Les données : le point aveugle de la littérature sur la simulation

Fabrizio Li Vigni

Les données : le point aveugle de la littérature sur la simulation

Les données : le point aveugle de la littérature sur la simulation

Cet article fait partie de l'atelier « Philosophie et numérique », qui poursuit les réflexions du dossier du même nom. Les essais et articles qui le composent constituent davantage des états de l'art ou des bilans sur un « work in progress ».

[box type= »bio »] Par Fabrizio Li Vigni. [/box]

[box type= »info »] **Résumé**

Cet article se penche sur un point aveugle des études philosophiques de la modélisation et simulation numérique, à savoir le travail sur les données, et contribue à déconstruire l'idée répandue selon laquelle les Big data seraient des agrégats informationnels qui vont d'eux-mêmes. En décrivant les phases du travail des scientifiques issus d'un domaine de recherche interdisciplinaire appelé « sciences des systèmes complexes », nous tâcherons de montrer que les opérations de collecte et préparation des données prennent plus de temps et d'énergie que les opérations sur lesquelles on a l'habitude de s'attarder davantage dans la littérature, à savoir la construction des hypothèses du modèle, la programmation du code informatique et la validation des output de celui-ci. Cette recherche se base sur une quarantaine d'entretiens et une dizaine de visites de laboratoire.

Mots-clés : données, sciences computationnelles, systèmes complexes, modélisation, simulation

Abstract

This article tackles a blind spot of philosophical studies of digital modeling and simulation, namely the work on data, and it contributes to the deconstruction of the common idea according to which Big data are self-evident informational aggregates. By describing the phases of the work of some scientists inscribed in the interdisciplinary research domain called “complex systems sciences”, we will try to show that the operations of data collection and preparation take more time and energy than the operations which it is usual to tackle the most in the literature, that is the construction of the hypotheses of the model, the coding of the software, and the validation of its outputs. This research is based on some forty interviews and a dozen laboratory visits.

Keywords: data, computational sciences, complex systems, modeling, simulation[/box]

Introduction

De toutes les sciences humaines et sociales, la philosophie des sciences est celle qui a le plus investi la réflexion sur la digitalisation des pratiques scientifiques[1]. Mais dans la vaste majorité des cas, ces travaux se sont occupés uniquement du modèle informatique : de son statut ontologique, de sa validité épistémique et de son usage pratique.

Cet article adresse un point aveugle des études épistémologiques de la modélisation et de la simulation numériques, à savoir le travail sur les données, et contribue à déconstruire l'idée répandue selon laquelle les *Big data* seraient des agrégats informationnels qui vont d'eux-mêmes. En effet, les données ne sont jamais déjà-là, déterminées, sans ultérieure intervention – elles sont au contraire filtrées, épurées, harmonisées, avant de pouvoir être utilisées.

En décrivant deux phases de l'activité de modélisation digitale – la collecte et la préparation des données –, à partir d'un travail de terrain mené sur un domaine de recherche interdisciplinaire appelé « sciences des systèmes complexes »[2], nous tâcherons de montrer que les opérations relatives aux données prennent plus de temps et d'énergie que les opérations sur lesquelles on a l'habitude de s'attarder dans la littérature. Nous soutenons que, en même temps que penser la structure et l'usage des simulations informatiques à visée de recherche, il est essentiel d'interroger le travail artisanal que les scientifiques doivent mener sur la matière qu'ils administrent aux simulations[3]. Ce questionnement présente au moins deux volets : d'une part, la description des pratiques liées à l'artisanat des données, qui est l'objet de ce texte ; d'autre part, la réflexion épistémologique, ontologique et axiologique sur les données – thématique qui reste en dehors du périmètre de cet article, mais à laquelle il invite à réfléchir.

I. Problématique et terrain

La littérature évoquée ci-dessus se distingue, nous semble-t-il, en deux grandes catégories : la réflexion pluri-thématique sur les *Big data* et la réflexion onto-épistémique sur la modélisation numérique. Quant au premier volet, la philosophie des sciences s'est occupée, parmi d'autres choses, de déconstruire le fameux article controversé « The End of Theory » de Chris Anderson[4], en montrant par exemple que les données ne sont justement pas « données »[5], que le travail de recherche inductif n'est jamais exempté d'hypothèses préalables[6] et que les corrélations dans le déluge des données sont souvent fallacieuses[7], ou bien elle s'est attardée sur l'usage des données pour corriger les modèles[8], sur la capacité prédictive des *Big data*[9] ou sur l'éthique liée à leur

utilisation[10]. Quant au deuxième volet, les philosophes ont surtout réfléchi à la manière de qualifier la simulation numérique, qui est le plus souvent vue comme une nouvelle voie au-delà de la théorie et de l'expérience[11], nécessitant une toute nouvelle épistémologie pour être pensée[12]. Ou bien, ils se sont penchés sur la validité de la simulation et sur son statut par rapport à l'expérience *in vivo* dans les laboratoires ou en plein air[13].

Face à cette profusion de travaux, il est légitime de se demander si la simulation, entendue ici comme le processus de codage informatique, est effectivement au centre des pratiques en sciences computationnelles comme cela semble ressortir des préoccupations des philosophes. Est-ce que la modélisation numérique occupe la plupart du temps de travail des scientifiques ou y a-t-il d'autres aspects, tout aussi importants, qui sont restés sous les radars de la réflexion épistémologique ? Il se trouve que dès que l'on entre dans un laboratoire qui fait usage de l'ordinateur pour enquêter sur ses objets d'étude, la plupart des choses que l'on entend dire aux scientifiques portent sur la collecte, la construction, la reconstruction, la standardisation, l'épurement et l'analyse des données – et assez peu sur la simulation. Le témoignage d'un épidémiologiste computationnel rattaché au Los Alamos National Laboratory (Los Alamos, NM, USA) résume bien ce qui ressort de notre terrain auprès des spécialistes des systèmes complexes en général :

> It's a huge pain. The rule of thumb in the data science world is that you spend like 85 to 90% of your time dealing with just the data. Cleaning the data, putting it into a form that's useful, and then the rest of it is running your model[14].

À cette réalité observée sur le terrain correspond une littérature primaire également engagée dans des questions relatives au traitement des données[15]. Pourtant, les textes philosophiques, mais aussi sociologiques et historiques, reviennent très peu sur cette préoccupation quotidienne des praticiens des *Big data*, même si c'est sur ce terrain-là que se jouent l'articulation et l'inscription des savoirs acquis, des ontologies, des hypothèses du modèle et des équations mathématiques qui traduisent ces dernières. La suite de ce texte est vouée à illustrer la citation de notre interlocuteur du LANL, en proposant une brève description des phases relatives à la fabrication des données dans le processus de modélisation.

Les « sciences des systèmes complexes » sont apparues dans les années 1970 en Europe et aux États-Unis et ont été consacrées sur le plan international par le Santa Fe Institute dans les années 1980[16]. Elles regroupent un ensemble de sciences naturelles, sociales et de l'ingénieur se proposant d'étudier les « systèmes complexes » : ensembles composés de nombreux éléments hétérogènes en interaction, donnant lieu à des propriétés émergentes (exemples : réseaux génétiques, écosystèmes, villes, marchés financiers, épidémies, etc.). Les représentants de ce domaine se proposent de les étudier par des outils informatiques, mathématiques et physiques comme la théorie des réseaux, les modèles à base d'agents, les automates cellulaires ou la physique statistique[17]. Le matériau sur lequel nous nous appuyons ici se constitue d'une quarantaine d'entretiens menés avec des épidémiologistes computationnels[18] et des géographes quantitativistes[19], et d'une dizaine de courtes ethnographies et visites de laboratoires[20].

Dans nos interviews, les scientifiques décrivent de nombreuses phases du travail de modélisation : la collecte des données, la préparation des données, l'analyse statistique des données, la formulation des hypothèses, la modélisation (écriture des hypothèses et des équations), la simulation (codage informatique du modèle), la calibration du modèle, la visualisation des données, l'analyse des données de simulation, la validation du modèle, etc. Si les noms et l'ordre de ces phases peuvent varier selon la discipline des scientifiques rencontrés, ces derniers mettent tous en garde contre une séparation nette et chronologique entre ces différents moments de leur travail, qui en réalité se chevauchent souvent, ou se précèdent et suivent selon des logiques contre-intuitives, contextuelles et aléatoires.

II. La collecte des données

La collecte des données apparaît, dès le départ, comme une construction plus que comme un ramassage. Pour être récoltées, les données doivent être en quelque sorte conçues au préalable : il faut en effet savoir ce qu'on cherche pour le chercher. Et selon l'instrument avec lequel on va les chercher, les données prendront cette forme plutôt qu'une autre. Cela concerne tantôt les données recueillies par des organismes publics ou privés, et tantôt celles récoltées directement par les scientifiques. Dans le premier cas, il s'agit par exemple de se diriger vers les instituts statistiques nationaux, qui mettent à disposition gratuitement leurs données en ligne, ou vers les plateformes des réseaux sociaux ou des moteurs de recherche pour acheter, souvent à un prix très élevé, des grandes quantités de métadonnées[21]. Des études géographiques, sociologiques, épidémiologiques, psychologiques et autres pourront être ensuite menées par les scientifiques, pourvu que ceux-ci arrivent à convaincre les comités éthiques indépendants de l'irréprochabilité de leurs recherches. Voilà d'emblée un filtre institutionnel – car chaque institut possède ses logiques de collecte et de mise en forme des données –, un filtre économique – qui empêche certaines équipes d'exploiter des bases de données trop coûteuses – et un filtre éthique – sur ce qui peut être collecté ou pas selon les lois en vigueur à un moment donné.

Dans le deuxième cas, la collecte est menée ou supervisée par les scientifiques eux-mêmes. Soit ceux-ci conduisent en première personne des expériences en laboratoire, réalisent des entretiens, administrent des questionnaires ou décrochent le téléphone pour collecter les informations dont ils ont besoin. Soit ils externalisent la collecte à des organismes ou à des groupes qui sont sur le terrain, par exemple un syndicat de transports pour une équipe de géographes ou bien un réseau de veille sanitaire composé de médecins déployés sur le territoire pour une équipe d'épidémiologistes. Dans tous chacun de ces cas, la préexistence d'un ensemble de questions, d'hypothèses de recherche, voire d'un modèle théorique, est nécessaire pour savoir quelles données collecter et comment. À lire Chris Anderson, il semblerait que la méthode d'enquête *data-driven* soit non seulement possible conceptuellement et pratiquement, mais aussi la meilleure disponible. Les terrains, en revanche, montrent toujours que chacune des phases est à la fois *data-driven* et *hypothesis-driven*, de manière indissoluble. Cette double difficulté – matérielle et conceptuelle – dans la collecte et l'exploitation des

données a été efficacement capturée par le concept de « friction de données » proposé par l'historien des sciences du climat Paul Edwards, concept qui se couple avec celui de « friction computationnelle » :

> Tandis que la friction computationnelle exprime la difficulté qu'implique la transformation des données dans une information et dans de la connaissance, le concept complémentaire de *friction des données* exprime une forme de résistance plus primitive. Tout comme la computation, les données ont toujours un aspect matériel. Les données sont des *choses*. […] La 'friction des données' se réfère aux coûts en termes de temps, énergie et attention requis pour simplement collecter, vérifier, emmagasiner, déplacer, recevoir et avoir accès aux données. Toutes les fois que les données voyagent – que cela soit d'un lieu de la Terre à un autre, d'une machine (ou d'un ordinateur) à une autre, ou bien d'un support (i.e. carte perforée) à un autre (i.e. bande magnétique) – la friction des données empêche leur mouvement. Et puisque la computation est une forme d'opération sur les données, la friction computationnelle et la friction des données interagissent souvent[22].

Cependant, une fois collectées, les données sont loin d'être prêtes à l'usage. Elles doivent d'abord être corrigées, standardisées et complétées. Ces trois opérations constituent la phase la plus longue du travail des modélisateurs numériques : celle de la préparation des données.

III. La préparation des données

Préparer les données demande plusieurs astuces et un long travail à l'œil et la main, souvent pénible et répétitif, mais toujours informé d'hypothèses théoriques, expériences passées et intuitions personnelles. Le sociologue Emmanuel Didier, qui a travaillé sur les statistiques agricoles américaines des années 1930, appelle « plasma » cet ensemble de notations, chiffres et documents épars que les statisticiens « soupçonne[nt] d'être intéressants à transformer en statistiques, mais qui [leur] échappent encore »[23]. Il nomme « pratiques d'agrégation » toutes les actions nécessaires à élaguer, systématiser et nettoyer les données – ensemble d'opérations que nos scientifiques appellent « préparation des données ».

L'une des actions les plus communes consiste dans la correction pure et simple des erreurs, des coquilles et des incohérences manifestes dans le matériau disponible. Didier appelle cette activité « apurement » ou *editing* en anglais[24]. À ce propos, le sociologue nous offre une remarque importante qui contribue à déconstruire une idée reçue :

> Ainsi, contrairement à ce que l'on croit souvent, le travail statistique n'est pas uniquement régi par des règles et des procédures mécaniques. L'apurement, qui en était une étape incontournable, en est un exemple. En effet, les données étaient presque toujours autocontradictoires. Par conséquent, le Statisticien n'avait pas d'autre solution que de procéder à une évaluation personnelle et non standardisée visant à accorder les données le mieux possible. […] C'est pourquoi il ne pouvait être effectué que par *jugement* du Statisticien[25].

Ensuite, tous les scientifiques de notre corpus soulignent l'importance de l'harmonisation des données. Parfois, comme en témoigne une mathématicienne du LANL, certaines modélisations ne peuvent pas être faites, non pas par manque d'informations, mais parce que celles-ci ne sont pas homogènes et résultent, de ce fait, inexploitables. La chercheuse en question prend l'exemple des moustiques, vecteurs de plusieurs maladies infectieuses que son équipe d'épidémiologistes tient sous contrôle. Collectées par différentes antennes sur le terrain, les données concernant ces insectes ne peuvent être standardisées qu'une fois qu'on a compris comment elles ont été collectées au cas par cas. Cette opération chronophage et cognitivement coûteuse présuppose d'aller parler avec ces équipes, pour voir quels pièges à moustique elles ont utilisé, quelles espèces sont collectées, dans quels pourcentages, dans quels contextes environnementaux, etc. Un géographe du CNRS résume bien ce point, en soulignant le fait que le problème des *Big data* « n'est pas tellement la masse des données, c'est plutôt le fait qu'elles sont dans des formats qu'on ne maîtrise pas »[26]. Un ex-doctorant en géographie à la Sorbonne-Paris 1 fait mention de la nécessité de croiser les bases de données pour les vérifier et, en cas d'incohérence ou incomplétude, de contacter les autorités publiques les ayant produites pour leur demander des éclaircissements ou des bases alternatives.

Un jeune biologiste du LANL dit avoir développé une expertise particulière pour traiter les disparités entre les informations des différents départements de santé publique du monde. Les juridictions changent d'un pays à l'autre et, avec elles, les définitions des termes techniques[27]. L'asymétrie dans la puissance et dans l'efficacité des infrastructures sanitaires nationales représente une autre source de difficulté pour lui et son équipe d'épidémiologistes : les *monitorings* et les *reports* sur les épidémies n'arrivent pas au même moment partout, et parfois ils ne sont pas du tout disponibles. Dans un article dédié à cette thématique et coécrit avec sept collègues[28], ce biologiste a plaidé pour que les départements de santé publique du monde entier se dotent d'infrastructures à la hauteur des enjeux épidémiologiques contemporains et pour qu'ils rendent accessibles leurs données de manière plus homogène, afin de faciliter le travail des modélisateurs et des décideurs politiques dans l'anticipation des maladies infectieuses. Entretemps, ces scientifiques se doivent d'activer toutes leurs capacités d'intuition et de discernement, aspect sur lequel Didier insiste plusieurs fois dans son ouvrage :

> il est impossible de faire des calculs sur des données hétérogènes. À partir du moment où les chiffres ne parlent pas de la même chose, toutes les formules algébriques deviennent inutiles. La seule opération qui reste possible pour synthétiser de telles données est alors le jugement. Les [statisticiens] ne calcul[ent] donc pas, ils décid[ent][29].

À part homogénéiser leurs données, les scientifiques doivent parfois les compléter en faisant un usage intelligent des informations disponibles. Un exemple de ce remplissage par abduction est donné par une épidémiologiste computationnelle basée à l'INSERM de Paris. Occupée dans la construction d'un modèle pouvant simuler les interactions entre tous les pays du monde à partir de données géographiques, démographiques et de mobilité, elle disposait d'une seule source homogène et exhaustive : celle des vols de ligne fournie par l'IATA[30]. En plus de cela, elle disposait d'une quarantaine de bases

hétéroclites provenant de différents instituts statistiques nationaux du monde pour ce qui concerne les déplacements pendulaires. Ces bases contenaient « l'Europe et l'Amérique du Nord, quelques pays de celle du Sud et certains pays asiatiques. L'Afrique était absente pour des raisons évidentes »[31]. Sachant que les États de la planète dépassent les deux cents unités, comment faire pour les absents ? La chercheure résume sa démarche ainsi : « Un modélisateur modélise les données qu'il a et puis il les applique aux points qui manquent »[32]. En se basant sur une théorie mutualisée aux géographes quantitativistes, elle a appliqué le « modèle *gravity law* » pour produire les données manquantes, car, au sens de la métaphore newtonienne, les déplacements pendulaires tendent à aller vers le point voisin le plus important, les grandes villes résultant ainsi les plus attractives : « nous sommes partis des aéroports, en les considérant comme des bassins d'attraction d'un certain type de mobilité […] En gros, le pendulaire entre Turin et Milan ne se comporte pas de manière différente par rapport à un pendulaire colombien »[33]. Pour préciser les données, notre interlocutrice a employé un autre expédient pour reconstruire la distribution de la population africaine sur le territoire. À partir d'une base de données satellitaires de la NASA, elle a calibré la densité lumineuse émise par les centres urbains de tous les continents sur les populations connues de ceux-ci.

Dans la littérature philosophique, sociologique et historique ici mobilisée, ce type d'opérations consistant dans la création d'une nouvelle donnée à partir de données existantes semble être sous-thématisée. Tout d'abord, il s'agit d'une forme de rapprochement, nécessaire à compenser l'impossibilité d'un recoupement. Comme l'explique le sociologue Francis Chateauraynaud, les recoupements « visent des opérations perceptuelles au contact des choses », tandis que les rapprochements « concernent les opérations intellectuelles qui associent […] des objets physiquement séparés »[34]. Or, « [l]e travail de l'enquête développe une économie cognitive qui consiste à maximiser les chances d'obtenir des recoupements et à réduire la liste des rapprochements nécessaires »[35]. Mais quand cela n'est pas possible, les scientifiques se tournent vers des solutions créatives comme celles que l'on vient de voir. Pour conceptualiser ce type de pratiques, on peut utiliser une métaphore neuroscientifique. En comparant les lacunes des données à des lésions cérébrales, on peut parler de datagénèse comme l'on parle de neurogénèse pour décrire ce mécanisme de neuroplasticité qui permet à un cerveau endommagé de récupérer partiellement ou totalement ses fonctions, grâce à la multiplication d'associations synaptiques alternatives. Tout comme le cerveau est parfois capable d'activer un processus de compensation menant au rétablissement d'une fonction, les scientifiques en manque de données peuvent combler les lacunes de manière indirecte, en mobilisant des données tierces pour extraire des informations d'une base de données qui ne les contient pas directement. Cet expédient inventif est plus qu'une extrapolation statistique classique, consistant à prolonger une série en introduisant à la suite des termes anciens un terme nouveau qui obéit à la loi de la série ; il est aussi plus qu'une interpolation, consistant à déterminer, dans une série, de nouvelles valeurs correspondant à des points intermédiaires pour lesquels manquent des mesures directes. Ce n'est pas non plus une forme de triangulation, qui supposerait de déduire une troisième donnée à partir de deux

disponibles. Il s'agit, en définitive, d'une génération sous contrôle théorique de données nouvelles à partir de données disponibles fragmentaires, par le biais d'un détour appuyé sur des données autres.

Conclusion

On a l'habitude de parler d'hypothèses quand on aborde la problématique de la construction du modèle. Mais, comme on vient de le voir, la reconstruction des données se fait déjà dans un cadre théorique et se fonde sur de nombreuses hypothèses – lieu par ailleurs d'intéressants croisements interdisciplinaires. Complexes et laborieuses, ces opérations demandent énormément d'investissement temporel et cognitif : l'épidémiologiste de l'INSERM citée ci-dessus déclare avoir eu besoin de « deux ans de travail de collecte de données et d'analyse-homogénéisation » (sur trois ans de contrat postdoctoral) pour construire sa première base de données[36]. Elle, comme tant d'autres, confirme donc l'extrait d'entretien cité en introduction à propos du temps passé par les scientifiques sur les données.

La primauté des données a émergé spontanément au cours de nos entretiens avec les spécialistes des systèmes complexes et a d'ailleurs été évidente lors de nos visites de laboratoire. Devant la requête de décrire les phases de leur travail, nous nous attendions à ce qu'ils dissertent en long et en large sur la structure et la validité de leurs modèles digitaux. De manière surprenante, les scientifiques ont porté notre attention sur d'autres opérations que celles sur lesquelles on a l'habitude de s'attarder dans la littérature – à savoir la construction des hypothèses du modèle, la programmation du code informatique et la validation des outputs de celui-ci. En conclusion, ce court article a voulu pointer un terrain à explorer par la réflexion philosophique, notre objectif n'étant pas la description fine du processus de modélisation numérique, d'ailleurs trop riche et complexe pour être décrit exhaustivement ici. Certes, dans ce processus, la simulation *in silico* reste l'horizon de nos scientifiques, mais, contrairement à l'idée qu'on se fait de leur action en lisant les travaux d'épistémologie, les opérations dédiées à la mise en opérabilité des données représentent l'écueil le plus grand, l'activité la plus longue et pénible. Elle se fait toujours en fonction et en parallèle de la construction d'un modèle, mais elle a ses propres logiques et ses propres pratiques qui méritent d'être enquêtées et pensées en tant que telles.

---

[1] La géographe Nicole Mathieu et la philosophe Anne-Françoise Schmid soutiennent, dans un livre collectif sur l'interdisciplinarité et l'épistémologie, que cette dernière, « après avoir été éclipsée quelque temps par les *science studies*, est en train de renaître autour des concepts de modélisation et de simulation » : Nicole Mathieu et Anne-Françoise Schmid (dir.), *Modélisation et interdisciplinarité. Six disciplines en quête d'épistémologie*, Versailles, Quæ, 2014. Tout en nous réjouissant de la renaissance de l'épistémologie autour de cet objet d'étude, le regard des *science studies*, avec leur obsession pour la matérialité, permet d'enrichir l'épistémologie avec une dimension qu'elle a parfois tendance à oublier : celle des pratiques matérielles.

[2] Fabrizio Li Vigni, *Les systèmes complexes et la digitalisation des sciences. Histoire des instituts de la complexité aux États-Unis et en France*, Thèse de doctorat en sociologie, Paris, École des Hautes Études en Sciences Sociales, 2018.

[3] Isabelle Stengers et Bernadette Bensaude-Vincent, *100 mots pour commencer à penser les sciences*, Les Empêcheurs de penser en rond, Paris, 2003, p. 382.

[4] Chris Anderson, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired*, 06.23.08. Disponible en ligne : https://www.wired.com/2008/06/pb-theory/.

[5] Léo Coutellec et Paul-Loup Weil-Dubuc, « *Big data* ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique », *Journal international de bioéthique et d'éthique des sciences*, vol. 28, n. 3, 2017, p. 63-79.

[6] Sabina Leonelli, S., « Introduction: Making sense of data-driven research in the biological and biomedical sciences », *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, n. 1, p. 1-3.

[7] Christian S. Calude et Giuseppe Longo, « The Deluge of Spurious Correlations in Big Data », *Foundations of Science*, vol. 22, n. 3, 2017, p. 595-612.

[8] Alisa Bokulich, « Using models to correct data: paleodiversity and the fossil record », *Synthese*, 2018, https://doi.org/10.1007/s11229-018-1820-x.

[9] Hykel Hosny et Angelo Vulpiani, « Forecasting in Light of Big Data », *Philosophy and Technology*, vol. 31, n. 4, 2018, p. 557-569.

[10] Wendy Lipworth, Paul H. Mason, Ian Kerridge et John P.A. Ioannidis, « Ethics and Epistemology in Big Data Research », *Journal of Bioethical Inquiry*, vol. 14, n. 4, 2017, p. 489-500.

[11] Marie Farge, « L'approche numérique en physique », *Fundamenta Scientiae*, 1986, vol. 7, n. 2, p. 155-175 ; Peter Galison, *Image and Logic. A Material Culture of Microphysics*, Chicago, London, The University of Chicago Press, 1997, p. 45-46 ; Deborah Dowling, « Experimenting on Theories », *Science in Context*, 1999, vol. 12, n. 2, p. 261-273 ; Eric Winsberg, « Simulated Experiments : Methodology for a Virtual World », *Philosophy of Science*, 2003, p. 70, p. 105-125.

[12] Paul Humphreys, *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*, Oxford, Oxford University Press, 2004.

[13] Stephen D. Norton et Frederick Suppe, « Why Atmospheric Modeling is Good Science », dans Clark Miller et Paul & Edwards (dir.), *Changing the Atmosphere : Expert Knowledge and Environmental Governance*, Cambridge, MIT Press, 2000 ; Eric Winsberg, « A Tale of Two Methods », *Synthese*, 2009, n. 169, p. 575-592 ; Corinna Elsenbroich, « Explanation in Agent-Based Modelling : Functions, Causality or Mechanisms ? », *Journal of Artificial Societies and Social Simulation*, 2011, vol. 15, n. 3 ;

Isabelle Peschard, « Les simulations sont-elles de réels substituts de l'expérience ? », dans Franck Varenne et Marc Silberstein (dir.), *Modéliser & simuler. Épistémologies et pratiques de la modélisation et de la simulation*, tome 1, vol. 1, Paris, Éditions Matériologiques, 2013, p. 161.

[14] Entretien avec un statisticien et épidémiologiste computationnel du LANL, 29.09.16.

[15] Jules Berman, *Principles of Big Data. Preparing, Sharing, and Analyzing Complex Information*, Boston (MA), Elsevier, 2013 ; Ahmed Elragal et Ralf Klischewski, « Theory-driven or process-driven prediction ? Epistemological challenges of big data analytics », *Journal of Big Data*, vol. 4, n. 19, 2017, https://doi.org/10.1186/s40537-017-0079-2 ; Gordana D. Crnkovic, Juan M. Duran et Davor Slutej, « Content Aggregation, Visualization and Emergent Properties in Computer Simulations », dans Kai-Mikael Jää-Aro & Thomas Larsson (dir.), *SIGRAD 2010 – Content aggregation and visualization*, Linköping, Linköping University Electronic Press, 2010, p. 77-83.

[16] Mitchell M. Waldrop, *Complexity. The Emerging Science at the Edge of Order and Chaos*, New York, Simon & Schuster Paperbacks, 1992 ; Stefan Helmreich, *Silicon Second Nature. Constructing Artificial Life in a Digital World*, Berkeley & Los Angeles, University of California Press, 1998.

[17] Des techniques de simulation qui intègrent des données dans des modèles, ceux-ci consistant en des règles d'interaction censées mimer le fonctionnement des systèmes étudiés.

[18] Sven Opitz, « Simulating the world: The digital enactment of pandemics as a mode of global self-observation », *European Journal of Social Theory*, vol. 20, n. 3, 2017.

[19] Sylvain Cuyala, *Analyse spatio-temporelle d'un mouvement scientifique. L'exemple de la géographie théorique et quantitative européenne francophone*, Thèse de doctorat en Géographie, Paris, Université Paris 1 Panthéon-Sorbonne, 2014.

[20] Pour l'épidémiologie computationnelle nous avons visité les laboratoires suivants : EPIcx, INSERM, Paris, en juillet 2015 et en mai 2017 ; Los Alamos National Laboratory (LANL), Los Alamos, en septembre 2016 ; Network Science Institute (NETSI) (Northeastern University), Boston, en septembre 2016 ; et l'Institute for Scientific Interchange (ISI), Turin, en février 2017. Pour la géographie quantitative nous avons visité : le Santa Fe Institute à Santa Fe en septembre 2016 et Géographie-cités, Université Sorbonne Paris 1 et CNRS, à plusieurs reprises entre 2015 et 2017.

[21] C'est-à-dire des données agrégées et anonymisées à partir des données personnelles des internautes.

[22] Paul Edwards, *A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming*, Cambridge, MIT Press, 2010, p. 84 (traduction de l'anglais par nos soins).

[23] Emmanuel Didier, *En quoi consiste l'Amérique ? Les statistiques, le New Deal et la démocratie, Paris,* La Découverte, 2009, p. 23.

[24] *Ibid.*, p. 40.

[25] *Ibid.*, p. 43.

[26] Entretien avec un géographe quantitativiste du CNRS, 01.03.17.

[27] Alain Desrosières, *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte, 2010.

[28] Geoffrey Fairchild, Byron Tasseff, Hari Khalsa, Nicolas Generous, Ashlynn R. Daughton, Nileena Velappan, Reid Priedhorsky et Alina Deshpande, « Epidemiological data challenges: planning for a more robust future through data standards », 2018, eprint arXiv:1805.00445.

[29] Emmanuel Didier, *En quoi consiste l'Amérique ? Les statistiques, le New Deal et la démocratie, Paris,* La Découverte, 2009, p. 64.

[30] International Air Transport Association.

[31] Entretien avec une épidémiologiste computationnelle de l'INSERM, 24.07.15.

[32] *Ibid*.

[33] *Ibid*.

[34] Francis Chateauraynaud, « L'épreuve du tangible. Expériences de l'enquête et surgissements de la preuve », *La croyance et l'enquête. Aux sources du pragmatisme. Raisons pratiques*, vol. 15, EHESS, 2004, p. 167-194.

[35] *Ibid.*

[36] *Ibid.*