exemple dans des cahiers annotés conservés à la maison et montrés aux amis et parents lors d'occasions particulières. Celui qui a ce hobby suit rarement des méthodes standardisées pour générer ses photos et ses notes écrites : comme cela a été largement documenté dans l'histoire des sciences botaniques, et malgré le succès des classifications taxonomiques comme celle de Linneo, chacun tend à construire son propre système d'observation selon ses préférences personnelles esthétiques, conceptuelles et affectives, et ses limites physiques et pratiques (une femme de quatre-vingt ans ne montera pas dans un arbre à la recherche d'une variété particulière de rampante ; un parent avec trois enfants en bas âge n'a pas deux heures par jour à consacrer au classement et à l'annotation des photos prises dans le parc). Il arrive souvent que ces amoureux de la nature décident de mettre sur Instagram ou sur un site web quelconque leurs photos et leurs observations, peut-être les plus belles du point de vue technique ou celles qui concernent les plantes et les lieux qui leur sont chers. Il arrive aussi que les scientifiques qui souhaitent documenter la morphologie des plantes d'un lieu particulier découvrent l'existence de ces photos et décident d'en utiliser quelques-unes dans le cadre de leurs recherches[68]. Les photos jugées pertinentes par les scientifiques en question pour leur étude morphologique (un critère très différent de ceux utilisés par les créateurs des photos lorsqu'ils choisissent lesquelles publier) sont donc extraites des sites web utilisés par les amateurs, formatées de manière à être interopérables avec les photos et les observations prises par d'autres dans d'autres lieux et insérées dans une banque de données botanique afin d'en permettre l'analyse. Un ensemble d'objets est ainsi généré, différent par le format et les contenus de celui originellement créé par les amateurs, mais aussi plus facile à comparer et à agréger avec des photos d'autres plantes déjà présentes dans le système. Une fois que les photos sont insérées dans ce contexte, et

[68] Considérer des objets produits sur le temps libre des citoyens ordinaires comme des données scientifiques est une composante fondamentale de la *citizen science* – la science des citoyens – qui fait partie du mouvement vers la Science Ouverte. La création et le partage de données sont une des méthodes les plus en vogue pour impliquer les non-professionnels dans la création de connaissance scientifique, comme l'a également montré l'histoire de l'astronomie et des sciences naturelles – où l'acquisition de données est depuis longtemps organisée parallèlement à une collaboration entre amateurs et chercheurs.

qu'elles deviennent ainsi des données potentiellement utilisables pour la production de connaissance, il est aussi parfaitement possible pour les autres scientifiques, en-dehors de tout intérêt ou spécialisation en morphologie des plantes, de sélectionner quelques photos et de les modifier ultérieurement pour leurs propres objectifs. Par exemple, certaines des images pourraient être agrandies et étudiées par des pathologistes intéressés par l'étude de la rapidité de diffusion d'une infection qui a pour effet de noircir les feuilles des plantes.

Nous assistons ainsi à une situation où une combinaison spécifique d'intérêts, de possibilités, d'accessibilité et de caractéristiques des objets en question déterminent ce qui est identifié comme une donnée – c'est-à-dire quelles photos sont considérées comme fondement empirique pour l'analyse et la résolution d'un problème particulier. Dans ce cas comme dans mille autres où les données circulent au sein de contextes différents, ce qui est valorisé comme une donnée – et ce qui est écarté comme élément non-pertinent pour l'analyse[69] – n'est pas toujours le même objet ou le même ensemble d'objets, mais continue de changer (par le format, par la partie de l'objet qui devient pertinente, ou par la manière dont l'objet se lie à d'autres[70]). Et comme l'objet change, la manière dont ses caractéristiques peuvent être conceptualisées et manipulées change aussi, ce qui génère des interprétations différentes sur ce que ces objets – ces données – peuvent représenter.

La vision représentative ne peut que lire ces manipulations comme une distorsion de la signification originale des données, et en réalité, ses partisans prônent souvent une claire distinction entre les données « brutes » (*raw data*) et celles qui ont été traitées ou bien ultérieurement élaborées pour en rendre l'analyse possible. L'élaboration des données brutes est reconnue comme une étape nécessaire pour pouvoir les utiliser comme source de connaissance, mais est généralement vue avec méfiance par celui qui interprète les données comme une représentation de la réalité : d'après cette perspective, la difficulté qu'il y a à interpréter les données consiste à s'assurer que les méthodes utilisées pour les formater ne trahissent pas leur signification d'origine. Dans la vision relationnelle en revanche, la distinction entre données « brutes » et données « élaborées » n'est plus pertinente du point de vue épistémique et est donc difficile à maintenir dans le cas des Big et des Open Data. En effet, (1) la différence entre données « originelles » et données « traitées » n'est ni évidente, ni claire, ni particulièrement utile pour l'analyse scientifique, où les données brutes peuvent être aussi problématiques que les données élaborées, selon les méthodes et les instruments utilisés pour les générer[71] ; (2) utiliser les données dans le format et dans l'ordre dans lesquels elles ont été originellement créées est souvent impossible, comme en attestent les nombreuses méthodes utilisées par les chercheurs pour « nettoyer » les données et les rendre traitables (y compris les techniques statistiques de réduction et de normalisation[72]) ; et (3) les mêmes objets qui sont sélectionnés et interprétés comme des données changent selon le contexte de recherche. D'après les partisans de la vision représentative des données, ces trois situations sont littéralement incompréhensibles : si les données constituent des représentations fidèles de la réalité, comment peuvent-elles être continuellement modifiées et, malgré cela, continuer à être utiles comme source de connaissance ? Cette réticence à reconnaître l'importance épistémique des processus d'élaboration des données se traduit en réticence à accorder de l'attention à de tels processus et

[69] McAllister 2007 ; Loettgers 2009 ; Woodward 2010 ; Boumans et Leonelli 2019.
[70] Voir aussi les analyses de Niccolò Tempini, Mary Morgan et James Griesemer contenues dans l'ouvrage à paraître *Varieties of Data Journeys* (Leonelli et Tempini, 2019).
[71] Voir aussi Gitelman, 2013.
[72] Mayo, 1996.

à les documenter afin de les rendre visibles et contestables. À cause de cette tendance à occulter ou à oublier les transformations subies par les données au cours de leurs circulations, les banques de données sont ensuite conçues comme une « boîte noire » ce qui, comme nous l'avons déjà vu, a des conséquences extrêmement problématiques pour le traitement et l'utilisation des Big Data.

Un troisième problème de la vision représentative des données est lié à la difficulté, discutée au chapitre précédent, qu'il y a à trouver des critères universels et absolus pour en évaluer la qualité et la fiabilité[73]. De tels critères n'existent pas, justement parce que la qualité des données doit être évaluée en fonction des objectifs précis de la recherche pour laquelle elles sont utilisées et du type de données en question (dont le format varie énormément selon les méthodes utilisées pour les générer). Le pluralisme des objectifs et des méthodes couramment utilisés dans le monde de la recherche reflète la variété et la diversité de la réalité qui nous entoure et que nous cherchons à comprendre et à contrôler, et se retrouve ensuite dans la variété des types de données produites par nos interactions avec le monde. Le choix de ce qui est considéré comme donnée fiable est donc inévitablement lié aux circonstances spécifiques de leur utilisation. Pour en revenir brièvement à notre exemple botanique, les photographies d'une plante peuvent être interprétées comme des données utiles à l'étude de beaucoup d'aspects différents, du développement morphologique de l'espèce en question, des symptômes d'infection, de l'effet de telles conditions météorologiques sur la coloration des feuilles, à la présence de parasites dans une localité déterminée. Chacune de ces interprétations est en partie conditionnée par les caractéristiques physiques des photographies (y'a-t-il des insectes sur les feuilles ? De quelle couleur sont les feuilles ?) et en partie par la façon dont celui qui utilise ces objets en accentue la traçabilité en tant que données (en retouchant les parties des photos où apparaissent les insectes, en extrayant les mesures des dimensions des feuilles grâce à des analyses informatiques, etc.). Que les données contiennent, de manière objective et indépendante du contexte d'utilisation, des informations précises sur le monde n'est donc pas une évidence. Les caractéristiques des objets qui sont considérés comme des données délimitent certainement le type d'utilisation et d'interprétation que l'on peut en faire mais, en même temps, il est possible d'obtenir des informations différentes des objets eux-mêmes selon la façon dont ils ont été gérés et interprétés – ce qui dément à nouveau l'idée que les données elles-mêmes sont des représentations fidèles, objectives et immuables de la réalité.

Si la fiabilité des données ne provient pas de leur capacité à représenter le monde de manière objective et immuable, alors d'où vient-elle ? L'alternative proposée ici est la conceptualisation *relationnelle* des données, plutôt que représentative. La donnée est conçue comme un objet mis en relation avec une question irrésolue, d'une façon et pour des raisons qui dépendent de la situation dans laquelle la question est posée, et non pas comme une représentation du réel. Dans la vision relationnelle, *n'importe quel objet peut endosser le rôle de « donnée » à condition (1) qu'il soit traité comme une source potentielle de connaissance empirique et (2) qu'il soit possible de le mobiliser afin de le rendre accessible à plus de personnes*. En d'autres termes, les intentions et les attentes des chercheurs déterminent fortement quels objets sont sélectionnés comme sources de connaissance potentielles, mais ces objets doivent pouvoir être visibles concrètement et inspectés par le plus de personnes possibles, de manière à servir de preuve aux allégations auxquelles ils sont liés. Ce n'est pas le cas des données si elles existent seulement dans l'esprit de celui qui les utilise : au moins en théorie, elles doivent être accessibles à

[73] Leonelli (2012), Canali (2016).

d'autres, qui peuvent en évaluer la valeur scientifique et en vérifier la fiabilité en tant que fondements empiriques de connaissance. Selon l'approche relationnelle, la signification assignée aux données ne dépend pas seulement de leurs caractéristiques physiques et de ce qu'elles représentent, mais aussi des intentions et des instruments utilisés pour les analyser et pour défendre certaines interprétations ; et la fiabilité des données dépend surtout de la crédibilité et de la rigueur des processus utilisés pour les produire et les analyser.

Cette vision reconnaît que n'importe quel objet peut fonctionner comme une donnée, ou bien arrêter de le faire, selon les circonstances – une observation bien connue de ceux qui traitent des données historiques, souvent conservées dans des archives oubliées par tous et donc réduites à des objets sans signification, ou des données provenant d'activités qui n'ont aucun rapport avec la recherche mais qui peuvent cependant être utilisées pour générer de nouvelles connaissances, comme par exemple la quantité de produits biologiques vendus en France, ou le nombre d'entreprises en faillite suite à la crise économique de 2008. De plus, l'approche relationnelle reconnaît que les objets qui sont considérés comme des données sont souvent modifiés au cours de leurs circulations entre sites de production, dissémination et réutilisation. Non seulement les données peuvent changer de format, mais ces changements peuvent avoir un fort impact sur comment, où et par qui elles sont utilisées comme sources de connaissance. En conséquence, il devient extrêmement important de documenter les processus de gestion et de transformation des données, particulièrement dans le cas des Big Data qui circulent sans cesse au sein de canaux digitaux, et sont regroupées, analysées et interprétées sous différentes formes et manières. La vision relationnelle encourage ainsi le traitement des données et l'attention à leur histoire, en soulignant leur qualité d'objets en perpétuelle évolution et soumis à des modifications parfois radicales, ainsi que les implications de cette qualité sur leur capacité à démentir ou à confirmer des hypothèses.

Une objection évidente à cette façon de penser les données comme des sujets changeants continuellement consiste à observer que ces transformations compliquent grandement le suivi des données pendant qu'elles circulent d'un lieu à l'autre. Comment pouvons-nous dire que les données utilisées dans un certain contexte sont les mêmes que celles qui ont été produites dans un contexte différent, si leur utilisation et leur forme même continuent à changer et à s'adapter aux nouvelles situations ? Et comment font les données pour conserver leur intégrité si, chaque fois qu'elles sont transférées et réévaluées, leurs caractéristiques physiques peuvent changer ? La réponse à cette objection est pragmatique[74]. Il est parfaitement possible de penser les données comme des entités historiques qui évoluent et changent à travers la reproduction et l'accumulation d'expériences différentes, mais dont la provenance peut et doit toujours être reconstruite, au moins en théorie, pour pouvoir en évaluer la validité. Le meilleur moyen de théoriser les mouvements et les transformations des données est celui que nous utilisons pour penser la reproduction des organismes vivants et la façon, imparfaite et imprévisible, certes, dont les caractéristiques de chaque génération sont transmises à celle qui suit. Les données peuvent donc être conceptualisées comme des *lignées* : des dynasties d'objets qui se transforment lors du passage d'une forme à l'autre mais dont l'étude dépend au moins en partie de notre capacité à en évaluer l'origine et la provenance.

[74] Ma théorisation est inspirée par le philosophe américain pragmatique John Dewey, comme je l'ai expliqué dans Leonelli (2016a).

L'idée que la donnée appartient à une lignée amène à une autre objection importante pour la vision relativiste de la donnée : elle porte sur le relativisme inévitable lorsqu'on considère les données uniquement en lien avec des contextes spécifiques plutôt que comme une source de connaissance objective. Ne risque-t-on pas de créer un concept de donnée complètement relativiste, où chaque chose utilisée comme donnée en devient automatiquement une, et tout objet peut être légitimement interprété pour fonder des affirmations de n'importe quel type ? Cette préoccupation est particulièrement vivace en raison du manque de respect flagrant, de la part de ceux qui insistent pour créer des vérités subjectives et complètement détachées du monde réel sur la base d'intérêts financiers ou de convictions personnelles[75], pour des faits absolument évidents pour tous (que notre planète est ronde et qu'elle n'est pas sous le contrôle de reptiles provenant d'une autre planète, par exemple).

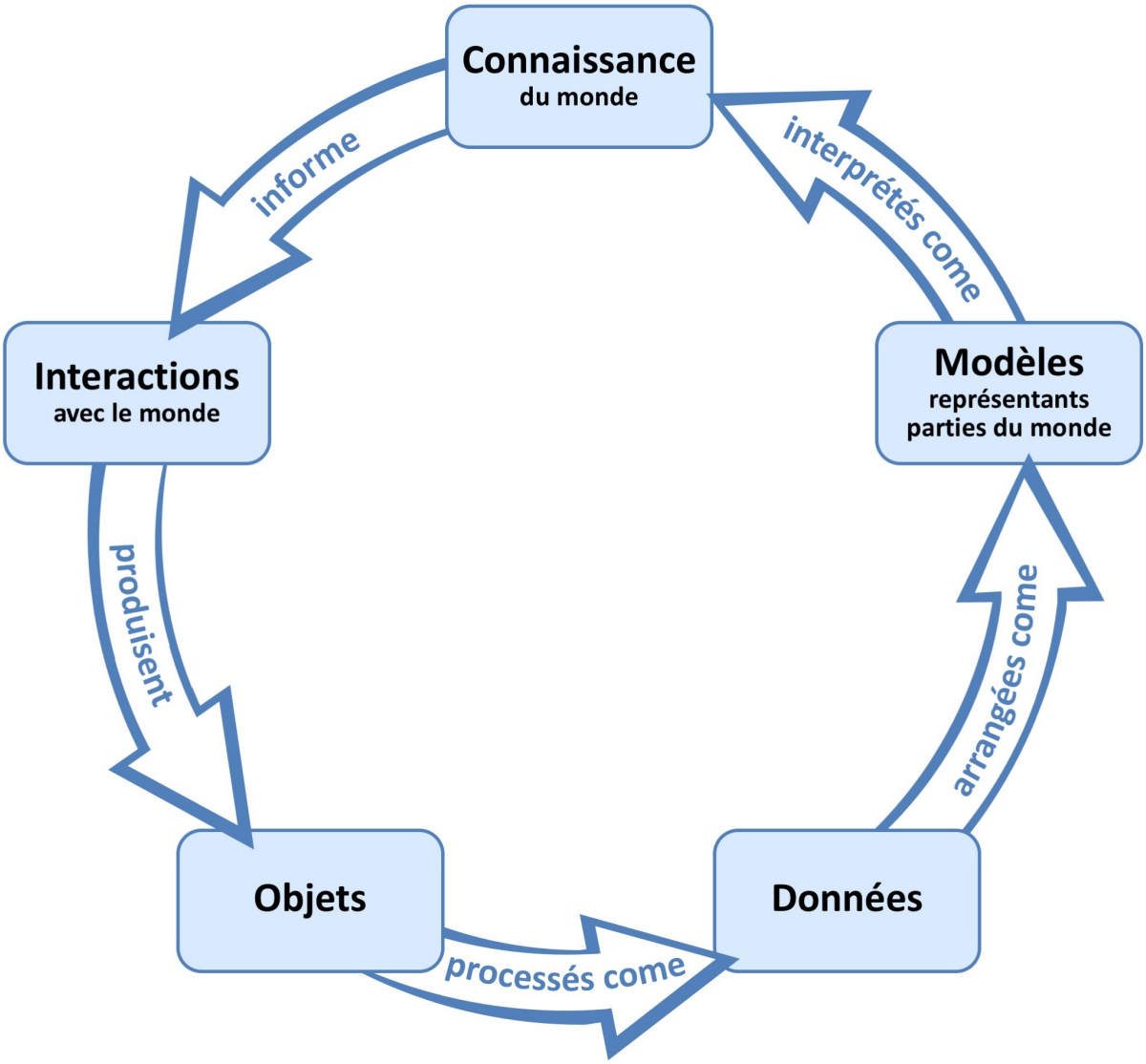
La vision représentative des données a une façon, en apparence simple et convaincante, de répondre à ces objections, ce qui explique pourquoi tant de gens s'obstinent à penser les données de cette manière : si les données sont une représentation objective d'un aspect déterminé du monde, alors il suffit de comparer les données au monde pour savoir si elles sont correctes et s'il est pertinent de croire à leur fiabilité. Pourtant, sans doute pour la plupart des données utilisées dans le domaine scientifique, une telle évaluation n'est malheureusement pas si facile. De quelle manière les données génétiques, comme par exemple la séquence GTTACCTGAAA, représentent de façon claire et irréprochable les informations héréditaires qui se trouvent à l'intérieur d'une cellule ? Que signifie, pour les nombres 34, 72 et 91, la représentation des dimensions d'un organisme ? En quoi une échographie représente un fœtus dans l'utérus maternel et fournit des informations sur son état de santé ? Bien souvent, les objets que les chercheurs utilisent comme des données ne ressemblent pas de manière évidente aux aspects du monde qui servent d'étude, et leur interprétation réclame de l'habileté et des connaissances précises (comme celles utilisées pour interpréter une échographie en tant que représentation d'une cardiopathie du fœtus). De même, les critères et méthodes utilisés pour attribuer un sens à ces objets changent souvent avec le temps et le contexte. Cela ne veut pas dire que ces critères et ces méthodes sont arbitraires, mais plutôt que leur sophistication et la précision avec laquelle ils sont appliqués dans des situations variées continue à croître petit à petit avec l'avancée des technologies et des connaissances scientifiques. Il existe des critères et des méthodologies toujours plus sophistiquées pour établir et vérifier qu'un objet peut être utilisé de manière fiable comme source de connaissance sur un phénomène déterminé, et à quelles conditions. Et en effet, il n'existe pas de méthode ou de critère scientifique qui permette d'interpréter les photographies satellites de notre planète comme preuve que la terre est plate, alors qu'il existe mille manières d'identifier des erreurs dans le raisonnement utilisé par ceux qui promeuvent cette idée, ce qui démontre que la connaissance produite de cette manière n'est pas solide. Accepter une conception relationnelle des données n'équivaut donc pas à accepter un relativisme total. Au contraire : cette vision encourage la vigilance continue sur les raisons, les méthodes et les objectifs pour lesquels certains objets sont identifiés, manipulés et proposés comme des données pour prouver des affirmations déterminées, en dehors ou au sein du monde scientifique.

[75] Les adeptes des théories de David Icke ou ceux qui sont persuadés que la Terre est plate sont seulement un des nombreux exemples contemporains de scénarios extrêmes.

Des données à la connaissance : une question de classement

Accepter la vision relationnelle des données a des répercussions sur la façon dont on conceptualise le processus complet de recherche empirique, qui est résumé dans le graphique de la figure 5.

Figure 5. Le processus de recherche associé à la vision relationnelle de la donnée (copyright Sabina Leonelli, réalisation Michel Durinx).



La recherche empirique commence, comme dans le cas de la vision représentative, par l'interaction entre l'homme et le monde. Ces interactions produisent des artefacts de genres variés, comme par exemple des nombres, des mesures, des symboles, des photographies, des descriptions et des graphiques. Certains de ces objets sont ensuite sélectionnés et traités afin de servir, au moins potentiellement, en tant que source de connaissance. Les objets ainsi manipulés

sont ceux que nous appelons les données. Comme je l'ai déjà expliqué, ce que ces données sont censées représenter n'est pas déterminé seulement par les caractéristiques physiques des données elles-mêmes, mais aussi par les suppositions et le contexte de ceux qui en évaluent la signification potentielle. Interpréter les données comme source de connaissance comprend donc deux passages ultérieurs : (1) la création de modes de classement des données – souvent appelés *modèles* dans le monde scientifique – qui en manifestent une certaine fonction représentative[76] et (2) l'utilisation de ces modèles comme fondement empirique pour la production de connaissance. Dans cette vision de la recherche, la fonction représentative des données continue donc d'être présente, mais ce ne sont pas les données en tant qu'objets en soi qui représentent – de manière plus ou moins fiable et exacte – une partie de la réalité. Ce qui remplit la fonction représentative, c'est le modèle de données que toute personne qui les interprète construit lorsqu'elle décide comment les organiser. En d'autres termes, c'est un certain *classement* des données – la manière dont elles sont visualisées et rendues ainsi pertinentes pour un certain type d'analyse – qui représente un aspect du monde et le rend plus accessible à l'étude scientifique[77]. Mettre les données en ordre est le moyen de les rendre utilisables comme preuves de faits déterminés et comme sources d'une nouvelle connaissance. Les données ne sont donc pas en soi une base objective pour le savoir : c'est la manière dont nous les organisons et les visualisons qui détermine la signification qui leur est assignée.

Prenons à nouveau l'exemple des données botaniques pour illustrer ce processus de manière plus concrète. Dans ce cas, l'amateur qui fait des photographies lors de sa promenade en forêt produit des objets grâce à son interaction avec le monde – les photos – qui sont ensuite traités par les chercheurs dans l'espoir qu'ils puissent servir de données (par exemple quand les photos sont formatées pour être insérées dans l'archive digitale). Les chercheurs organisent et classent les données ainsi obtenues d'une manière qui les aide à représenter différents phénomènes : dans le cas des morphologies, la forme d'une espèce particulière dans une certaine localité ; dans le cas des pathologies, les symptômes potentiels d'une infection des feuilles. Ces modèles sont ensuite testés pour en vérifier la fiabilité et la pertinence pour les phénomènes qu'ils documentent – par exemple, les chercheurs vérifient que le modèle des symptômes de l'infection obtenu à partir de l'analyse des images trouvées *online* reflète les caractéristiques des modèles de données qui proviennent d'autres sources, et ils retournent, quand c'est possible, sur le lieu en question pour vérifier la justesse du modèle[78]. Si les modèles sont jugés adéquats, ils sont utilisés comme source de connaissance sur la façon dont l'infection se manifeste chez les plantes en question. S'ils ne sont pas jugés adéquats, les chercheurs réanalysent les données et essaient de les classer différemment – ce qui implique ensuite un changement radical du type d'objet qui est considéré comme une donnée et/ou l'aspect de la réalité qui est étudiée[79].

Nous avons déjà vu comment l'une des principales caractéristiques des Big Data, qui les rend particulièrement intéressantes comme sources de connaissance, est la possibilité d'examiner et

[76] Leonelli (en cours d'impression), Bokulich (2018) et Suppes (1962).
[77] Comme je l'ai déjà mentionné dans le premier chapitre, cette vision des données trouve ses racines dans le travail de nombreux philosophes, et de manière peut-être évidente dans celui de Michel Foucault et de Jacques Derrida. Ce n'est pas mon objectif d'en retracer les origines dans ce livre, mais plutôt d'en spécifier ses caractéristiques et son importance dans le contexte contemporain.
[78] Par exemple Shavit et Griesemer (2009) et Leonelli et Tempini (2018).
[79] Comme détaillé dans Leonelli (en cours d'impression), l'identification des phénomènes étudiés par les chercheurs est fortement conditionnée par le choix et la gestion des données – un point épistémologiquement significatif qui a également été étudié par Bogen et Woodward (1988), McAllister (2007), Teller (2010), Massimi (2011) et Feest (2011).

de comparer tant de types de données différents, obtenus par des moyens souvent très variés mais potentiellement pertinents pour l'analyse du même phénomène. L'analyse comparative de ce que l'on peut apprendre de la juxtaposition de types de données différents est souvent appelée *triangulation*, et est fortement associée à l'idée que plus il y a de types de données qui confirment une certaine interprétation, plus cette interprétation peut être jugée fiable et empiriquement correcte. Il est pourtant essentiel de préciser que la triangulation fonctionne seulement si les données qui sont juxtaposées ont des origines différentes entre elles (et proviennent donc de lignées différentes). Si les données sont créées par le même groupe de chercheurs, avec les mêmes instruments et sur la base des mêmes hypothèses théoriques, il est en réalité probable qu'elles soient interprétées de manière similaire, mais il est difficile de savoir si l'interprétation dépend des similarités entre les données elles-mêmes ou de leur fiabilité comme source de connaissance. Si les données qui ne partagent pas la même histoire et n'appartiennent pas à la même dynastie tendent dans la même direction, nous avons en revanche une confirmation plus crédible du fait que l'interprétation est correcte. La philosophe Alison Wylie a étudié en détail les manières dont les données de différents types sont utilisées pour trianguler les interprétations et renforcer ainsi la base empirique du raisonnement scientifique, et elle conclut par l'idée que connaître l'origine et l'histoire de la façon dont les données sont gérées permet de vérifier de manière adéquate si et comment certains modes de classement (modélisation) des données sont fiables en tant que sources de connaissance[80].

Dans la vision représentative des données, il est difficile de reconnaître le rôle fondamental joué par l'histoire des données en tant qu'objets dans leur interprétation comme sources de connaissance. Pour celui qui pense les données comme des représentations objectives et immuables du monde, les conditions dans lesquelles elles sont formatées et classées importent peu : ce qui importe est de réussir à dévoiler leur réelle signification. Cette approche s'accompagne facilement de l'idée que le regroupement d'autant de données constitue en soi une augmentation de la base empirique de la connaissance. L'accumulation des données équivaut à l'accumulation de nombreux faits, un véritable trésor dont on peut extraire de nouvelles découvertes grâce à des techniques reposant sur l'induction et les statistiques. Il est simple de comprendre comme celui qui adopte cette vision des données est facilement la cible des fausses promesses liées à l'utilisation des Big Data, comme par exemple celle selon laquelle elles sont universellement fiables, impartiales et utilisables pour n'importe quel type d'analyse[81].

La vision relationnelle des données s'accompagne en revanche d'une vision moins utopique des conditions dans lesquelles les Big Data peuvent être utilisées comme sources fiables et efficaces de connaissance. Dans la vision relationnelle, l'obtention de connaissance comporte le *positionnement* d'objets choisis pour remplir la fonction de données (et donc de leurs caractéristiques physiques) *en relation* avec d'autres éléments cruciaux pour l'interprétation, comme l'objectif de la recherche, les hypothèses conceptuelles sur lesquelles elle est fondée et

[80] Wylie (2002, 2017) ; Chapman et Wylie (2015).

[81] Je ne veux pas dire que la vision représentative des données soit nécessairement incompatible avec une vision des Big Data bien plus sophistiquée, mais plutôt que la vision relationnelle s'accorde bien mieux à une gestion et à une utilisation des Big Data toujours attentives à de potentiels problèmes méthodologiques – ce qui la rend préférable à mes yeux.

le type de connaissance – théorique ou pratique – qui est en est tirée[82]. Ce positionnement comporte donc des présupposés et des choix bien plus larges que ceux impliqués dans l'application de méthodes statistiques. Les procédures avec lesquelles les données sont traitées et classées sont essentielles à leur utilisation comme source de connaissance, et au choix et à l'utilisation de critères pour en juger les potentialités représentatives vis-à-vis de la réalité. La vision relationnelle des données reconnaît donc l'énorme travail nécessaire pour documenter les circulations des données et en rendre possible l'observation au cours des processus d'interprétation.

[82] Ma conception de la connaissance s'accorde avec celle, d'inspiration pragmatique également, proposée par Chang (2017) : connaissance comme capacité à agir, dont la connaissance exprimée en forme linguistique est seulement une des composantes possibles.

Chapitre 4. Comment encourager à faire mieux : vers une science participative et responsable

Retournons maintenant aux questions fondamentales avec lesquelles nous avons commencé notre étude des Big Data. Comment peuvent-elles être utilisées pour se protéger de la *post-vérité* et des si nombreuses forces qui cherchent à manipuler les faits pour leur bénéfice personnel ? Comment pouvons-nous nous assurer que la connaissance scientifique à laquelle nous nous fions chaque jour ait un fondement empirique solide et fiable ? La réponse à ces interrogations ne consiste pas à agiter une baguette magique et à invoquer une certaine méthode ou une figure professionnelle comme la solution unique et parfaite à la crise épistémique de notre époque. Il n'existe pas de magie qui puisse résoudre d'un seul coup les tensions et l'incertitude inhérentes à la multitude de voix, de secteurs et d'intérêts impliqués dans les circulations des données. Comme illustré dans le chapitre précédent, partir d'une vision relationnelle des données signifie accepter qu'il n'existe pas de référents fixes ni de techniques infaillibles pour en juger la qualité et la valeur scientifique et épistémique. Ce qui compte est la manière dont elles sont classées et visualisées en fonction de la situation et de l'objectif de l'analyse.

Cette observation ne doit pas décourager ceux qui se fient à la science – y compris la médecine et la technologie – comme point de référence fondamental pour comprendre le monde et soi-même. La vision relationnelle des données a en réalité une conséquence extrêmement importante pour la gestion de la recherche scientifique et de l'intelligence artificielle : celle de souligner le lien étroit entre les décisions prises lors du classement des données et la manière dont elles sont interprétées. Des techniques comme l'apprentissage automatisé rendent toujours plus facile l'automatisation de certaines de ces décisions, qui peuvent être déterminées et mises en œuvre par des algorithmes capables d'évoluer et de s'améliorer au cours des expériences réalisées. Mais le jugement humain continue à déterminer la manière dont les algorithmes privilégient certaines sources et certains types de données par rapport à d'autres ; à déterminer le choix de techniques de visualisation et d'analyse statistique ; ainsi que les objectifs, les présuppositions et les préférences présentes dans le processus d'analyse.

Ces décisions sont tout autant scientifiques qu'éthiques, et démontrent l'importance de l'éthique et de la participation sociale dans le développement des systèmes qui permettent la mobilisation et la réutilisation des données. Des choix qui peuvent sembler purement techniques – quelle forme de calcul de probabilité utiliser, à quel type de classification se fier – sont en réalité chargés d'implications pour la manière dont la connaissance qui en découle peut transformer la société. Et dans la mesure où les scientifiques impliqués dans l'analyse des Big Data sont seuls responsables des décisions qu'ils prennent, leurs choix ne peuvent être compris en-dehors du contexte social dans lequel la connaissance est produite et utilisée. Au sein d'une société démocratique, cela implique de rechercher continuellement le dialogue et la confrontation entre chercheurs et autres groupes sociaux, dont l'expérience de certaines situations les rend capables de contribuer de manière décisive à l'évaluation des hypothèses et des choix effectués dans la production, la sélection, la dissémination et l'interprétation des données. Parents, entrepreneurs, patients, enseignants, chercheurs ont une connaissance unique et précieuse de ce que signifie s'occuper de ses enfants, gérer un agenda, vivre au quotidien avec la maladie, éduquer les nouvelles générations et satisfaire les préférences du public tout en préservant sa santé. Dans la vision relationnelle des données, ce type de connaissance doit

être confronté aux méthodes scientifiques utilisées pour analyser les Big Data, et être incorporé autant que possible lors des circulations des données. La capacité des données à contourner des séparations rigides entre différents types d'expertise – que nous avons vue à l'œuvre chaque fois que les Big Data commencent à circuler – peut et doit être exploitée pour modifier et améliorer la communication entre des secteurs sociaux qui sont devenus trop spécialisés et incapables de se confronter entre eux de manière constructive et nécessaire pour le développement d'une société fonctionnelle. Les données circulent inévitablement à travers de nombreux mondes sociaux différents, et c'est lors de ces circulations qu'elles acquièrent une valeur épistémique en tant que source de connaissance : reconnaître cette réalité est un pas important vers une conception de la production scientifique qui n'exclut pas le monde extérieur à la recherche, mais au contraire l'embrasse et l'inclut pour juger et établir ce qui constitue une connaissance – à la fois fiable et juste – du point de vue scientifique et éthique.

Certains des initiés rappellent à ce stade que, justement à cause de la nature fortement divisée et technique du travail informatique, il n'est jamais possible pour les experts impliqués de prévoir quelles peuvent être les conséquences éthiques et sociales de leurs choix, ou tout au moins d'impliquer des secteurs sociaux différents dans cette évaluation – et donc le lien entre éthique et science, bien que convaincant du point de vue théorique, se rompt une fois confronté aux limites pratiques de l'utilisation des données dans la vie de tous les jours. Dans le monde concret de l'analyse des Big Data, disent-ils, il est impossible d'évaluer les implications de nos actions, parce que ces implications émergent seulement une fois qu'un programme, un instrument ou un type d'analyse déterminé est mis en œuvre. Mais à ce stade, continuent les critiques, il n'est plus possible de modifier les hypothèses sur lesquelles les techniques d'analyse ont été construites, en partie parce qu'elles sont habituellement incorporées dans un appareil technologique et informatique extrêmement complexe, mais aussi parce que presque personne, parmi ceux qui appliquent cet appareil à des situations de recherche, n'a une bonne compréhension de la « boîte noire » créée par cette technologie. Voici un exemple concret : d'une part, celui qui construit des algorithmes d'analyse pour des tests Google ne peut se préoccuper de toutes les manières possibles dont ces algorithmes peuvent être détournés, parce qu'habituellement, l'abus potentiel émerge seulement quand les algorithmes sont incorporés dans l'énorme appareil de Google et rendus accessibles au public ; d'autre part, une fois que les algorithmes sont assimilés et publiés de cette façon, il est très difficile de les modifier et d'identifier exactement lesquelles de leurs caractéristiques deviennent problématiques du point de vue social. La conclusion de ce raisonnement est qu'aucun des initiés – ni ceux qui produisent les données et les infrastructures qui y sont liées, ni ceux qui en font usage pour produire de la connaissance – ne semble pouvoir assumer la responsabilité des conséquences des choix effectués dans la gestion des données. Nous nous retrouvons ainsi en proie au déterminisme technologique : bien que l'on soit conscient des implications fortement négatives de l'analyse des Big Data, il ne semble pas possible d'arrêter l'abus. Dans cette optique, toutes les technologies impliquées dans l'analyse des Big Data – qu'elles soient officiellement reconnues comme intelligence artificielle ou non – peuvent être interprétées comme étant supérieures à l'homme, par le simple fait qu'elles en dépassent désormais la capacité de jugement et de délibération.

Je crois personnellement qu'il est non seulement possible mais absolument nécessaire d'éviter de se résigner au déterminisme technologique, qui est dangereux aussi bien pour la qualité et la crédibilité de la connaissance scientifique que pour son impact social – et que la vision relationnelle contient des pistes essentielles pour nous y aider. Les deux prochaines parties

examineront brièvement deux stratégies que peuvent utiliser ceux qui travaillent avec les Big Data, pour tenir compte des possibles implications éthiques et sociales de leur propre travail, et qui font valoir une forme limitée mais cruciale de contrôle sur la connaissance produite, atténuant ainsi – si ce n'est évitant complètement – les risques épistémiques discutés jusqu'à maintenant.

L'intégration de l'éthique dans la recherche scientifique

La première stratégie est l'adoption de procédures pour intégrer l'éthique dans les choix techniques de gestion et d'analyse des données. Cela requiert avant tout d'abandonner l'idée selon laquelle il est nécessaire de disposer, pour évaluer la valeur éthique d'une innovation, d'un pronostic précis de son impact social potentiel. C'est une chose de dépenser du temps et des ressources à imaginer et vérifier comment une certaine innovation peut être incorporée dans divers contextes – un effort fondamental pour la production de connaissance et de technologies adaptées aux exigences sociales et aux déterminismes culturels de celui qui les utilise. C'en est une autre de chercher à élaborer des moyens de prévoir, quantifier et contrôler complètement les implications exactes – une exigence impossible pour n'importe quelle nouveauté. Cela vaut même pour des innovations dans le domaine médical, où l'approbation de nouveaux types de traitement est soumise à des années de tests et de vérifications sévères, mais qui ne garantissent pas d'éviter les effets collatéraux inattendus, surtout à long terme. Il est donc absurde de penser que l'analyse des implications éthiques et sociales potentielles a un sens seulement quand celle-ci se fait dans un cadre précis.

En second lieu, il faut abandonner l'idée selon laquelle la recherche doit se focaliser seulement sur des innovations aux effets socialement positifs. Ce n'est une hypothèse réaliste pour aucun type d'innovation : elles peuvent toutes être utilisées à des fins dommageables, d'une manière ou d'une autre, à certaines parties de la société, et toutes sont liées à des situations de risque et d'incertitude. Il suffit de penser à l'impact qu'ont des technologies comme l'intelligence artificielle sur le marché du travail, où des millions de personnes – des chauffeurs de taxi aux enseignants, en passant par les ouvriers et les avocats – risquent d'être remplacés par une série d'algorithmes, comme cela s'est déjà produit pour tant d'autres avant eux lors de l'apparition du *personal computer*. Dans ce cas, il est impossible d'éviter l'impact social sans arrêter complètement le développement technologique : solution extrême qui priverait l'humanité des avantages et des opportunités offertes par ces mêmes technologies. L'acquisition de connaissances amène toujours des avantages et des inconvénients, et le problème éthique réside dans l'évaluation de la manière dont ceux-ci se lient entre eux, ainsi que le poids qu'ils peuvent avoir sur des réalités sociales différentes. Dans le cas de l'impact de l'intelligence artificielle sur le marché du travail, l'évaluation éthique concerne les types de caractéristiques et de fonctionnalités à privilégier pour le développement de nouveaux algorithmes, les secteurs qui peuvent le mieux bénéficier de ces changements et les types de réorganisation culturelle, pédagogique et sociale qu'il est préférable d'encourager pour favoriser un impact positif sur la population (par exemple en admettant que la gestion des Big Data a l'énorme potentiel de créer de nouvelles formes d'emploi, qui requièrent pourtant une formation différente de celle que la très grande majorité de la population âgée de moins de trente ans a reçue dans le système éducatif français).

En d'autres termes, l'évaluation critique consiste à tenir compte à chaque étape, autant que possible, des circonstances dans lesquelles un certain résultat pourrait être utilisé, ainsi que de

la situation et des attentes des utilisateurs – et à utiliser cette connaissance comme support des décisions techniques qui concernent les sources, les formats, la classification et l'analyse des Big Data. Cela implique de rechercher ceux qui seront potentiellement intéressés par la connaissance produite ; pour quel motif ; qui est inclus ou exclu de cette utilisation potentielle ; et comment modifier le processus de recherche de manière à rendre les résultats moins discriminants, plus durables ou moins inclusifs selon les exigences.

Il est important de noter que poser ces questions ne garantit pas l'existence de réponses adéquates, et génère souvent des compromis et des doutes plutôt que des solutions optimales. L'évaluation éthique ne fournit aucune certitude sur le futur, mais se poser ces questions demeure une étape fondamentale vers l'acquisition d'une meilleure conscience de l'éventuel impact éthique et social de la gestion des données. Avec cette évaluation, celui qui produit, mobilise et analyse les données est d'autant plus responsable de l'impact de ses procédures. De cette manière, la production de connaissance est orientée selon qui peut en bénéficier et l'éthique devient partie intégrante du processus de recherche.

Il y a de nombreux exemples d'innovations développées grâce à l'analyse des Big Data dont l'effet néfaste aurait pu être modéré si seulement celui qui les avait développées s'était interrogé de manière sérieuse et systématique sur l'impact des choix effectués. Prenons de nouveau le cas de Facebook qui, surtout dans ses premières années d'existence, a ostensiblement absorbé et revendu les données personnelles de ses utilisateurs sans se préoccuper des conséquences potentielles, devenant ainsi un véritable Big Brother – un instrument de surveillance utilisé par de nombreuses entreprises et institutions comme source d'informations sur les citoyens, qui négligent le fait que ce type de *social media* ne représente pas toujours la vie réelle des internautes et peut donc générer une connaissance totalement erronée. Ce type d'abus, pour lequel Facebook rechigne encore à assumer ses responsabilités, a initialement aidé l'entreprise à croître, mais s'est retourné contre elle sur le long terme, en abîmant énormément son image et le rapport de confiance avec les utilisateurs. De nombreux autres exemples sont liés à la tentative d'utiliser les Big Data à des fins médicales. Un cas bien connu est celui de Google Flu Tends, un programme lancé en 2008 avec pour objectif d'utiliser les Big Data générées par les recherches effectuées sur Google pour prévoir l'apparition d'épidémies de grippe. L'idée était de profiter du fait que beaucoup des utilisateurs qui recherchent les mots « grippe », « symptômes » et « fièvre » le font bien avant d'appeler un médecin, et parfois même comme une alternative au système sanitaire. Google espérait donc analyser ces données pour en tirer des pronostics bien plus fiables que ceux provenant de l'analyse des données médicales officielles, et déclara même en 2012 que le programme permettait d'identifier des foyers d'infection cinq jours avant que celle-ci ne devienne visible par les services de santé. Le programme ne tenait pourtant pas compte de la variété des terminologies utilisées par les utilisateurs pour décrire les symptômes, ainsi que de la quantité de recherches, formellement similaires à celles de ceux qui sont malades, effectuées en réalité pour une toute autre raison – en d'autres termes, trop peu d'études avaient été faites sur les utilisateurs potentiels, sur les usages de Google et sur la manière dont ceux-ci pouvaient pervertir l'efficacité de l'analyse des données. Et cette incarnation d'apprentissage automatisé est justement devenue un emblème de connaissance erronée : alors qu'en 2013 le programme ne parvînt pas à prévoir une épidémie particulièrement grande, une analyse indépendante des conclusions de Google montra que le

nombre de cas diagnostiqués par Google Flu Trends était deux fois supérieur à ceux effectivement vérifiés[83].

Ce type d'erreur dans l'identification des données pertinentes, et dans leur classement et analyse, démontre l'impossibilité de séparer l'éthique de l'évaluation de la solidité, de la sophistication technique et de la crédibilité des méthodes scientifiques[84]. Dans le cas de Google Flu Trends, la connaissance obtenue est aussi douteuse que discriminante pour les individus qui ont une véritable grippe mais qui sont exclus de ce type de recherche. Le plus grand défi pour l'utilisation des Big Data est justement de générer des mécanismes de réflexion et de responsabilité, à chaque étape de la gestion des données, qui aident à identifier au plus vite les potentielles sources d'erreur et de discrimination, et permettent de corriger et, si nécessaire, de sanctionner les décisions qui finissent par être problématiques du point de vue social. Nous sommes sûrement encore loin – si cela est un jour possible – de créer des technologies qui peuvent remplacer la capacité humaine à évaluer le contexte et les implications de leurs actions. Pour le moment, la manière dont les Big Data sont gérées est en train de créer un écart grandissant entre la croissance exponentielle des banques de données et autres algorithmes pour l'analyse des Big Data, et le manque de procédures et de principes qui permettent une évaluation sérieuse de leur impact. Au cours des cinq dernières années, une centaine de programmes semblables à Google Flu Trends ont été créés, souvent sans aucun type de contrôle et sans capacité ni volonté de prendre le temps de réfléchir sur leurs effets sociaux et leur fiabilité scientifique. La production, le commerce et l'analyse des données sont souvent faits « parce qu'on peut le faire » et non sur la base de critères solides du point de vue technique et éthique.

La participation sociale et l'importance de ralentir les temps de recherche

L'introduction de procédures explicitement finalisées pour l'identification de tels critères et la réflexion sur leur pertinence à chaque phase de circulation des données est une alternative possible. On entre ici au cœur de la seconde stratégie pour éviter le déterminisme technologique, qui consiste à adopter des processus délibératifs fondés sur la consultation sociale comme support des décisions techniques prises par ceux qui analysent les Big Data. Ce type de consultation peut paraître utopique aux initiés, qui travaillent souvent avec peu de ressources et sous de grandes pressions financières. Et pourtant, la mise en place de procédures permettant un dialogue social plus large sur le traitement des données est un moyen immédiat et constructif d'explorer les implications éthiques et sociales de la recherche, en invoquant l'aide de ceux qui vivent directement ces implications.

Cette leçon a été apprise depuis longtemps par le monde des services digitaux et des réseaux sociaux, où l'avis des utilisateurs est régulièrement requis et utilisé pour améliorer la qualité et l'utilité des technologies en question. Obtenir des données sur les éléments techniques impliqués dans l'analyse des Big Data pose cependant deux autres difficultés : la première concerne le *manque d'incitations* qui pourraient aider à impliquer dans le développement de systèmes digitaux des personnes déjà accablées par d'autres responsabilités ; l'autre consiste à trouver des modes intelligents d'*impliquer* des personnes sans formation scientifique dans des décisions qui peuvent paraître incompréhensibles – hormis pour un ensemble réduit

[83] Lindstrom (2016).
[84] Sur l'idée du rôle fondamental des valeurs éthiques dans la science, voir aussi Douglas (2009), et Elliott *et al.* (2016) dans le cas de la recherche data-centrée.

d'informaticiens. Encore une fois, la vision relationnelle des données peut nous aider à affronter ces obstacles.

Avant tout, la construction de procédures, réglementations et instruments dont le but explicite est d'augmenter le dialogue social sur les systèmes de production, de gestion et d'interprétation des données est une base fondamentale pour prendre des décisions sur ce qui est éthique, dans quelles situations, et pour qui. Le RGPD, c'est-à-dire la législation européenne créée en 2018 pour protéger les citoyens de l'exploitation de leurs données personnelles, est un pas en avant dans ce sens, puisqu'il réclame à celui qui réutilise les données de documenter exactement la manière dont elles sont gérées et de concevoir et maintenir un dialogue entre l'analyste et l'objet des données. En préparation à l'entrée en vigueur de cette législation, beaucoup de gestionnaires des données dans le secteur public et privé ont été contraints de réexaminer les hypothèses et modalités avec lesquelles ils les organisent et les analysent, et de trouver des moyens d'améliorer la communication et le dialogue avec leurs utilisateurs. Cet exercice fastidieux a sûrement pour effet de ralentir et de limiter la production de connaissance à court terme, mais peut aussi fortement en améliorer la qualité et l'impact social à long terme – et démontre comment l'implantation d'un dialogue social sur la gestion des données est bien plus simple à mettre en œuvre durant le processus de construction des banques de données, plutôt que de manière rétroactive.

Dans un deuxième temps, nous avons déjà vu comment même à l'intérieur du monde de la recherche personne n'a une compréhension parfaite et totale des systèmes utilisés pour gérer les données. Nous nous trouvons donc déjà dans une situation où des personnes aux points de vue et aux capacités diverses doivent collaborer pour créer un système qui fonctionne dans son ensemble. Participer n'implique donc pas de rééduquer totalement la population et de transformer tous les citoyens en experts informaticiens, mais plutôt de créer des canaux de communication où les analystes se confrontent à des groupes d'utilisateurs potentiels de manière à examiner et à discuter de leur travail. Cette communication doit être la plus libre possible, afin de faciliter un échange équitable entre public et techniciens, et encourager ainsi les techniciens à modifier leurs systèmes digitaux en tenant compte des exigences et des objections survenues lors du dialogue. Parallèlement, tous les participants travaillent sous des limites techniques et méthodologiques précises, que les experts en ingénierie électronique et programmation doivent communiquer aux autres, et ceux qui ne comprennent rien à ces choses doivent les considérer avec l'esprit ouvert. Il n'y a pas de symétrie dans les échanges de ce genre, ni de garanties que la communication fonctionne bien et produise un apprentissage réciproque – et dans divers moments et situations du développement des Big Data, ces échanges requièrent sûrement des efforts différents de la part de ceux impliqués. Mais l'ouverture au dialogue et à la confrontation sociale la plus large possible demeure fondamentale pour la gestion et l'analyse des Big Data à des fins de recherche.

Un excellent exemple de ce type d'échanges, et de la manière dont ils peuvent contribuer à la qualité de la recherche, est le système avec lequel l'État anglais gère l'expérimentation sur les animaux – un autre secteur où l'application de principes éthiques dans le processus de recherche dépend énormément de la spécificité du cas et est extrêmement controversé du point de vue social (rappelons-nous qu'en Angleterre il existe de nombreux groupes dédiés à la sauvegarde de la vie animale, dont certains ont adopté des techniques d'intimidation et de violence envers les biologistes qui utilisent des animaux pour leurs recherches). Aussi bien l'État que les communautés scientifiques se sont engagées à créer des espaces de dialogue au sein desquels

les chercheurs peuvent discuter des raisons qu'ils ont à utiliser des animaux et recevoir des contributions pour en réduire le nombre et améliorer leur traitement par des moyens compatibles avec les objectifs de leurs projets. En outre, tous les projets qui emploient des animaux sont examinés régulièrement par des inspecteurs du gouvernement qui ont des compétences en biologie et en questions éthiques et légales. La confrontation perpétuelle entre les chercheurs et les inspecteurs n'est pas simplement vouée à implanter des règles déterminées, mais surtout à encourager des réflexions sur la façon dont la recherche se développe et sur les implications possibles pour les animaux utilisés, ainsi que l'élaboration de justifications explicites des choix effectués et des compromis ainsi atteints. Cela aboutit à des moments lors desquels les chercheurs peuvent suspendre temporairement leur frénétique activité de recherche et prendre le temps de s'interroger sur la manière d'améliorer les méthodes et leurs conséquences, par exemple en modifiant le traitement des animaux selon les résultats obtenus jusqu'alors[85].

Ce type d'encouragement à prendre du temps de réflexion et d'évaluation de son propre travail peut sembler banal, mais est en réalité révolutionnaire par rapport aux raisons pour lesquelles la recherche sur les Big Data est habituellement menée et aux structures institutionnelles et économiques dans lesquelles elle prend place. Trop souvent les Big Data sont considérées comme un moyen d'accélérer fortement la production de connaissance au détriment de tout « scrupule ». La construction de procédures qui encouragent la gestion éthique des Big Data peut en revanche aider à améliorer la fiabilité et la qualité de la connaissance produite, sa valeur méthodologique, la responsabilisation des chercheurs impliqués, la durabilité des banques de données impliquées, et inciter à n'utiliser les données qu'à des fins réellement innovantes. Comme dans le cas de la *slow food*, la *slow science* constitue justement une alternative valable au modèle d'utilisation des Big Data qui prévoit une aliénation croissante entre les procédures de recherche et les préférences, exigences et défis qui en caractérisent le contexte social.

La plupart des experts qui travaillent avec les Big Data sont les premiers à souligner la contradiction entre la complexité de la gestion des données et l'espoir qu'elles fournissent une connaissance fiable de manière simple, rapide et socialement acceptable. Il y a des ingénieurs, des informaticiens et des archivistes qui, justement pour cette raison, insistent sur l'adoption d'un *code du comportement* pour la science des données qui, de façon semblable au serment d'Hippocrate effectué par les médecins, encourage ceux qui analysent les Big Data à prendre leurs responsabilités face aux conséquences éventuelles de leurs choix – y compris celle de se confronter aux autres autant que possible afin d'identifier plus efficacement et de manière plus participative quelles peuvent être ces conséquences[86].

Il existe également des institutions créées justement pour défendre les droits des citoyens dont les comportements sont étudiés grâce aux Big Data. La banque de données *Secure anonymized information linkage* (SAIL) au Pays de Galles, par exemple, fut mise en place il y a quinze ans pour conserver et rendre anonymes les données sensibles utilisées dans la recherche médicale, mais a progressivement évolué en un centre capable d'organiser des consultations entre de nombreux types d'experts, de concilier les exigences des patients, médecins et chercheurs, et de conseiller les scientifiques sur le type de recherche à faire sur ces données, et de quelle manière. Cette fonction de médiation sociale a permis à SAIL, simple ressource purement

[85] J'ai détaillé cet exemple et sa pertinence pour la science des données dans Leonelli (2016b).
[86] Plusieurs exemples de ce que ce code pourrait contenir sont donnés par Boyd (2012), Dove *et al.* (2016) et Zook *et al.* (2017) ; voir aussi les suggestions à la fin de ce chapitre.

instrumentale à l'origine, de devenir partie intégrante du processus de recherche[87]. Cette expérience est partagée par beaucoup des banques de données avec qui j'ai travaillé au cours des années, dont la survie et la progressive reconnaissance comme point de référence crucial pour des communautés entières de chercheurs est en grande partie due à leur capacité à faciliter la communication entre les divers secteurs impliqués dans les circulations des données et à fournir des possibilités de comparaison et d'élaboration de solutions communes. L'organisation de la Science Ouverte reconnaît et valorise aussi le rôle de la médiation et de la communication sociale ; et l'institution de consortiums justement voués à favoriser ces échanges a été identifiée comme une composante cruciale pour le développement d'autres formes de Science Ouverte[88].

En outre, l'attention croissante aux questions éthiques sert de stimulation pour le développement de nouveaux types d'algorithmes et de technologies d'analyse, dont l'objectif est justement de permettre l'examen de données potentiellement utiles pour la recherche sans pour autant mobiliser de grandes quantités de données considérées comme sensibles ou privées. Un exemple est la création d'algorithmes et d'accords entre des banques de données qui permettent de réunir et d'analyser certaines caractéristiques des données conservées en différents lieux sans devoir accéder aux banques de données dans leur intégralité. Cela permet d'analyser les données sans nécessairement devoir les partager ou les déplacer d'un endroit à un autre, et de conserver le contrôle sur ceux qui peuvent y avoir accès et sur les objectifs pour lesquels elles peuvent être utilisées[89].

Aucune de ces solutions n'est idéale ou universelle dans ses conséquences, et chacune d'entre elles a plus ou moins de sens selon le type de données et de situation sociale et culturelle dans laquelle ces dernières sont utilisées. Toutes ces solutions sont pourtant des moyens d'encourager la participation sociale à la recherche, et de conceptualiser la recherche elle-même comme un dialogue permanent entre les valeurs et les problèmes sociaux. Il est clair qu'il existe des moments lors des circulations des données – par exemple celles où les nouvelles techniques de programmation et de conservation de celles-ci sont élaborées – où les chercheurs travaillent de manière séparée et indépendante des questions sociales immédiates, et cette indépendance relative joue une fonction importante dans le développement de technologies et de connaissances qui vont au-delà du moment historique lors duquel elles ont été conçues. Parallèlement, pourtant, les chercheurs ont une forte responsabilité vis-à-vis des applications immédiates de leurs résultats, qui comprend l'utilisation des méthodes et des concepts scientifiques développés au cours des siècles afin d'améliorer la qualité et l'intégrité de l'analyse des données ; et tous les autres secteurs sociaux ont la responsabilité de s'intéresser et de s'ouvrir au débat sur les suppositions et les choix avec lesquels sont réalisés les instruments de production de connaissance. Il y a déjà tellement de cas où les personnes qui n'ont ni formation ni rôle professionnel dans le monde de la recherche contribuent de manière décisive à l'analyse des Big Data : il suffit de penser à la création des « *health apps* » qui servent à quantifier notre forme physique ; aux données sur l'environnement et le climat ; aux données produites par les services sociaux et démographiques ; aux données fournies par les patients pour la recherche biomédicale. Il est fondamental de reconnaître que les experts en Big Data ne sont pas seulement ceux qui sont payés pour faire des analyses des données et qui connaissent

[87] Jones *et al.* (2014) ; nous avons déjà analysé l'histoire et les caractéristiques de cette infrastructure dans Tempini et Leonelli (2018).
[88] Voir Vallance *et al.* (2016) et Leonelli (en cours d'impression).
[89] Voir aussi Richards *et al.* (2015) et l'exemple de Data-Schield (Burton *et al*. 2015).

des méthodes statistiques et informatiques. La *data expertise* comprend la connaissance des conditions dans lesquelles les données sont recueillies, la manière dont elles sont recueillies, et les implications de leur utilisation – et cette *expertise* joue un rôle fondamental pour procéder aux choix techniques.

Principes guides pour faciliter la transformation des Big Data en connaissance fiable

Comment peut-on traduire cette analyse en pratique ? Cette section, qui conclut le chapitre, identifie huit « principes guides » pour la gestion pratique des Big Data, chacun desquels ayant une ou plusieurs conséquences pratiques pour différents secteurs sociaux (certainement non exhaustives, mais il s'agit de montrer au moins quelques-unes des ramifications possibles). Ces principes proviennent de l'analyse développée dans ce livre et dans plusieurs autres études, citées ici, de l'impact social des Big Data. Cette liste, bien que grandement simplifiée et incomplète, se veut d'être un point de référence initial pour ceux impliqués dans la création, la dissémination ou la réutilisation des données dans le monde de la recherche ou ailleurs.

Principe 1 : La « donnée » est une catégorie relationnelle.

Il n'y a pas de donnée sans relation : les données, Big Data ou non, sont interprétables seulement d'après un réseau de relations conceptuelles, matérielles et sociales qui doit être rendu explicite de manière à justifier les résultats de l'analyse.

- *Conséquences pratiques* : une gestion des données attentive aux circonstances dans lesquelles elles ont été produites ou à celles dans lesquelles elles sont mobilisées est indispensable à leur réutilisation. L'histoire des données – les étapes de leurs circulations, les matériaux dont elles ont été extraites et les transformations qu'elles ont subies – doit être documentée de manière explicite et facilement accessible par ceux qui souhaitent les analyser.

Principe 2 : L'entretien régulier et à long terme des infrastructures est nécessaire pour justifier la confiance accordée aux Big Data.

L'accumulation et l'interopérabilité des Big Data requièrent un énorme appareil conceptuel, matériel et institutionnel sous la forme d'infrastructures, de banques de données, de réglementations et de programmes de formation adaptés. Appareil qui requiert lui-même des financements spécifiques et considérables destinés à maintenir et à régulièrement mettre à jour cet appareil sur le long terme.

- *Conséquence pratique* : les institutions de recherche doivent travailler avec le gouvernement au niveau national et avec des consortiums et des associations internationales pour développer et maintenir des systèmes efficaces pour le traitement et l'entretien des données (comme par exemple dans le cas de l'*European Open Science Cloud*, avec laquelle de nombreux gouvernements et associations de recherche collaborent dans le sens de la création d'un système fédéral pour la conservation et le traitement de données produites par des financements européens[90]).

[90] Voir le site de la Commission Européenne dédié à l'EOSC : http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

Principe 3 : Des infrastructures et des compétences en gestion des données sont essentielles à l'extraction de connaissance à partir des Big Data.

Il est crucial que les chercheurs impliqués dans l'analyse des Big Data s'intéressent au fonctionnement des banques de données et des algorithmes utilisés pour mobiliser et analyser les données, afin de pouvoir évaluer de manière critique l'incidence de ces instruments, des méthodologies qui y sont liées et des systèmes de classification sur la connaissance extraite des données. Parallèlement, les institutions de recherche doivent reconnaître l'importance d'employer des experts de la gestion et de l'analyse des données qui peuvent aider les chercheurs à composer avec les difficultés inhérentes à leur activité.

- *Conséquence pratique* : les experts des données (ceux qu'on appelle les *data scientists*) doivent être valorisés par les universités et par le monde industriel comme un nouveau type de figure professionnelle indispensable à l'utilisation correcte des Big Data. Les chercheurs de tous les domaines doivent recevoir une éducation minimale pour l'utilisation de technologies, d'infrastructures et de méthodes relatives à l'analyse des Big Data.

Principe 4 : L'espace pour la recherche explorative doit être préservé.

Il est important de ne pas s'engager dans l'utilisation de certains types de données uniquement parce que celles-ci sont plus faciles à produire ou à mobiliser.

- *Conséquence pratique* : l'analyse des Big Data doit être destinée à l'identification d'aires de connaissance pour lesquelles il n'y a pas de données disponibles, et donc stimuler le recyclage des vieilles données autant que la production de nouvelles données.

Principe 5 : La recherche scientifique doit tirer profit d'autant de sources de données que possible, en tenant compte des risques de discrimination et d'inégalité liés à l'utilisation des Big Data.

La triangulation des Big Data aide à produire de la connaissance plus fiable seulement dans les cas où les données utilisées ont des histoires différentes et séparées (c'est-à-dire, comme discuté dans le chapitre 3, qui proviennent de lignées diverses). Dans le cas où cela ne serait pas possible, les chercheurs doivent mettre en évidence les manières dont différents types de données peuvent ou non être comparés, en s'inspirant des méthodes déjà bien développées dans les domaines scientifiques de référence.

- *Conséquence* : le choix des données à insérer dans une banque de données doit être documenté de manière explicite et avisée par des méthodes d'échantillonnage déjà utilisées depuis longtemps par les sciences naturelles et sociales.

Principe 6 : L'éthique, la sécurité et la responsabilité sociale sont partie intégrante de la recherche data-centrée.

Il n'existe aucun moyen de séparer la valeur éthique des données de la valeur épistémologique et scientifique et, puisque l'analyse des données peut être réglementée et standardisée, celui qui se charge d'organiser et d'analyser les Big Data demeure responsable de la manière dont les règles sont appliquées dans chaque cas précis. Les choix fondamentaux effectués lors de l'extraction de connaissance des Big Data ont des implications décisives pour le niveau de conservatisme, de fiabilité, de partialité, de corruption et de sensibilité sociale présent dans les processus et les résultats de recherche. Dans la mesure où elles sont réparties entre de

nombreuses personnes aux compétences différentes et hautement spécialisées, toutes les phases du travail associé aux circulations des données – y compris leur planification, leur création, leur mobilisation et leur analyse – impliquent donc une certaine responsabilité vis-à-vis des conséquences sociales qu'auront les instruments et la connaissance ainsi obtenus.

- *Conséquence pratique* : chaque élément de la gestion des Big Data doit être évalué aussi bien pour sa valeur technique que pour ses implications éthiques et sociales, comme indiqué par les nombreux codes de comportement pour *data scientists* créés récemment pour guider leur travail.

Principe 7 : L'utilisation des Big Data à des fins de recherche est liée au dialogue social sur les hypothèses utilisées pour les analyser dans divers contextes d'application.

La participation citoyenne est un prérequis essentiel à la dissémination et à l'utilisation des Big Data pour produire de la connaissance scientifique, étant donné qu'elle aide les chercheurs à identifier des problèmes potentiels en lien avec des situations de réutilisation spécifiques, et à trouver des solutions adaptées.

- *Conséquence pratique* : une ou plusieurs formes de consultation et de dialogue social doivent être intégrées au sein des procédures de construction et de maintien d'infrastructures et de techniques d'analyse dédiées aux Big Data. Des institutions tournées vers la médiation entre différents utilisateurs des Big Data peuvent être extrêmement utiles à la gestion des données elles-mêmes comme à la production de connaissance fiable et appropriée au contexte social.

Principe 8 : Il est fondamental que chaque secteur social impliqué dans l'utilisation de connaissances et de technologies provenant de l'analyse des Big Data s'intéresse au fondement empirique de ces connaissances et ait les instruments nécessaires pour interagir techniquement avec les choix effectués.

- *Conséquence pratique* : la création de formes de dialogue et de participation sociale dans les processus scientifiques doit être évaluée et promue par les gouvernements, les universités et financeurs, en tant que partie importante de la recherche, ce qui améliorera ainsi les incitations pour impliquer les scientifiques à ces efforts. Le programme scolaire national doit comprendre, pour chaque niveau (des écoles élémentaires à l'enseignement supérieur), des discussions et de l'éducation pratique sur la gestion et l'interprétation des données, tandis que l'État et la société civile doivent s'engager à organiser des cours de rattrapage et des occasions de dialogue pour les adultes.

Conclusion

Cet ouvrage a cherché à démontrer comment ce qui est considéré comme une donnée, et la manière dont les Big Data génèrent de la connaissance, dépend des technologies impliquées dans la production, la mobilisation et l'analyse des données – et donc des décisions et des choix effectués par les individus responsables de la gestion des données, des ressources économiques et sociales à leur disposition, ainsi que des incitations et des objectifs pour lesquels les données sont recueillies et analysées. J'ai souhaité transmettre deux messages fondamentaux. Le premier est que le pluralisme et la variabilité du type de connaissance et des méthodes utilisées dans le monde de la recherche sont précieux, et doivent être reconnus et exploités plutôt qu'éliminés en tant que source de complications. Le second est que l'impact des Big Data sur la recherche et sur le futur de la connaissance humaine dépend de la manière dont nous tous, et particulièrement ceux qui s'occupent des données dans leur travail, nous confrontons à trois environnements strictement liés entre eux :

1. La gestion des banques de données : comment et par qui sont gérées les infrastructures responsables de la conservation et du classement des données, quelles modalités d'accès aux données sont préférables, et quels types de données (et d'informations qui en découlent) devraient être recueillies et réutilisées ?
2. La fiabilité des données : qui juge de la qualité des données et des technologies utilisées pour les analyser, comment, et à quelle étape du processus de recherche et d'interprétation ? Comment pouvons-nous nous assurer que les données disponibles en ligne sont gérées correctement et qu'il existe des solutions durables pour mettre à jour les collections et les algorithmes actuellement existants ?
3. La participation et l'interaction avec les données : comment peut-on organiser un dialogue entre tous ceux intéressés par la gestion et l'utilisation des données, ou par la connaissance qui en est tirée ? Quel rôle ont les différents types de compétences et d'*expertises* impliquées dans l'interprétation des données, et comment peuvent-elles interagir de manière constructive ? Qui est exclu, exploité ou persécuté par l'utilisation des Big Data, et que peut-on faire pour limiter l'impact négatif qu'il subit ?

Passer du temps à réfléchir à ces questions peut être perçu par les chercheurs comme une perte de temps précieux et un obstacle supplémentaire à des études déjà très complexes et dont la résolution est difficile. Certains chercheurs sont aussi – et de manière compréhensible – préoccupés par le fait qu'un poids excessif conféré à des considérations éthiques constitue un obstacle à la production de certaines formes de connaissance, car cela ajoute d'autres barrières administratives à des processus de recherche déjà alourdis par la bureaucratie (par exemple lorsque les chercheurs doivent demander l'autorisation pour réutiliser des données personnelles dans de nouveaux contextes de recherche), ou bien parce que certaines des manières dont les données sont gérées en réduisent la capacité à être utilisées comme source de connaissance (par exemple quand les procédures d'« anonymisation » réduisent fortement la granularité des données produites par la recherche sur des sujets humains). La protection des données personnelles peut être assez problématique, par exemple dans les cas de recherches effectuées dans plusieurs lieux ou cultures différents, et décourager les chercheurs à se poser certains types de questions, ce qui rend ainsi impossibles certains types de recherche et donc, de connaissance.

Ce sont des faiblesses considérables de l'approche que j'ai proposée, mais il est impossible de les éviter sans encourir le risque bien plus grand de produire de la connaissance douteuse et dommageable, conformément à ce qui a été décrit dans le deuxième chapitre. De plus, nous avons vu que certaines de ces faiblesses sont liées à des opportunités d'améliorer à la fois les procédures et les résultats de la recherche, en favorisant un ralentissement du rythme de la production scientifique au bénéfice de sa qualité éthique et scientifique. Prêter une plus grande attention et fournir plus d'efforts pour déterminer quelles données sont plus appropriées à un type de projet déterminé, et comment elles doivent être traitées afin d'être réutilisées par d'autres aux intérêts différents, stimule la création de mécanismes qui rendent les banques de données plus durables ; améliore la qualité et la fiabilité des données elles-mêmes ; et reconfigure les relations entre ceux qui sont impliqués dans la recherche, comme les relations qui existent par exemple entre les patients et les chercheurs cliniques, de manière à rendre le processus d'enquête plus ouvert à l'intégration de sources de connaissance extérieures à celles reconnues traditionnellement comme « scientifiques », et donc potentiellement plus réfléchi et pertinent pour les futures implications possibles de l'utilisation des Big Data.

Bibliographie

Anderson C., *The end of theory. The data deluge makes the scientific method obsolete*, in « Wired », juin 2008. URL: https://www.wired.com/2008/06/pb-theory/.

Ankeny, R.A. *The overlooked role of cases in causal attribution in medicine* in « Philosophy of Science » *81*(5), 2014, pp.999-1011.

Ankeny R.A., Leonelli S., *Repertoires : A Post-Kuhnian Perspective on Scientific Change and Collaborative Research*, in « Studies in the History and the Philosophy of Science : Part A », 60, 2016, pp.18-28.

Aronova E., van Hoerzen C., Sepkoski D., *Introduction : Historicizing Big Data*, « Osiris », 32 (1), 2018, pp.1-17.

Barnes B., Dupré J., *Genomes and What to Make of Them*, University of Chicago Press 2008.

Beer D., *Metric Power*, Palgrave Macmillan, Basingstoke 2016.

Bezuidenhout L., Leonelli S., Kelly A., Rappert B., *Beyond the Digital Divide. Towards a Situated Approach to Open Data*, in « Science and Public Policy », 44, n°4, 2017, pp.464-475.

Bezuidenhout L., Kelly A., Leonelli S., Rappert B., « *$100 Is Not Much To You* » *: Open Science and Neglected Accessibilities for Scientific Research in Africa*, in « Critical Public Health », 2016, pp.1-11.

Bogen J., Woodward J., *Saving the Phenomena*, in « *The Philosophical Review* » 97(3), 1988, pp.303-352.

Borgman C.L., *Big Data, Little Data, No Data*, MIT Press, Cambridge 2015.

Boulton G., Campbell P., Collins B. *et al.*, *Science as an open enterprise. The Royal Society Science Policy Centre Report 02/12*, The Royal Society Publishing, London 2012.

Boumans M., Leonelli S., *From Dirty Data to Tidy Facts : Practices of Clustering in Plant Phenomics and Business Cycles*, in Leonelli S., Tempini N. (dir.), *Varieties of Data Journeys*, 2019.

Bowker G.C., *Science on the run : information management and industrial science at Schlumberger, 1920-1940*, MIT Press, Cambridge 1994.

Bowker G.C., Star S.L., *Sorting Things Out*, MIT Press, Cambridge 1999.

Boyd D., *Critical Questions for Big Data*, in « Information, Communication & Society », 4462, June 2012, pp.37-41.

Boyd D., Crawford K., *Six Provocations for Big Data*, in « Data & Society Paper » 2011.

Broadbent A., *Philosophy of Epidemiology*, Palgrave Macmillan, Basingstoke 2013.

Bokulich A., *Using Models to Correct Data : Paleodiversity and the Fossil Record*, in « Synthese », 2018.

Burton P.R., Murtagh M.J., Boyd A. *et al.*, *Data Safe Havens in Health Research and Healthcare*, in « Bioinformatics », 31, n°20, 2015, pp.3241-3248.

Cai L., Zhu Y., *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*, in « Data Science Journal », 14, p. 2, 2015.

Cambrosio A., Keating P., Nelson N., *Régimes Thérapeutiques et Dispositifs de Preuve en Oncologie. L'organisation des Essais Cliniques, des Groupes Coopérateurs aux Consortiums de Recherche*, in « Sciences Sociales et Santé », 32, 2014, pp.13-42.

Canali S., *Big Data, Epistemology and Causality. Knowledge in and Knowledge out in EXPOsOMICS*, in « Big Data & Society », 3, n°2, 2016, pp.1-11.

Caporael L.R., Griesemer J.R., Wimsatt W.C., *Developing Scaffolds in Evolution, Culture and Cognition*, MIT Press, Cambridge 2014.

Chang H., *Is Water H2O?*, Springer Netherlands, Dordrecht 2012.

Chang H., *Pragmatist Coherence as the Source of Truth and Reality*, in « Proceedings of the Aristotelian Society », CXVII, n°2, 2017.

Chapman R., Wylie A. (dir.), *Material Evidence. Learning from Archaeological Practice*, Routledge, Oxon and New York 2015.

Curioni A., *La Protezione dei Dati. Guida Pratica al Regolamento Europeo*, Mimesis Editore, Milano 2017.

Daston L., *The Moral Economy of Science,* in « Osiris », 10, 1995, pp.2-24.

Derrida J., *Mal D'Archive : Une Impression Freudienne*, Éditions Galilée, Paris 1995.

Directorate-General for Research and Innovation (European Commission), *Incentives and Rewards to Engage in Open Science Activities. Thematic Report No. 3 for the Mutual Learning Exercise Open Science : Altmetrics and Rewards*, Publications Office of the European Union, Luxembourg 2017.

Directorate-General for Research and Innovation (European Commission), *Implementing Open Science. Strategies, Experiences and Models. Thematic Report No. 4 for the Mutual Learning Exercise on Open Science : Altmetrics and Rewards,* Publications Office of the European Union, Luxembourg 2018.

Douglas H., *Science, Policy and the Value-Free Ideal*, University of Pittsburgh Press, Pittsburgh 2009.

Dove E.S., David T., Meslin E. *et al.*, *Ethics Review for International Data-Intensive Research*, in « Science », 351, n°6280, 2016, pp.1399-1400.

Dupré J., *The Disorder of Things. Metaphysical Foundations of the Disunity of Science*, Harvard University Press, Cambridge et Londres 1983.

Ebeling M.F.E., *Healthcare and Big Data. Digital Specters and Phantom Objects*, Palgrave Macmillan, New York 2016.

Edwards P.N., *A vast machine : Computer models, climate data, and the politics of global warming*, MIT Press, Cambridge 2010.

Edwards P.N., Mayernik M.S., Batcheller A.L. *et al.*, *Science Friction. Data, metadata, and collaboration*, in « Social Studies of Science », 41, n°5, 2011, pp.667-690.

Elliott K.C., Cheruvelil K.S., Montgomery G.M., Soranno P.A., *Conceptions of Good Science in Our Data-Rich World*, in « BioScience », 66, n°10, 2016, pp.880-889.

Fecher B., Friesike S., Hebing M., *What Drives Academic Data Sharing ?*, in « PLoS ONE », 10, n°2, e0118053, 2015.

Feest U., *What Exactly is Stabilized When Phenomena are Stabilized ?*, in « Synthese » 182(1), 2011, pp.57-71.

Floridi L., *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, trad. it. M. Durante, Raffaello Cortina Editore, Milan 2017.

Floridi L., Illari P. (dir.), *The Philosophy of Information Quality. Synthese Library 358*, Springer, Cham Heidelberg, New York Dordrecht London 2014.

Foucault M., *Le Parole e le cose*, trad. it. E. Panaitescu, Rizzoli, Milan 1967.

Gitelman L., *« Raw data » is an Oxymoron*, MIT Press, Cambridge 2013.

Harris A., Kelly S., Wyatt S., *CyberGenetics. Health Genetics and New Media*, Routledge/Taylor & Francis Group, Londres 2016.

Hey T., Tansley S., Tolle K., *The fourth paradigm. Dataintensive scientific discovery*, Microsoft Research, Redmond 2009.

Hilgartner S., *Constituting large-scale biology. Building a regime of governance in the early years of the Human Genome Project*, in « BioSocieties », 8, 2013, pp.397-416.

Hine C., *Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work*, in « Social Studies of Science », 36, n°2, 2006, pp.269-298.

Jones K., Ford D.V., Jones C. *et al.*, *A Case Study of the Secure Anonymous Information Linkage (SAIL) Gateway. A Privacy-Protecting Remote Access System for Health-Related*

Research and Evaluation, in « Journal of Biomedical Informatics », 50 (Special Issue on Informatics Methods in Medical Privacy), August, 2014, pp.196-204.

Kellert S.H., Logino H.E., Waters C.K. (dir.), *Scientific Pluralism*, University of Minnesota Press, Minneapolis 2006.

Kitchin R., *The data revolution. Big data, open data, data infrastructures and their consequences*, SAGE, Londres 2014.

Kitchin R., McArdle G., *What Makes Big Data, Big Data ? Exploring the Ontological Characteristics of 26 Datasets*, in « Big Data & Society », 3, n°1, 2016, pp.1-10.

Landecker H., *Culturing Life : How Cells Became Technologies*, Harvard University Press, Cambridge 2007.

Latour B., *Science in Action : How to Follow Scientists and Engineers through Society*, Harvard University Press, Cambridge 1987.

Leonelli S., *When Humans Are the Exception. Cross-Species Databases at the Interface of Clinical and Biological Research*, in « Social Studies of Science », 42, n°2, 2012, pp.214-236.

Leonelli S., *What Difference Does Quantity Make ? On the Epistemology of Big Data in Biology*, in « Big Data and Society », 1, 2014, pp.1-11.

Leonelli S., *Data-Centric Biology : A Philosophical Study*, Chicago University Press, Chicago 2016a.

Leonelli S., *Locating ethics in data science : responsibility and accountability in global and distributed knowledge production*, in « Philosophical Transactions of the Royal Society: Part A », 374, n°2083, 2016b, 20160122.

Leonelli S., *Biomedical Knowledge Production in the Age of Big Data. Analysis conducted on behalf of the Swiss Science and Innovation Council SSIC*, Bern 2017a, URL : https://www.swir.ch/it/pubblicazioni.

Leonelli S., *Global Data Quality Assessment and the Situated Nature of « Best » Research Practices in Biology*, in « Data Science Journal », 16, n°32, 2017b, pp.1-11.

Leonelli S., *The Time of Data. Time-Scales of Data Use in the Life Sciences*, in « Philosophy of Science », 85 (5), 2018.

Leonelli S., *Scientific Agency and Social Scaffolding in Contemporary Data-Intensive Biology*, in Wimsatt W., Love A.C. (dir.), *Beyond the Meme. Articulating Dynamic Structures in Cultural Evolution*, University of Minnesota Press, Minneapolis 2019.

Leonelli S., en cours d'impression. *What Distinguishes Data from Models ?*

Leonelli S., Diehl A.D., Christie K.R., Harris M.A., Lomax J., *How the Gene Ontology Evolves*, in « BMC Bioinformatics », 12, 2011.

Leonelli S., Tempini N., *Where Health and Environment Meet. The Use of Invariant Parameters in Big Data Analysis*, in « Synthese », Special issue on Philosophy of Epidemiology, 2018.

Levin N., Leonelli S., *How Does One « Open » Science ? Questions of Value in Biological Research*, in « Science, Technology and Human Values », 42, n°2, 2016, pp.280-305.

Levin N., Leonelli S., Weckowska D. *et al.*, *How Do Scientists Understand Openness ? Exploring the Relationship between Open Science Policies and Research Practice*, in « Bulletin for Science and Technology Studies », 36, n°2, 2016, pp.128-141.

Lindstrom M., *Small Data: The Tiny Clues that Uncover Huge Trends*, St Martin's Press, New York 2016.

Loettgers A., *Synthetic Biology and the Emergence of a Dual Meaning of Noise*, in « Biological Theory », 4, n°4, 2009, pp.340-355.

Marr B., *Big Data. Using SMART big data, analytics and metrics to take better decisions and improve performance*, John Wiley & Sons, Hoboken 2015.

Massimi M., *From Data to Phenomena : A Kantian Stance*, in « Synthese » 182(1), 2011, pp.101-116.

Mauthner N.S., Parry O., *Open Access Digital Data Sharing : Principles, Policies and Practices*, in « Social Epistemology », 27, n°1, 2013, pp.47-67.

Mayer-Schönberger V., Cukier K., *Big Data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, trad. it. R. Merlini, Garzanti, Milan 2013.

Mayo D., *Error and the Growth of Experimental Knowledge,* University of Chicago Press, Chicago 1996.

McAllister J.W., *Model Selection and the Multiplicity of Patterns in Empirical Data*, in « Philosophy of Science », 74, n°5, 2007, pp.884-894.

Mittelstadt B.D., Floridi L. (dir.), *The Ethics of Biomedical Big Data*, Springer Switzerland, Basel 2016.

Mongilli A., Pellegrino G. (dir.), *Information Infrastructure(s). Boundaries, Ecologies, Multiplicity*, Cambridge Scholars Publishing, Cambridge 2014.

Müller-Wille S., Charmantier I., *Natural history and information overload. The case of Linnaeus*, in « Studies in History and Philosophy of Science Part C », 43, 2012, pp.4-15.

Müller-Wille S., Rheinberger H., *A Cultural History of Heredity*, University of Chicago Press, Chicago 2012.

Murphy M., *The Economization of Life*, Duke University Press, Durham 2017.

Normandeau K., *Beyond volume, variety and velocity is the issue of big data veracity*, « Inside Big Data », 2013, URL : http://insidebigdata.com/2013/09/12/beyondvolume-variety-velocity-issue-big-data-veracity.

November J., *Biomedical Computing : Digitazing Life in the United States*, The John Hopkins University Press, Baltimora 2012.

Oreskes N., Conway E.M., *Merchants of Doubt*, Bloomsbury Press, Londres 2010.

Ossorio P., *Bodies of data : Genomic data and bioscience data sharing*, in « Social Research », 78, n°3, 2011, pp.907-932.

Pestre D., *Regimes of knowledge production in society. Towards a more political and social reading*, in « Minerva », 41, 2003, pp.245-261.

Plan National pour la Science Ouverte, 2018, URL : http://www.enseignementsup-recherche.gouv.fr/cid132531/plan-national-pour-la-science-ouverte-discours-de-frederique-vidal.html

Prainsack B., *Personalised Medicine. Empowered Patients in the 21st Century ?*, New York University Press, New York 2017.

B. Prainsack, A. Buyx, *Solidarity in Biomedicine and Beyond*, Cambridge University Press, Cambridge 2017.

Rappert B., Selgelid M.J., *On the Dual Uses of Science and Ethics : Principles, Practices and Prospects*, ANU Press, Canberra 2013.

de Regt H.W., *Understanding Scientific Understanding*, Oxford University Press, Oxford 2017.

de Regt H.W., Leonelli S., Eigner K., *Scientific Understanding : Philosophical Perspectives*, University of Pittsburgh Press, Pittsburgh 2009.

Rheinberger H., *Infra-Experimentality : From Traces to Data, From Data to Patterning Facts*, in « History of Science » 49(164), 2011, pp.337-348.

Ribes D., Polk J.B., *Organizing for Ontological Change. The Kernel of a Research Infrastructure*, in « Social Studies of Science », 45, n°2, 2015, pp.214-241.

Richards M., Anderson R., Hinde S. *et al.*, *The collection, linking and use of data in biomedical research and health care. Ethical issues*, Nuffield Council on Bioethics, Londres 2015.

Sætnan A.R., Schneider I., Green S. (dir.), *The Policy and Politics of Big Data*, Routledge, Oxon 2018.

Sansone S.A., Rocca-Serra P., Field D. *et al.*, *Toward Interoperable Bioscience Data*, in « Nature Genetics », 44, n°2, 2012, pp.121-126.

Science International, *Open data in a big data world. An international accord*, ICSU, ISCC TWAS, & IAP, Paris 2015.

Shavit A., Griesemer J.R., *There and back again, or the problem of locality in biodiversity surveys*, in « Philosophy of Science », 76, 2011, pp.273-294.

Srnicek N., *Platform Capitalism*, Polity Press, Cambridge and Malden 2017.

Stevens H., *Life out of Sequence. A Data-Driven History of Bioinformatics*, University of Chicago Press, Chicago (IL) 2013.

Strasser B.J., *The Experimenter's Museum : GenBank, Natural History, and the Moral Economies of Biomedicine, 1979-1982*, in « Isis », 102, 2011, pp.60-96.

Strasser B.J., Edwards P., *Big Data is the Answer… But What is the Question ?*, in « Osiris » 32 (1), 2017, pp.328-345.

Sunder Rajan K., *Pharmocracy. Value, Politics, and Knowledge in Global Medicine. Duke University Press*, Durham 2017.

Suppes P., *Models of data*, in Nagel E., Suppes P., Tarski A. (dir.), *Logic, methodology and philosophy of science*, Stanford University Press, Stanford, 1962.

Teller P., *Saving the Phenomena Today*, in « Philosophy of Science »" 77(5), 2010, pp.815-826.

Tempini N., *Till Data Do Us Part. Understanding Data-based Value Creation in Data-Intensive Infrastructures*, in « Information & Organization », 27, 2017, pp.191-210.

Tempini N., Leonelli S., *Concealment and Discovery : The Role of Information Security in Biomedical Data Re-Use*, « Social Studies of Science », en cours d'impression.

Timmermans S., Epstein S., *A World full of Standards but not a Standard World : Toward a Sociology of Standardization*, in « Annual Review of Sociology », 36, 2010, pp.69-89.

Thrift N., *Knowing capitalism*, SAGE, Londres 2005.

Vallance P., Freeman A., Stewart M., *Data Sharing as Part of the Normal Scientific Process. A View from the Pharmaceutical Industry*, in « PLoS Medicine », 13, n°1, e1001936, 2016.

Vayena E., Tasioulas J., *The Dynamics of Big Data and Human Rights. The Case of Scientific Research*, in « Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences », 374, n°2083, id 20160129, 2016.

Vermeir K., Leonelli S., Shams Bin Tariq A. *et al.*, *Global Access to Research Software : The Forgotten Pillar of Open Science Implementation. A Global Young Academy Report*, Global Young Academy, Halle 2018.

Ward J.S., Barker A., *Undefined by data. A survey of big data definitions*, School of Computer Science at the University of St Andrews, St. Andrews 2013.

Wilkinson M.D. *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*, in « Scientific Data » 3:160018 doi: 10.1038/sdata.2016.18, 2016.

Woodward J., *Data, Phenomena, Signal, and Noise*, in « Philosophy of Science », 77, n°5, 2010, pp.792-803.

Wouters P., Beaulieu A., Scharnhorst A., Wyatt S. (dir.), *Virtual Knowledge. Experimenting in the Humanities and the Social Sciences*, MIT Press, Cambridge 2013.

Wylie A., *Thinking from Things. Essays in the Philosophy of Archaeology*, University of California Press, Berkeley 2002.

Wylie A., *How Archaeological Evidence Bites Back. Strategies for Putting Old Data to Work in New Ways*, in « Science, Technology, and Human Values », 42, n°2, 2017, pp.203-225.

Zook M., Barocas S., Boyd D. *et al.*, *Ten Simple Rules for Responsible Big Data Research*, in « PLoS Computational Biology », 13, n°3, e1005399, 2017.

Remerciements

Les recherches qui ont mené à cet ouvrage ont été financées par les institutions publiques suivantes, que je remercie vivement pour leur soutien : European Research Council (DATA_SCIENCE grant award 335925, « The Epistemology of Data-Intensive Science »), Leverhulme Trust (award RPG-2013-153), Australian Research Council (award DP160102989), U.K. Medical Research Council and Natural Environment Research Council (award MR/K019341/1) et U.K. Economic and Social Research Council (award ES/P011489/1).

Giovanni Boniolo m'a donné l'idée d'écrire ce livre, qui n'existerait pas sans ses encouragements et son talent éditorial. Koen Vermeir et Nicolas Filicic m'ont aidée à trouver ma traductrice Faustine Galicia, qui a travaillé sans relâche pour faciliter la traduction rapide de ce livre. Je suis reconnaissant d'avoir travaillé avec elle. Le Centre de Logique et de Philosophie des Sciences de l'Université de Ghent en Belgique, qui m'a accueillie au printemps 2018, m'a fourni un formidable espace de travail pour rédiger le manuscrit – et sans le café du bar Emmy, je n'aurais sans doute pas pu le finir dans les temps. Michel Durinx m'a aidée à concevoir les figures et la bibliographie avec une efficacité et une disponibilité considérable. Malheureusement, il est difficile de remercier convenablement la centaine de scientifiques et de collègues en philosophie, sociologie, anthropologie et histoire des sciences avec lesquels j'ai eu le privilège de discuter ces idées au cours des dix dernières années. Les personnes suivantes méritent d'être mentionnées (par ordre alphabétique) pour le rôle fondamental qu'elles ont joué dans mes recherches sur les données au sein du projet « DATA_SCIENCE » : Rachel Ankeny, Elizabeth Arnaud, Ruth Bastow, Bill Bechtel, Louise Bezuidenhout, Marcel Boumans, Alberto Cambrosio, Hasok Chang, Adrian Currie, Gail Davies, John Dupré, Lora Fleming, Luciano Floridi, Jean-Paul Gaudillière, James Griesemer, Gregor Halfmann, Mary Morgan, Rebecca Lovell, Staffan Müller-Wille, Barbara Prainsack, Hans Radder, Ed Ramsden, Brian Rappert, Hans-Jörg Rheinberger, Beckett Sterner, Kaushik Sunder Rajan, David Teira, Niccolò Tempini, Sally Wyatt et Alison Wylie.

L'affection de mes amis continue à me donner force et inspiration chaque jour - et je dois tout au soutien de ma famille et à la patience et à l'amour de Leonardo, Luna et Michel.