Projet ANR-14-CE22-0006

ELITISME

ELITISME WORKING PAPER SERIES

ANR-ELITISME-2017-001

Big Data : Tour d'horizon et repérages

André de Palma

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR
Labex MME-DII
Modèles Mathématiques et Économiques de la Dynamique, de l'Incertitude et des Interactions
UNIVERSITÉ de Cergy-Pontoise
ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY
ENS CACHAN
université PARIS-SACLAY

Big Data : Tour d'horizon et repérages

André DE PALMA

Ce document entend exposer, puisqu'on ne le fait guère, ce que peuvent dire les scientifiques sur l'irruption des données massives. Les capteurs de données numériques sont aujourd'hui peu coûteux. La course aux machines *exaflops*, qui effectuent à chaque seconde des milliards de milliards d'opérations en virgule flottante, est aujourd'hui engagée. La fibre optique envahit les villes. Des milliards de téléphones portables circulent, qui contiennent de petits ordinateurs. La liste « collecte, communication, traitement et usage » définit les fronts à surveiller pour tous ceux qui voudront suivre cette évolution ou l'anticiper.

Cette évolution intéresse l'ensemble des secteurs économiques. Certains limitent parfois son impact, voire son centre moteur, au secteur tertiaire. C'est ignorer les perspectives et les enjeux. Car l'accélération de la concurrence pousse à des décisions réalistes. Ainsi, la Commission européenne encourage les industriels à concentrer leurs efforts. On parle de tenter avec les calculatrices rapides ce qu'on a réussi avec Airbus ; c'est le but du programme BXI du consortium CEA/Bull-Altos. Pendant ce temps, Volkswagen filtre et teste des données collectées sur 10 000 taxis à Pékin, et se propose d'optimiser le trafic en temps réel à l'aide d'ordinateurs quantiques, dont la puissance de calcul franchit de nouvelles échelles encore.

L'information, toujours produite en quantité finie, est depuis longtemps une composante intrinsèque des secteurs primaire et secondaire, secteurs qui nous confrontent à toutes sortes d'incertitudes plus ou moins réductibles sur la nature et sur les technologies. Mais l'incertain n'est pas le résidu de ce que la science n'a pas conquis. Il est le champ de manœuvre de la chercheuse, comme nous l'ont appris des observateurs attentifs comme Abraham Moles avec ses *Sciences de l'imprécis* (1990) et Georges-Théodule Guilbaud avec ses *Leçons d'à peu près* (1995). On retrouve ici la distinction classique des techniques et des technologies. Une technique applique à une fin déterminée un moyen contrôlable et approprié : l'outil (définition d'Aristote). Une nouvelle technologie permet une gestion plus vague des rapports moyens / fins, puisqu'elle dégage un champ de possibles à défricher et invite au bricolage, où François Jacob a proposé de voir une stratégie du vivant (voir *Le jeu des possibles*, 1981, et sa référence au contraste de l'ingénieur et du bricoleur esquissé par Claude Lévi-Strauss dans *La pensée sauvage*, 1962).

Ces rappels seront utiles pour baliser notre futur. Ce seront dans le cas général de grands opérateurs industriels qui stimuleront ces nouveaux développements, et non pas des sociétés de conseil et de sous-traitance du tertiaire, moins encore des exploitants de services liés aux réseaux sociaux, quelles que soient les captations de chalandise. L'électronique, l'informatique et les communications restent des secteurs concentrés à l'extrême, produits par un monde oligopolistique qui a assimilé Bell et Edison. Par exemple, IBM et la SNCF peuvent attendre beaucoup de leur nouveau programme de maintenance industrielle en collaboration autour de la plate-

forme IBM Watson, alimentée par des milliers de capteurs implantés sur des centaines de trains et bientôt sur les voies et dans les gares.

Il est vrai qu'il existe un secteur foisonnant de petites sociétés de services flexibles au client où la concurrence est vive. Un spécialiste de la diversité du plancton marin a proposé une lecture écologique de ces évolutions foisonnantes : la nature devient baroque dans les situations où peu d'énergie est disponible pour de grands changements (Margalef, 1974). Certains de ces entrepreneurs réussissent à devenir eux-mêmes de grands acteurs industriels. Google/Alphabet et Apple sont à présent les deux premières capitalisations boursières du monde. Néanmoins, cela ne change pas la pyramide écologique des puissances de calcul.

Considérons l'échelonnement de cet univers. Le téléphone et l'ordinateur personnel, en passant par l'ordinateur portable, occupent les degrés inférieurs d'une gamme en constante évolution ; mais le centre de calcul et le super-ordinateur, qui occupent d'autres degrés, seront demain plus présents que jamais, précisément parce qu'on peut désormais mobiliser les données en masse. À moins d'un bouleversement technologique majeur, le tempérament de cette gamme relèvera longtemps encore de stratégies industrielles lourdes. Je me rappelle la surprise de cet ingénieur qui avait promis de remplacer le thermostat du frigo, et qui trouva un vieux processeur 68 000, objet-culte de ses années d'étude. Les objets passent, la segmentation du marché évolue plus lentement et amortit sur le bas de gamme des procédés et des équipements désormais obsolètes dans leur niche originaire.

On propose ici quelques remarques sur la collecte, la communication, le traitement et l'usage des données massives ainsi que sur les structures logiques qui permettent aux machines de répondre à nos questions. Au préalable, on décrit rapidement l'activité du secteur public dans ce domaine de la « datamasse » ; on discute la nouveauté du phénomène ; et on brosse quelques tableaux historiques sommaires pour baliser les horizons qui s'éloignent et ceux qui se rapprochent. Le tout sera formulé du point de vue d'un petit groupe mal défini : les gens du calcul des données massives. Je mentionnerai donc quelques épisodes et un vocabulaire, parfois des préoccupations ou des curiosités qui pourront surprendre. Certes, les comptables aussi comptent, et les juges, et les dévots. Mais il existe des hommes et des femmes incapables de résister devant une série numérique curieuse ou un protocole ingénieux de correction d'erreur. J'ai cru utile de les convoquer ici pour aider le lecteur à comprendre quelles portes vont s'ouvrir, et quelles discussions concernent ou pas ces gens-là.

1. Tableau d'ensemble

Les données massives subissent-elles des effets de mode, comme jadis la tulipe et naguère l'intelligence artificielle ? Une bulle spéculative comparable à la « bulle Internet » des années 2000 est possible. Ce risque même n'en fera pas une vulgaire panacée foraine. Pour prendre un peu de recul, lisons la notice du *Journal officiel* sur le vocabulaire des *Big Data*[1] : elle parle de « données structurées ou non dont le très grand volume requiert des outils d'analyse adaptés », ce qui indique un matériau hétérogène.

Commençons ce tour d'horizon par le secteur public. Il est très présent sur ce front. Seul capable d'initiatives proportionnées à ces échelles nouvelles d'infrastructures, il peut encourager la nécessaire normalisation des données et des protocoles.

1. Le *Journal officiel* (voir n° 0193 du 22/8/2014, p. 13972 texte 89) propose pour les *Big Data* l'expression « données massives », avec « mégadonnées », qu'on peut trouver moins heureux puisque le préfixe pourrait suggérer un ordre de grandeur déterminé. On parle aussi de « datamasse ».

1.1. Le secteur public et les corps d'ingénieurs des données

Les services du Premier ministre ont nommé un Administrateur général des données, Henri Verdier. Dans son premier rapport, celui-ci commente une affirmation ravageuse de deux chercheurs (« Google en sait plus que l'INSEE sur la France », Grumbach et Frénot, 2013) :

« La formule, largement reprise, fait mouche … [elle] masque pourtant la complémentarité entre les sources. L'INSEE produit les statistiques par des méthodes scientifiques robustes, de manière objective et transparente. Les statistiques publiques présentent une cohérence interne et dans le temps. Elles permettent les comparaisons … Les grands acteurs du numérique, pour leur part, collectent des données de manière non scientifiquement contrôlée**,** grâce à des capteurs, au recueil de traces d'utilisation, ou encore à la contribution des internautes. Elles n'ont ni la robustesse, ni la complétude des données scientifiques. Elles créent cependant progressivement, de par leur seul volume, une forme d'empreinte du réel qui peut à son tour être interprétée et être utilisée : produire un savoir activable. » (Verdier, 2015, p. 15 ; on consultera aussi le site data.gouv.fr.)

Rappelons donc d'abord les missions de l'État sur la collecte de l'information. Elles ont été redéfinies depuis le XVIII^{e} siècle (recensement, cartographie, cadastre, état civil, bibliothèques). Elles se sont élargies depuis (démographie, santé, économie, transports, météo, bibliographie, annuaire téléphonique) et sont assurées par de grands établissements spécialisés.

Les masses de données ne font pas à elles seules les données massives. Certains demanderont combien ces organismes récoltent de données et on posera des questions d'efficience rétroactive ; mais ce n'est pas notre sujet, qui est de reconnaître et situer les communautés des calculateurs, et de voir ce qu'elles ont à dire sur les données massives. C'est surtout l'émergence de compétences reconnues et de machines simples (dont l'exemple type reste la règle à calcul) qui a provoqué par poussées la spécialisation d'ingénieurs affectés à la partie numérique du traitement au sein des établissements, et bientôt dans des équipes spécifiques. Voilà repérés par l'État « les gens du calcul des données massives ». On a fondé pour eux des organismes dotés de moyens considérables : le Commissariat à l'Énergie Atomique et aux Énergies Alternatives, l'Institut du Développement et des Ressources en Informatique Scientifique (Orsay), le Grand Équipement National de Calcul Intensif, le Centre Informatique National de l'Enseignement Supérieur, l'Institut National de Recherche en Informatique et en Automatique ; l'Office National d'Études et de Recherches Aérospatiales.

La Puissance publique tentera de valoriser cette richesse de compétences et d'informations accumulée, comme celles que détiennent les EPIC et autres organismes chargés de fonctions d'intérêt public, comme la SNCF ou la RATP. Aujourd'hui, les données de ces établissements, de leurs sous-traitants et de leurs concurrents sont parfois difficiles à consulter : leur portabilité est précaire, et leur accès est parfois barré. L'État essaie de dégager les voies et les moyens du futur sur ce point préoccupant. Signalons d'abord un rapport de Laurence Monnoyer-Smith (2016), qui propose de définir la fonction d'un superviseur général des données dans le cadre du ministère de l'Environnement, lequel a dans ses attributions les transports. Ensuite, des remarques de la Cour des comptes au Premier ministre, qui confirment que « l'État devrait … imposer l'ouverture la plus large des données du transport ». Celui-ci montre la généralité du problème, en exposant un des axes de la stratégie de la compagnie d'Uber, qui refuse à ses partenaires l'usage de comparateurs de prix, utilisés pourtant dans le transport aérien et l'hôtellerie. L'État doit donc se pencher sur la libre circulation de données. Suivons son regard.

1.2. La nouveauté cruciale : les capteurs branchés

De nombreux gestionnaires des données massives avaient somme toute une philosophie proche de celle du service public. Après les artilleurs, cartographes, astronomes et météorologues, qui sculptaient depuis des siècles des colonnes de nombres immenses (toutes proportions gardées), économistes, psychologues, sociologues, criminologues, médecins ou agents de circulation et autres sont invités à faire face à leurs datamasses respectives. On prévoit le surgissement de nouveaux acteurs économiques spécialisés dans telle ou telle étape du traitement des données, par sous-traitance occasionnelle ou par externalisation régulière. Chaque acteur pourra soupçonner les autres de tenter de saisir l'opportunité de détourner à leur profit une part des travaux à effectuer.

Ce nouveau monde est-il né sous nos yeux ? Il existe un argument pour l'affirmer. On avait compris naguère que l'Internet offre un accès renouvelé aux données massives. On voit aujourd'hui comment il pousse à les produire, en raison des transactions qu'il suscite puisqu'il peut les enregistrer et les communiquer. Que je m'intéresse à la musique baroque espagnole est en soi une information anecdotique et surtout sans valeur ; que ce trait de comportement se traduise par le clic d'une souris, et voilà des programmes d'édition, de compilation, d'indexation et de consultation qui se mettent en route.

Les économistes sont bien placés pour évaluer le coût et la rentabilité des transactions. Votre voisin, quoi qu'on dise, ne produit pas de données chaque fois qu'il respire ; il faut une transaction, et une transaction enregistrable. Ces transactions ne sont pas forcément marchandes dans l'immédiat. Un téléphone portable permet d'enregistrer le nombre de pas, le pouls et la respiration de l'usager d'un logiciel de santé. Cette fureur d'encodage reste souvent mystérieuse pour l'observateur, on dira : virtuelle. Celui qui met un tel logiciel en circulation changera peut-être d'opinion, et donc de stratégie, sur la façon dont il valorisera à bon droit ou non ces données captées sur autrui. Il pourra en cours de route, comme dans toute bonne intrigue, imputer son crime à un autre, émarger au chômage, et tout ce qu'on voudra. Les informaticiens ont des procédures pour le *garbage collection* (la vidange des espaces-mémoire). Mais il faudra d'abord leur donner ce mandat explicitement.

Certains acteurs offrent à titre gracieux des données qu'ils collectent ; entre-temps ils vendent à des annonceurs publicitaires des profils d'usagers pour permettre une publicité ciblée. D'autres valorisent sans intermédiaire la clientèle ainsi constituée, qu'ils traiteront ou non comme un marché captif, selon leur intérêt plus ou moins bien compris.

On ne discutera pas ici les conditions économiques, sociales, politiques ou juridiques de leur apparition, pas plus que leurs stratégies. Soulignons que la transaction, au moins virtuelle, est le critère qui motive le travail de collecte et la conservation des données. Celles-ci n'ont pas plus et pas moins de valeur que les minerais en attente d'exploitation. Le cours du pétrole module les campagnes de forage comme la valorisation espérée des données module leur collecte. Il se trouvera toujours des artistes ou des fous pour collecter des données sans les exploiter. Ces entreprises d'une autre sorte sont me semble-t-il souvent valorisées sur un autre plan ; conclura-t-on à un économisme opiniâtre dans mon chef ?

Selon ce schéma, Internet serait le premier moteur de la montée en puissance des données massives : la communication donnerait une valeur à des données jusqu'à hier insignifiantes mais peu coûteuses et utilisables à diverses fins. En pratique, l'ensemble des sous-circuits qui parcourent la liste « collecte, communication, traitement et usage » concourent à déterminer une causalité réciproque où les progrès de chaque segment stimulent ceux des autres. Examinons à présent la structuration souvent faible des requêtes adressées aux moteurs de recherches.

1.3. Première approche de la datamasse

Pour introduire les *Big Data*, on évoque volontiers les moteurs de recherche sur Internet. Ce n'est pas faux. Cependant, pour indexer les pages consultées par les robots sur la Toile comme pour répondre aux requêtes, les concepteurs de Google n'ont pas cherché à tout savoir, ni à répondre sur le sens du contenu.

Ce sont là des tâches du futur si l'on néglige, à tort, certains sites traditionnels ; après tout, l'usager des bases bibliographiques des années 1980 avait sur la redéfinition progressive de son cœur de cible documentaire un contrôle manuel plus fort que celui qui interroge aujourd'hui un moteur de recherche : les données étaient plus structurées.

Fondateurs de Google, Brin et Page (1998) ont su se contenter d'idées brillantes sur le comptage des liens entre pages[1]. Il ne s'agit pas ici pour l'opérateur industriel d'effectuer sur les données un traitement de haut niveau, mais seulement de repérer les pages vers lesquelles convergent via les liens HTTP un grand nombre de pages porteuses des mots-clés visés. Les programmes construisent une sorte de pertinence probabiliste. Dans ces tables d'index effectivement gigantesques et localisées dans les « fermes de données » qu'exploitent les opérateurs de ces moteurs de recherche, repérer des sous-ensembles de lignes pose surtout des problèmes de balayage des fichiers. Pour ne pas soumettre les machines à des tâches infernales, les opérateurs distribuent le travail, et balaient des sous-fichiers différents sur cent machines en parallèle.

Quand vient le résultat, on parle de miracle, et le mot est juste, puisqu'en bonne doctrine, avant de violer les lois de la nature, le miracle fait surtout signe au croyant. Ce dieu sorti de la machine n'est pas très subtil, même s'il apporte souvent une résolution aux petits drames de notre vie documentaire quotidienne. L'usager se contente d'ailleurs de peu : il n'utilise guère les opérateurs logiques proposés par ces moteurs comme l'union et la négation. Le plus souvent il jette quelques mots pour chercher une intersection. Ce qui lui importe c'est de retrouver la bonne réponse, et souvent elle apparaît dès la première page après quelques essais. S'agit-il ici d'intelligence artificielle ou de données massives ? Cette démarche quotidienne illustre-t-elle vraiment l'irruption des puissances documentaires émergentes ? En fait, ces dernières promettent beaucoup plus.

Certains affirment que les machines seront toujours dépourvues d'imagination et de bon sens, mais pas de logique ni de talent pour le calcul. Cette mauvaise querelle a eu des suites (Gams *et al.*, 1997). Pour contribuer utilement au débat, Hubert Dreyfus a tenu à discuter l'apport possible d'idées inspirées par certains courants de la philosophie contemporaine (corps, monde, intention, signification) aux programmes de recherche en intelligence artificielle. Il intitule son article : « Pourquoi l'intelligence artificielle heidegerrienne est en panne, et comment pour la réparer il faut la refaire en plus heidegerrien » (Dreyfus, 2007). Quelle que soit sa portée, cette intervention ne correspond guère au rôle qu'on assigne à cet auteur dans des querelles de principes.

L'évolution des rapports homme-machine sera donc plus complexe. Les machines peuvent acquérir un peu de toutes ces choses, si on tient à l'effet de style, au moins dans l'esprit d'une remarque sur les petits mensonges, elle aussi constructive. Comme le dit Donald Knuth dans la préface du TeXbook (1984) : « En général, les chapitres suivants contiennent davantage d'informations fiables que les précédents. L'auteur pressent que cette technique de mensonge délibéré rendra plus facile l'apprentissage des idées ».

On hésite à supputer pourquoi les ennemis des machines ont concédé qu'elles savent calculer. En fait, elles ne savent pas : elles voient passer des codes représentant les nombres et les opérations à effectuer. Surtout, elles

1. Cette technique a un coût : au début de ce millénaire, leurs machines mettaient des heures pour traiter un immense tableau produit pendant l'indexation des données du réseau. La procédure de pertinence probabilisée est expliquée simplement dans le chapitre 14 d'un manuel consacré aux mathématiques élémentaires des réseaux sociaux et de l'Internet, Easley et Klinberg (2010).

n'avaient pour être des calculatrices aucun besoin de se donner une représentation de ce qu'est un nombre. C'était là l'étape décisive.

Des machines, aujourd'hui, ont reçu quelques idées sur ce sujet et sur d'autres. Pour qu'elles nous aident à débroussailler l'univers des données massives, il faudra les aider à se donner d'autres représentations encore. Autrement dit, elles ne seront pas de simples outils appropriés à leur usage, mis par un bon maître à portée de notre main. On n'a pas fini d'explorer les effets de ce décentrement.

Pour qu'une machine distingue ne serait-ce qu'un nom d'auteur d'un titre de livre, il faut lui dire comment faire. Le miracle, certes, est que Google n'en a pas besoin ; mais les miracles suffisent-ils ? Dans les moteurs de recherche spécialisés, chaque case des tables est bien définie, et cette information a été accumulée et mise en forme par des êtres humains. Depuis deux siècles, de nombreuses disciplines ont entrepris de construire des tables de bibliographies, de poids atomiques, d'affinités chimiques, de pH, de positions des objets célestes, et autres données bien spécifiées. Leurs auteurs ont tout naturellement pris pied sur Internet, et sont engagés dans des entreprises qui relèvent sans aucun doute des données massives. On écrit le bilan des banques avec plus de chiffres qu'on n'en connaît pour de nombreuses mesures physiques ou chimiques ; je vous fais grâce des remarques d'Oskar Morgenstern (1972) sur l'abus de précision dans les données économiques.

Pour conclure ce tableau très grossier, reprenons la liste des caractéristiques de la datamasse revisitée par Pierre Senellart[1] :

– Volume : on est souvent plusieurs ordres de grandeur au-dessus de ce qu'une machine peut traiter de manière centralisée.

– Variété : on gère divers types de données (textes, multimédia, données structurées et autres).

– Vitesse : les données évoluent et sont produites à grand débit.

– Véracité : la qualité de l'information est très variable selon des contextes dont la définition est malaisée.

Le quatrième V, pour véracité, qui rallonge la liste classique des 3V, n'est pas là pour avertir face à la précarité des données. On conçoit que certains soient intéressés à une redéfinition normative de la véracité. On connaît par exemple la publicité comparative et les dossiers pour consommateurs.

Mais ici la véracité est en passe de devenir un critère de travail au service des programmes de construction de grands ensembles de connaissances utilisables par les machines. Ce travail ne résoudra aucun des problèmes juridiques et philosophiques qu'on vient d'évoquer. Il pourra aider à les formuler. Le lecteur trouvera sur le site pierre.senellart.com/talks/ un exposé sobrement intitulé « Que savent sur vous les entreprises de l'Internet ? » (*Journée de réflexion sur les monopoles*, DG Concurrence, Commission européenne, 2016).

2. Repères historiques

2.1. De la Préhistoire aux temps contemporains

Pour mieux cerner la problématique de l'histoire du traitement des données massives, voici quelques exemples dispersés dans le temps. La situation en histoire des sciences est un peu ce qu'elle est en paléontologie : les échelles temporelles sont immenses et les données rares. Les dossiers isolés sont donc moins mal connus que les ensembles, sans parler des séquences.

1. Je reprendrai au paragraphe 3.1 quelques éléments de réflexion proposés par mon collègue à l'ENS Cachan, Serge Abiteboul. Senellart, un de ses élèves, enseigne à Télécom ParisTech.

L'histoire des sciences est fragmentaire et discontinue, et pas seulement dans les documents positifs qui sont à notre disposition. De nombreux chapitres sont dépareillés : ce fut sans doute aussi vrai dans la réalité des explorations de nos ancêtres que dans les documents qui nous en restent : perdre, retrouver, relire.

Cherchera-t-on ici les données massives sous d'autres noms ? Non. Une rumeur de grands nombres accompagne pourtant l'espèce humaine. Les hommes sont à la peine face à un travail de *pattern matching* où il faut distinguer d'un coup d'œil deux collections de 10 et 11 objets dans un laboratoire de psychologie. Ils ont développé des systèmes de numération et des opérations d'évaluation et de calcul. Autrement dit, s'ils n'ont pas d'intuition des nombres au-delà de la demi-douzaine, ils se donnent des outils pour dénombrer.

Ne parlons pas ici des calculateurs prodiges, même s'ils existent[1]. Parlons seulement des bons calculateurs, puisque c'est leur point de vue que nous tentons d'exposer ici. Certains travaux historiques ou ethnographiques permettent de deviner combien ils peuvent être parfois appréciés.

Certaines nations de Nouvelle-Guinée éprouvent une « obsession pour la quantification ». « Avec d'autres nations, [ils] traitent souvent de grandes quantités d'objets de valeur, qui demandent des méthodes efficaces de comptage. On trouve un … système de numération complexe – au moins traite-t-il des grands nombres, en base 60 – chez les Kapauku ». Ce n'est pas une furie de comptage abstrait. Lors de grandes fêtes, « des milliers de cochons changent de propriétaire en un seul jour … l'ensemble des prestations est enregistré par l'un des membres du noyau dur des donneurs : il s'agit d'hommes importants qui balaient les rangées en courant sur un mode stylisé, tout en dénombrant les objets du cadeau collectif … Les milliers de spectateurs peuvent être si émerveillés par un compte bien mis en forme qu'ils se bousculeront pour envoyer leurs filles à marier dans le clan qui distribue » (Bowers et Lepi, 1975 ; Bowers, 1977). Je cite ce cas parce qu'il est bien connu (Firth, 1965 et surtout Pospisil et Price, 1966).

Je renonce à vous imposer les techniques de pesage des commerçants itinérants, computs astraux, comptages sur des cordes à nœuds, estimations des navigateurs, carrés magiques, devinettes et autres témoins mal connus mais attestés d'une activité qui précède de très loin le miracle grec. Pour un relevé informé de ces archipels, je renvoie le lecteur au dossier compilé par Marcia Ascher dans ses *Mathématiques venues d'ailleurs* (1998).

L'apparition des ordinateurs a encouragé la réévaluation du statut du calcul dans l'activité scientifique, laquelle ne confronte pas seulement théorie et expérience. Ainsi des tables de nombres réciproques babyloniennes, des séries d'éclipses, des calculs d'héritage et autres témoins d'une frénésie numérique et logique reconnaissable au fait que souvent elle dépasse les besoins du problème pratique mentionné.

L'algorithmique est un genre littéraire ancien qui n'a reçu que récemment sa reconnaissance. Des chercheuses ont par la suite pu entreprendre de décrire l'activité mathématique comme un processus de relecture et de reformatage de tables, de listes, d'opérations et autres structures textuelles, et non pas seulement comme une recherche de résultats calculables ou démontrables (Chemla, 2004). Ce travail de mise en forme du raisonnement calculateur appelle parfois une sereine indifférence aux aspérités des calculs locaux. Chez les gens du calcul les nombres grands et nombreux ne sont pas forcément issus de l'expérience : souvent les procédures prolifèrent généreusement en résultats intermédiaires[2].

Ce n'est pas un hasard si la question des données recoupe souvent celle du calcul. Lisons dans les *Nouveaux essais sur l'entendement humain* de Gottfried Wilhelm Leibniz (1705, Livre 4, 7.19) cette remarque qui compare

---

1. De 1958 à 1976, le CERN de Genève a payé les services de Willem Klein, artiste marginal qui avait entre autres mémorisé la table des logarithmes de 1 à 150 (à cinq décimales seulement).

[2] Ainsi, l'algorithme de Héron pour l'extraction de la racine carrée part tout simplement de la définition du problème pour écrire une série convergente qui progressera à coup de divisions : il est donc clair dans son principe et lourd de résultats inutilisables.

les programmes qui voudraient compresser mille pages de droit ou mille pages d'observations biologiques : « Car je crois que la millième partie des livres des jurisconsultes nous suffirait, mais que nous n'aurions rien de trop en matière de médecine, si nous avions mille fois plus d'observations bien circonstanciées. C'est que la Jurisprudence est toute fondée en raisons à l'égard de ce qui n'est pas expressément marqué par les lois ou par les coutumes ».

On trouvera des contre-exemples. On taxera Leibniz de rationalisme, et l'opinion commune demeurera que la biologie brasse des généralités et que les situations juridiques sont infiniment distinctes ; cela pourrait se soutenir quelquefois. Lisons plutôt : « Et les lois de chaque pays sont finies et déterminées, ou peuvent le devenir ; au lieu qu'en Médecine les principes d'expérience, c'est-à-dire les observations, ne sauraient être trop multipliés, pour donner plus d'occasions à la raison de déchiffrer ce que la nature ne nous donne à connaître qu'à demi. ». Il conclut triomphalement : « on perfectionnera l'art de faire de telles observations, et encore celui de les employer pour établir des aphorismes[1] ».

Leibniz a créé le formalisme du calcul différentiel en traitant comme un objet de rêve, de spéculation, de travail et de discussion des listes, et des listes de listes. Les gens du calcul voient ici des colonnes dont la première donne des nombres bruts, la suivante les différences premières entre les items de la première, et ainsi de suite. Ainsi devint lisible la magie des dérivées successives.

Traiter les nombres est un art. Ceux qui y réussissent ne sont pas toujours reconnus comme ayant contribué à la théorie ; et parfois on leur reproche de ne pas avoir l'esprit assez expérimental. Mais ils contemplent les nombres. Aussi n'ont-ils pas toujours besoin de grands tableaux. Le travail qui devait convaincre les médecins que le SIDA est surtout une maladie vénérienne à transmission hétérosexuelle s'appuyait sur moins de quarante observations (Piot *et al.*, 1984).

2.2. L'Internet, puis la Toile

Voici un autre moment singulier : Vannevar Bush, au sortir de la guerre où il avait géré la politique de recherche des États-Unis, esquisse le schéma des temps qui viennent. *The Atlantic Monthly* publie son essai « As we may think » (Bush, 1945), où sont préfigurés les ordinateurs personnels connectés, les documents hypertextes, les bases de données, et surtout les encyclopédies miniaturisées. Il ne s'agit pas ici de romans d'anticipation (comme chez Robida et Verne) ; non plus des prophéties d'un bibliothécaire (Otlet, 1934) : Bush vit dans un milieu d'ingénieurs.

La télécommunication des données constitue un nouveau règne technique. Pendant les années 1970, l'idée de relier les réseaux locaux constitués autour des centres de calcul relevait de l'exploration futuriste ou d'un privilège réservé aux militaires et aux grandes industries. Que le clavier d'un terminal soit connecté avec un ordinateur central, cela allait de soi : le terminal n'étant pas un ordinateur, il n'y avait pas là de communication entre machines. Et quand les premières modalités de communication entre ordinateurs sont apparues, on a d'abord songé à partager leur temps de calcul ou à échanger des logiciels. Un nouveau continent devait surgir. C'était rien moins qu'évident au cours des années 1960 et 1970, et c'est bien pourquoi Bush avait écrit.

Pendant que biologistes, météorologues, astronomes et physiciens échangeaient leurs programmes et parfois leurs données en transportant des bandes magnétiques dans des camionnettes, les spécialistes des

1. Leibniz a correspondu avec d'autres chercheurs, allemands, italiens, anglais et autres, qui compilaient des tables de nombres sur la santé publique et la population. Dans une lettre fameuse au *Journal des Savants* (1694), il invite les médecins à s'associer à ce travail de collecte et de discussion.

télécommunications démarchaient les grandes agences pour financer leurs travaux. Lorsqu'en 1974 apparut la définition d'un protocole d'adressage individuel des machines, le TCP/IP, on parla d'*Internetworking*, en abrégé *Internet*.

3. Données et stratégies

3.1. Vers la structure de l'information

On aborde ici les terminologies qui structureront les données massives. Certains célèbrent un *Data mining* supposé construire, sans architecte, des châteaux en Espagne avec des milliards d'aiguilles collectées dans des bottes de foin non indexées. Ce tableau est un peu forcé. Car pendant ce temps la vérité du travail des professionnels est plus prosaïque, ainsi que l'explique Abiteboul : les machines ne peuvent que compter sur les êtres humains. Ensuite, et très lentement d'abord, elles tenteront de se débrouiller.

Commençons par une remarque de premier ordre. Si l'on prend pour référence les données produites par la recherche scientifique en Europe, les fonctionnaires aiment à répéter que la moitié est « non extractable » et que plus des trois quarts ne sont pas archivés de manière fiable. Inutile de demander aux machines de travailler dans ces conditions, et la course en avant vers les réseaux sociaux et les smartphones ne corrigera pas ce qui reste provisoirement un immense manque à gagner sur des données scientifiques déjà produites.

Écoutons ici quelques suggestions d'un observateur privilégié, Tim Berners-Lee, formulées lors de sa *TED lecture* de 2009 : *The next web*. Chacun sait pourquoi son nom est associé à la naissance du Web : sur une machine NeXT, qui lui assurait un environnement de programmation efficient et confortable[1], il avait eu l'occasion de développer le premier serveur hypertexte pour l'Internet, dont le Cern, où il travaillait alors, était un nœud majeur.

Le schéma du WWW fit le tour du monde scientifique en quelques semaines, tant il répondait à des besoins en passe de devenir cohérents. Parmi les outils retenus, le langage HTML se prêtait mieux à la mise en page des documents et à leurs fonctions hypertexte qu'à la structuration des données. SGML, dont HTML est un dialecte, était conçu pour gérer des informations structurées dispersées dans des documents connectés. Ce langage a d'abord servi à des industriels comme Boeing, lesquels comptent, il faut le savoir, parmi les plus gros éditeurs de manuels techniques. XML a repris une partie du projet SGML, il a connu de nombreuses applications, et il figure dans les fichiers de paramètres d'usage des smartphones, c'est-à-dire dans leur composante industrielle. Les interfaces homme-machine évoluent en effet souvent plus lentement que la gestion industrielle de l'information. Il existe des tentatives pour leur donner accès aux données structurées, comme le langage de recherche proposé par Yahoo à ses développeurs : developer.yahoo.com/yql/, dérivé de SQL, le Structured Query Language des bases de données relationnelles. Ces essais aboutiront. C'est un des enjeux du débat sur les données massives.

Ce point prend son importance si l'on retient qu'au nom de la communauté scientifique, Berners-Lee insiste : les êtres humains ont besoin de documents bien mis en page, tandis que les machines préfèrent des informations

1. Faute de développer ici l'histoire des gens de l'informatique, on peut au moins mentionner un témoignage de Stephen Wolfram (2011) sur Steve Jobs, qui venait de mourir (« Il a toujours préféré les interfaces utilisateurs aux langages, mais il essayait d'être utile ») ; le langage Mathematica ne devait plus quitter les machines de Jobs. La même notice obituaire a paru dans le *Guardian* un peu plus tard, plus brève, mais rallongée d'un détail que connaissaient les lecteurs du blog : « J'ai rencontré Steve Jobs quand il avait quitté Apple, et travaillait à construire sa machine NeXT ».

convenablement décrites et présentées de manière cohérente. Il appelle donc de ses vœux le développement d'un Internet des informations structurées. La discussion mondiale alimentée par les *Ted Lectures* l'a entendu.

On a compris que l'idée de données massives correspond au croisement entre des données rationnellement collectées et bien définies, et d'autres données collectées de manière désordonnée, souvent informes, mais nombreuses et peu coûteuses.

La structure des bases de données relationnelles reconnue par Codd dans son texte inaugural de 1970 suggère que nous vivons aujourd'hui des surprises du même ordre que celle qui a permis aux lecteurs de Leibniz de ne pas voir en lui un simple émule de Newton quand il a écrit le calcul des dérivées sous la forme *dx/dt.* Les enfants devraient savoir – ils sauront bientôt – que ces bases effectuent sur des tableaux rectangulaires des opérations semblables à celles du calcul ensembliste élémentaire.

Cette splendeur lumineuse n'empêche pas la consultation des banques de données de poser des problèmes techniques. Dans sa leçon inaugurale au Collège de France, Abiteboul (2012 ; voir aussi 2014) donne des perspectives, mais il signale aussi les contraintes sévères du calcul pratique : «…les optimiseurs de systèmes [de bases de données relationnelles] font des merveilles sur des requêtes simples. C'est bien moins glorieux pour les requêtes complexes, par exemple celles qui mettent en jeu des quantificateurs universels comme la question : quels sont les acteurs qui n'ont joué que dans des comédies ? Heureusement, en pratique, la plupart des questions posées sont simples ».

Reprenons avec lui le b-a-ba de cette phénoménologie du monde sensible à l'usage des machines. Il propose, comme le font volontiers ses collègues, une hiérarchie en trois niveaux : données, informations et connaissances :

– Les données sont des mesures élémentaires, souvent des nombres.

– Les informations sont obtenues en structurant ces données pour en dégager du sens, par exemple des tableaux de mesures.

– Les connaissances reprennent les informations, qu'elles permettent d'interpréter. Ce peuvent être des faits, considérés comme vrais dans un univers donné, ou des lois de cet univers (des règles qui lisent ces faits, et peuvent en générer d'autres).

On ne parlera pas d'anthropomorphisme. Ces trois types de données sont des objets informatiques maniables et non pas des robots pensants. Je vous renvoie à la littérature de ce programme de travail.

Pour permettre au lecteur d'évaluer ce qui est en jeu dans ce traitement des données relues et recodées, on peut mentionner ici la classification des signes en indices, icônes et symboles, proposée par Charles Sanders Peirce. Un signe peut présenter plus d'une relation à la situation qu'il évoque. Si l'on veut forcer les contrastes, il peut exister des cas presque purs.

Une trace sur une carrosserie atteste qu'un autre objet a été en contact avec la voiture. À un certain moment du temps, ces deux objets ont été contemporains et contigus. La trace est un indice.

Si le signe présente une ressemblance possible avec l'objet il évoque un héritage possible, un passé commun ou une coïncidence. À ce titre, c'est une icône.

Mais le symbole comme tel n'a pas avec son objet de contact dans un présent factuel, ni de ressemblance possible dans le souvenir : il est associé à un objet ou à une opération en vertu d'une convention arbitraire. Il fait donc plus que renvoyer à un objet, ce que fait assez bien la paire « ressemblance-contact » des deux autres signes. Il relève du domaine des règles. Il permettra à plusieurs niveaux des opérations logiques.

Les conventions symboliques sont tournées vers les futurs ; l'indice atteste une co-présence ; l'image rappelle quelque chose. Ces fonctions ne s'excluent pas, mais peuvent se contrarier ; ainsi, les chiffres romains sont

intuitifs, mais ils se prêtent mal au calcul. La classification de Peirce montre en quoi les symboles ne sont pas des signes comme les autres. On parle donc de calcul symbolique quand un ordinateur fait de l'algèbre[1].

Le symbole introduit dans la vie humaine une historicité spécifique : « Les symboles pullulent… Nous pensons seulement avec des signes. Ces signes mentaux sont d'une nature composée, et nous appelons concepts leur partie symbolique. C'est seulement sur fond de symboles qu'un nouveau symbole peut surgir… Une fois parvenu à l'existence, le symbole se répand. Dans l'usage et l'expérience, sa signification se développe » (Peirce, 1893, je cite les *Collected Papers,*1931, vol. 2, § 298).

Si les machines doivent raisonner avec nous, il faudra peut-être songer à leur donner quelques rudiments des résultats d'un programme de recherche linguistique en pleine explosion, celui qui traite la structure de l'information. Le lecteur se reportera à un texte visionnaire de Manfred Krifka que je paraphrase sommairement : prendre un sujet dans le fonds commun des connaissances partagées par les interlocuteurs pour en dire quelque chose qui soit significatif, et contribuer ainsi à enrichir ce fonds commun, c'est là un mode d'intervention spécifiquement humain. Cette pragmatique *Topic-Comment* fait écho à la pratique archaïque de la coordination asymétrique entre les deux mains dans le travail manuel : l'une cadre et fixe l'objet, l'autre le travaille pour lui apporter les spécifications recherchées[2].

La structure de l'information n'est donc pas une formule toute faite. Le programme de travail qui en traite étudie les signes linguistiques qui marquent la pratique de l'intervention des interlocuteurs dans la conversation commune. Saurons-nous y trouver une place pour les machines et leurs sectateurs ?

---

1. Je signale le choc que provoqua chez les gens du calcul l'apparition de langages comme Macsyma, Maple ou Mathematica. Cette intelligence artificielle n'a pas tranché de question philosophique, mais elle a changé le travail de plusieurs communautés professionnelles – et leur vision du monde.

2. L'auteur ajoute que ce schéma « *Topic-Comment* » n'est d'ailleurs pas la seule façon praticable de communiquer entre soi ou de catégoriser des données ; cela rend cette analyse plus intéressante encore (Krifka, 2007). Pour la démarche et les premiers résultats du programme général d'*Information Structure*, voir par exemple Krifka et Féry (2008).

3.2. Le grand Jeu

Ces remarques éparses devraient convaincre les plus réticents que les ordinateurs auront bientôt mieux à faire que gagner aux échecs ou reconnaître un visage. Une lame de fond se lève et il est raisonnable de dire que les données massives et leur structuration constituent une des technologies de rupture des prochaines années (OECD, 2016, p. 83).

Parlons donc des décisions stratégiques souvent mentionnées aujourd'hui chez ceux qui refusent les méthodes paresseuses de l'extrapolation pour parler du futur. Elles nous rapprochent de la communauté des militaires.

Ceux qui n'ont pas le temps d'annoter le chapitre 6B du livre VIII du *Traité de la guerre* de Clausewitz (1832) peuvent lire le chapitre 1 du livre I, qui parcourt les caractéristiques de la guerre. Le petit paragraphe 24 se termine sur l'injonction : « La visée politique est le but, la guerre est le moyen, et le moyen ne peut tout simplement pas être pensé sans la fin »[1]. Son titre, qui est devenu *La Formule* de Clausewitz, a trompé de nombreux lecteurs : « La guerre est la continuation simple de la politique par d'autres moyens ». Il ne signifie pas que la guerre relaie la politique quand on a épuisé les autres moyens. Moins encore que la politique est suspendue pendant le temps de guerre. À l'inverse, Clausewitz dit que la guerre sans fins politiques n'est que violence aveugle et somme de hasards. Raymond Aron (1976) a longuement et courageusement corrigé les erreurs de lecture de cette maxime, qui ont entraîné de terrifiantes conséquences, et d'abord en 14-18.

Non que le hasard soit exclu, au contraire, il constitue l'élément dans lequel s'exerce le travail du militaire. La guerre, dit-il, est l'élément de l'incertitude ; et dans sa théorie de la guerre (Livre II Ch. 2) figure la belle image de la transformation crépusculaire des données : « un éclairage nébuleux ou lunaire donne aux choses des proportions déformées, une allure grotesque ».

La guerre confronte le décideur à des données incohérentes, fausses ou incertaines. Le militaire n'en sait donc jamais assez. Mais il sait déjà qu'il lui faudra trancher, parce qu'il n'aura pas de bon modèle du rapport des forces en présence et de leur distribution quand il faudra engager tout ou partie des siennes dans la bataille.

Cette philosophie du hasard permet de renverser une autre erreur de perspective, celle qui explique paresseusement l'appétit des militaires pour les calculatrices par leur habitude du calcul. En effet, ce n'est pas seulement parce que le levé de terrain topographique manie un calcul des erreurs, ou parce que la balistique étudie les ellipses de dispersion des boulets, que les militaires sont familiarisés avec l'incertitude et l'accumulation des données. C'est parce qu'à l'inverse ces disciplines, cartographie et artillerie, manifestaient, dans une mer de données, la volonté de toucher juste[2].

Ce n'est donc pas par atavisme qu'on retrouve certains militaires lors de la première apparition des ordinateurs, autour de la seconde guerre mondiale. Après tout, jusqu'au Moyen-âge latin, les Matematici étaient tout simplement les astrologues : ils apportaient leur contribution à la levée des incertitudes du futur. Le lecteur reconnaît ici encore les gens du calcul.

À quoi ressemblera un monde plongé dans les données massives et qui prétend dénoncer la technocratie ? Il nous faudrait pour achever ce tour d'horizon un tableau de la société dite du *Big Data*. Les abus qui accompagnent celle-ci construisent par retouches, souvent impondérables et quelquefois brutales, un portrait du Dorian Gray collectif qu'interpellait Oscar Wilde dans son roman. La magie du peintre est-elle le sujet ?

---

1. Je reprends très rapidement ces phrases fameuses ; voir aussi Herbert Rosinski (2009).

2. Les moyens de cette volonté se propagent : les bilans sénologiques traitent les images selon des algorithmes parfois recyclés de travaux sur la reconnaissance de parties dures (camions) en environnement mou (forêt vietnamienne). Les gens du calcul se parlent entre eux.

Les réseaux sociaux, leurs solutions centralisées, leurs archivages intrusifs, souvent mercantiles et dominateurs, ne sont pas un effet obligé des techniques. Si certains serveurs inquiètent, il ne faut pas pour autant interpeller les gens du calcul, mais peut-être les écouter quand ils en parlent. Pourquoi est-il si bien porté de ridiculiser la maxime des entrepreneurs de Google « *Don't be evil* », au point de pousser ses auteurs à s'excuser ? Ne serait-il pas prudent de se protéger surtout des *fake news* ? Faut-il écrire *fake data* pour être entendu ?

Trouvons donc un tel portrait. Comme souvent dans les données massives, nous avons l'embarras du choix. Posons le problème dans de bonnes conditions. Soit un rectangle sur lequel se projettent des jeux d'équipe : une télévision. Le suivi numérique via des capteurs attachés au corps de chacun des joueurs dans les sports d'équipe (football, cricket, rugby, hockey) est en passe de devenir la norme dans une industrie qui manie des ressources financières notables.

Ces dispositifs, qui prolongent sur un mode inouï les vieux gestes démonstratifs du cirque, utilisent la liste de technologies que nous avons appris à reconnaître (capteurs-communications-calculs-utilisations). Assis devant l'écran de la télévision, nous devinons qu'ils savent collecter des flots de données et les encoder, les lire, les interpréter et les comparer avant que des vagues successives submergent les capteurs et leurs ordinateurs traitants.

Ils font donc l'objet de campagnes d'investissements et d'ouverture de nouveaux marchés dont les effets se font déjà sentir sur le recrutement des joueurs, sur la composition des équipes, et sur les commentaires qui accompagnent la diffusion télévisée. Des firmes spécialisées dans la communication comme Cisco sont prêtes à brasser cet immense marché émergent.

Pendant que les contextes des commentaires s'enrichissent en même temps que les capitaines prédisent mieux les performances, il reste pour le public des moments d'enchantement : quand il y a du jeu dans les calculs, et que se précipite un joueur pour tenter sa chance.

Ce seront ces dispositifs qui nous donneront pour conclure un bon échantillon de la société des données massives. Et ils suggèrent de rendre aux hommes leur place, décentrée mais décisive, devant les machinations diverses dont les promesses mènent le train du spectacle.

3.3. L'homme hanté par les données

Décrire la société des données massives est bien difficile. Je ne l'ai pas promis. J'espère avoir évoqué des hommes peu connus, les calculateurs qui s'y intéressent. Ils partagent des rêves, des chagrins, des archives, des problèmes et des solutions, mais tant que le hasard des colloques et des publications ne leur aura pas donné la chance de se connaître, ils ne sont pas encore une communauté. Elle se constituera au cours des années à venir.

Disons enfin un mot sur « l'homme des données », nouvelle figure de la postmodernité qu'on nous annonce depuis quelques décennies[1]. Il n'aime guère les grands mouvements de la datamasse. Il préfère évaluer des situations simples.

Il a connu les forêts et les marécages, les nuages d'oiseaux et les insectes grouillants. Son corps n'a pas eu le temps d'évoluer pour accompagner les signaux souvent désordonnés qui sourdent des machines. Les

1. La préface de l'essai fameux d'Ulrich Beck (*La société du risque. Sur la voie d'une autre modernité*, 1986) commence par une analyse incisive de l'abus du préfixe post- (post-industrialisme, post-modernisme, post-Lumières). C'est une mise en scène de la thématique du livre, en même temps qu'un avertissement : l'histoire continue, et les problèmes que nous ne résoudrons pas reviendront tôt ou tard.

mécanismes ingénieux de la perception logarithmique qui filtrent dans le corps humain les données de la sensation ne suffisent pas. Les ergonomes qui construisent les tableaux de bord, et les psychologues qui testent les biais d'évaluation nous le confirment : l'être humain a devant les grands nombres une faculté d'intuition faible, souvent biaisée.

Il faut donc que les données massives se structurent avant de submerger ses sens. À défaut, son rapport au corps et aux opérations ne peut que basculer, sur un mode parfois brutal, comme le montre une curieuse page de Claude Lévi-Strauss (1971, 610 sq.), qui digresse d'une discussion sociologique des rapports entre rite et mythe pour déboucher sur la comparaison de l'homme et de l'animal devant l'imprévu :

« Cet homme roule à vive allure … rien ne sollicite … son attention … il s'en remet à l'automatisme du conducteur … pour accomplir des gestes menus [qui] relèvent … d'une seconde nature. Mais que, soudain, un objet … placé sur le siège et oublié … bouge, produisant un bruit inattendu : aussitôt … une tension anxieuse saisit son organisme … dans l'appréhension d'un désordre inintelligible et qui pourrait tourner … au désastre. Dans un laps de temps aussi court, l'inventaire des explications possibles défile, les parades sont mobilisées, la mémoire sommée de remplir son office : l'effet est rattaché à sa cause. Sans doute, ce n'était rien ; et cependant, pour un instant, il aura fallu qu'un système nerveux fait pour un corps ordinaire se mesurât avec les risques inhérents au surcroît énorme de puissance que lui allouait le moteur … on voit [que] l'usage d'un engin fabriqué … rapproche … l'homme de sa condition animale : bien qu'incomparablement supérieures … ses capacités symboliques se trouvent en quelque sorte minimisées par la commande d'un corps artificiel dont la puissance physique surpasse celle de son corps naturel ».

Il s'agit là pour lui d'effets qu'on décrit volontiers sur le mode métaphorique : « Les journaux disent parfois qu'au volant d'une automobile, l'homme redevient une bête fauve… ». Le travail du mythologue ne consiste pas à récuser les images, mais à comprendre leur structuration en reconstruisant le contexte – un peu comme une machine chargée de lire nos documents. Sans aucun doute, dans vingt ans, les machines brasseront-elles des données et des calculs à des échelles pour nous inouïes. Entre-temps, les êtres humains doivent subir les pressions que nos conquêtes techniques permettent de leur imposer jusqu'à ce que nous apprenions à mieux gérer cette écologie nouvelle.

Guy Deniélou, fondateur de l'Université de Technologie de Compiègne, considérait son métier d'ingénieur comme une science. Il m'expliquait un jour des années 1980 qu'il avait déposé un projet d'évaluation des risques pour les salles de contrôle des centrales nucléaires. Soit un hyper-parallélépipède découpé selon les limites des plages de sécurité définies sur chaque dimension de l'espace des états possibles du réacteur. Des lampes rouges s'allument quand le point représentatif de l'état observé franchit une des faces. Le projet de Deniélou était d'allumer d'autres lampes dès que, sans être pour autant sorti de l'hyper-boîte, le point navigue près d'un des hyper-coins. Ayant énoncé cela, il regardait son interlocuteur en bourrant sa pipe.

Devant des données nombreuses, le travail de l'automate est de nous aider à éviter la crise permanente, et à regarder là où il convient. Les meilleurs systèmes d'alarme sont ceux qu'on ne doit pas écouter, pour avoir, après de nombreux calculs, trouvé le moyen de faire sans, puisqu'on en a construit d'autres, moins alarmants. Il va de soi que le problème admet souvent plus d'une solution.

Références

Abiteboul S. (2012). « Sciences des données : de la logique du premier ordre à la Toile », *Leçon inaugurale* prononcée le jeudi 8 mars 2012 à la Chaire d'Informatique et sciences numériques du Collège de France, en ligne : books.openedition.org/cdf/529

Abiteboul S. (2014). « À la découverte des connaissances massives de la Toile », *La datamasse : directions et enjeux pour les données massives*, Conférence de l'Académie des sciences, organisée par Serge Abiteboul et Patrick Flandrin le 18/2/2014, en ligne : www.academie-sciences.fr/fr/Colloques-conferences-et-debats-par-et-pour-la-communaute-scientifique/la-datamasse-directions-et-enjeux-pour-les-donnees-massives.html.

Aron R. (1976). *Penser la guerre, Clausewitz*, Gallimard.

Ascher M. (1998). *Des mathématiques venues d'ailleurs. Nombres, formes et jeux dans les cultures traditionnelles*, Le Seuil.

Beck U. (1986). *Risikogesellschaft: Auf dem Weg in eine andere Moderne*, Suhrkamp, Frankfurt am Main, traduction française *La société du risque. Sur la voie d'une autre modernité*, Aubier, 2001.

Bowers N., Lepi P. (1975). « Kaugel Valley Systems of Reckoning », *The Journal of the Polynesian Society*, 84, p. 309-24.

Bowers N. (1977). « Kapauku numeration: reckoning, racism, scholarship and melanesian countings systems », *The Journal of the Polynesian Society*, 86, p. 105-116.

Brin S., Page L. (1998). « The anatomy of a large-scale hypertextual Web search engine », *Computer Networks and ISDN Systems*, 30, p. 107-117.

Bush V. (1945). « As we may think », *The Atlantic monthly*, July.

Chemla K. ed. (2004). *History of Science, History of Text*, Springer.

Clausewitz C. P. G. von (1832). *Vom Kriege*, Werner Hahlweg (ed.), Bonn, Dümmler. On lira *De la guerre*, tr. fr. J.-B. Neuens, sans négliger les éditions et commentaires plus récents.

Codd E.F. (1970). « A Relational Model of Data for Large Shared Data Banks », *Communications of the ACM*, 13, p. 377-387.

Dreyfus H.L. (2007). « Why heideggerian AI failed and how fixing it would require making it more heideggerian », *Philosophical Psychology*, 20, p. 247-268.

Easley D., Klinberg J. (2010). *Networks, Crowds, and Markets. Reasoning about a Highly Connected World*, Cambridge University Press.

Firth R. (1965). Compte-rendu de Pospisil L. (1963), *American anthropologist*, 67, 1965, p. 122.

Gams M., Paprzycki M., Wu X. (1997). *Were Dreyfus and Winograd right?*, IOS Press.

Grumbach S. et Frénot S. (2013). « Les données, puissance du futur », *Le Monde*, 07.01.2013.

Guilbaud G.-T. (1985). *Leçons d'à peu près*, Christian Bourgeois.

Jacob F. (1981). *Le jeu des possibles*, Fayard.

Knuth D. (1984) *The TeXbook*, Addison-Wesley.

Krifka M. (2007). « Functional similarities between bimanual coordination and topic/comment structure », *Interdisciplinary studies in Information structure* 08, p. 61-96.

Krifka M., Féry C. (2008). « Information structure. Notional distinctions, ways of expression », in: *Unity and diversity of languages*, van Sterkenburg P. ed., Amsterdam, John Benjamins, p. 123-136.

Leibniz G.W. (1694). « Lettre sur la manière de perfectionner la médecine », *Journal des Savants*, p. 162-163.

Leibniz G.W. (1705). *Nouveaux essais sur l'entendement humain*, Garnier Flammarion.

Lévi-Strauss C. (1962). *La pensée sauvage*, Plon.

Margalef R. (1974). « Diversity, stability and maturity in natural ecosystems », *Proceedings of the 1st International congress of Ecology, Structure, functioning and management of ecosystems*, Den Haag, September 8-14, Wageningen, Centre for Agricultural Publishing and Documentation.

Moles A. (1990). *Les sciences de l'imprécis*, Le Seuil.

Monnoyer-Smith L. (2016). *Rapport de préfiguration de la fonction de superviseur général des données du ministère de l'Environnement*, remis à Ségolène Royal, ministre de l'Environnement, de l'Énergie et de la Mer, en charge des relations internationales sur le climat.

Morgenstern O. (1972). *Précision et incertitude des données économiques*, Dunod.

OECD (2016). *Science, Technology and Innovation Outlook*, Paris, OECD.

Otlet P. (1934). *Traité de documentation : le livre sur le livre, théorie et pratique*, Bruxelles, Éditions Mundaneum.

Peirce Ch.S. (1931). *Collected Papers*, Cambridge, MA: Harvard University Press.

Piot P, Taelman H., Minlangu K.B., Mbendi N., Ndangi K., Kalambayi K., Bridts C., Quinn T.C., Feinsod F.M., Wobin O., Mazebo P., Stevens W., Mitchell S., McCormick J. (1984). « Acquired immunodeficiency syndrome in a heterosexual population in Zaire », *The Lancet*, 324, p. 65-69.

Pospisil L. (1963). *Papuan economy*, Yale University publications in anthropology.

Price D.J. de Solla, Pospisil L. (1966). « A Survival of Babylonian Arithmetic in New Guinea? » *Indian Journal of History of Science*, 1, p. 30-33.

Rosinski H. (2009). « La structure de la stratégie militaire », *Stratégique*, 97-98, p. 17-50.

Verdier H. (2015). *Les données au service de la transformation de l'action publique. Rapport au Premier ministre sur la gouvernance de la donnée*. Administrateur général des données, secrétariat général pour la modernisation de l'action publique, services du Premier ministre, Paris, La Documentation française. Voir le site agd.data.gouv.fr

Wolfram S. (2011). « Steve Jobs: A Few Memories », consulter le site : blog.stephenwolfram.com/2011/10/.