

Big data : no best way

Christophe Bénavent
Université Paris-Ouest Nanterre La Défense

La technique vient par vagues et, pour paraphraser un de ses philosophes, se concrétise en cascades d'objets dont la vie tient moins à leurs fonctionnalités qu'à la valeur qu'en tire le marché. Le big data est au mieux le nom donné à la dernière de ces vagues, qui charrie un ensemble divers de technologies, de concepts, plus ou moins aboutis, reliés, co-déterminés. Certains de ces ensembles sont en voie d'intégration profonde comme les *Digital Management Platforms*, dont celle du *cloud*, les machines virtuelles, les bases *hadoop*¹, les architectures de serveurs ; d'autres restent encore largement désagrégés comme les objets connectés ; d'autres enfin demeurent à l'état de briques désordonnées comme c'est le cas de l'*open data* ou du *Vendor Relationship Management* (VRM).

De ce point de vue, c'est plus une couche supplémentaire qui s'ajoute, participant à un modèle centralisé qui ressemble d'une certaine manière au modèle des *mainframes* recentralisant données et calculs. On restera conscient qu'un modèle alternatif, aujourd'hui silencieux, pourrait s'exprimer.

Sans aucun doute, les discours professionnels entretiennent l'illusion d'une pure instrumentalité qui en célèbre la puissance et les ressources (il faut relire le texte de Chris Anderson de 2008 ainsi que, de manière plus nuancée, McAfee & Brynjolfsson, 2012) mais omettent les limitations intrinsèques des données à l'exception de Michael Jordan (Gomes, 2014). Le symétrique de ce point de vue est le récit apocalyptique d'une société de surveillance qui limiterait nos libertés, notre pensée ou l'espace public (un bon exemple de ce type d'analyse d'inspiration foucaldienne est par exemple le texte de Zeynep Tufekci, 2014). Un troisième point de vue est celui du *pharmakon*, désormais classique en matière de technologie : sans vertu particulière, c'est le contexte et la relation qui en définiraient les propriétés. Sa version critique est largement défendue par le projet organologique de Bernard Stiegler (2014). Nous pourrions aussi favoriser un autre point de vue, celui défendu en son temps par Ciborra et Lanzarra (1999), qui voient dans la technologie un corps étranger que l'organisation doit accueillir. À vrai dire, la littérature académique en sciences de gestion reste relativement silencieuse. La revue *Management Information System Quarterly* n'a publié un appel à communication que depuis le mois d'octobre de cette année (Goes, 2014). En *marketing*, comptons le programme de recherche proposé par Weinberg *et al.* (2012). On complétera ce rapide aperçu en mentionnant l'histoire personnelle du big data de Diebold (2012).

Le commentaire relatif à cette vague technique se partage entre critiques et prosélytisme. Pour échapper aux sirènes de l'utopie et aux injonctions de l'idéologie,

1. Système de fichiers conçu pour stocker des données très volumineuses et facilitant la création d'applications échelonnées et distribuables.

il est nécessaire de revenir aux rapports que la technique entretient avec les organisations. Les problèmes posés par le big data sont celui des conditions sociales qui en permettent l'expression et celui des nouvelles formes d'organisations qu'il l'autorise.

Ce point de vue s'oppose à ce qui est devenu une thèse classique : celle du co-alignement qui suppose que la performance vient moins des propriétés spécifiques de la technologie et de la stratégie que de leur mutuel ajustement (le *fit*) dont on ne sait s'il vient de la parfaite adéquation des formes comme les pièces d'un puzzle ou de l'élasticité d'un bas sur une jambe. Dans la pensée de Ciborra et Lanzarra, et au-delà dans celle de Simondon (1958), il y a cette idée que la technique, au cours de sa concrétisation, s'autonomise, et que, par conséquent, son intégration dans l'organisation nécessite des adaptations de cette dernière, et si celles-ci ne sont pas possibles, suscite d'autres organisations. C'est cette interrogation qui est au cœur de cet essai.

On s'attachera en premier lieu à l'examen du phénomène technique pour envisager dans un second temps ses modèles d'usage et conclure sur les questions que soulèvent ces modèles.

Data : de leur obtention à leur distribution

Si l'on comprend parfaitement désormais ses caractéristiques fonctionnelles – mettre à disposition de manière presque instantanée le résultat de requêtes sur des ensembles qui peuvent comprendre une grande variété de formats et atteindre des tailles telles que ce sont des architectures multi-serveurs qui sont nécessaires pour les maintenir (les fameux 3V) –, cette représentation du big data ne dit rien quant à l'utilité des données dans les processus d'affaires.

Au-delà de la volumétrie, dont les informaticiens ont trouvé les solutions techniques (qui ne se limitent pas à *MapReduce*² et *hadoop*), c'est par le processus des données que l'on peut en avoir une meilleure compréhension. Ce processus s'appuie d'abord sur la prolifération des capteurs et sur l'intensification du processus de capture. Si, à l'évidence, ce mouvement de capture est le propre des plates-formes de données sociales dont l'intérêt premier est de capturer le flux de conversations partagées par des centaines de millions d'internautes, mais aussi de rares entreprises (*telecom* et *utilities*) pour lesquelles la mesure fait partie du processus opérationnel, il promet de s'étendre avec l'Internet des objets.

La diversité des capteurs est certainement le phénomène le plus important. La question des méthodes de stockage, même si elles sont spectaculaires de par la distribution sur des milliers de serveurs d'une information qui se compte en plus de pétaoctets (10^{15}) par période, est finalement secondaire. À vrai dire un téraoctet (10^{12}) tient dans une pièce de monnaie. Ces solutions – *hadoop* est un des termes les plus fréquemment cités – quittent aujourd'hui les vaisseaux amiraux des données (GAFA – Google, Apple, Facebook, Amazon) pour envahir les cales de tous les systèmes d'information. Les capteurs, quant à eux, saisissent désormais en continu les variations d'états des clients et du marché. Si les données de comportements – achat et réactions aux messages publicitaires – sont acquises depuis une vingtaine d'années, les conversations le sont depuis une dizaine d'années. Et, avec la montée en charge du parc des *smartphones* et tablettes, ce sont les données géolocalisées qui ont le vent en poupe. De nouvelles générations d'appareils livrent désormais des données

2. *MapReduce* est un modèle d'architecture de développement informatique, inventé par Google, qui permet de traiter des données potentiellement très volumineuses en les distribuant dans des *clusters* parallèles.

de mouvement et de température, voire la composition de nos humeurs. Ces données physiologiques renseignent sur l'état de nos corps et de leurs satellites : voitures, maisons, et même animaux domestiques. Ce seront plus tard nos visages, ou nos émotions.

Ces données nécessitent ensuite d'être corrigées, redressées, filtrées pour être modélisées, simulées et finalement introduites dans des systèmes de décisions et d'actions. En effet, on restera attentif au fait que la donnée dite brute est produite par des processus qui ne sont pas toujours parfaitement maîtrisés. Il peut y avoir des biais de sélection lorsqu'une partie spécifique d'une population est concernée, des interruptions dans les séries, parfois même des éléments erronés, sans compter les erreurs de capture. L'exemple des données provenant des réseaux sociaux est le plus simple à comprendre. Le biais de sélection peut être très important dans la mesure où la motivation des individus les plus actifs est très différente de celle des *lurkers* (ceux qui se contentent d'observer), sans compter les jeux de masques, omissions et autres travestissements auxquels les comptes sont amenés ne serait-ce que par la pression sociale à laquelle ils sont exposés. C'est certainement la plus grande difficulté de l'analyse de sentiment avant même les difficultés linguistiques de la désambiguïsation.

On comprendra que la donnée brute n'existe pas. L'utilité de la donnée provient bien plus du sens qu'on lui donne que de sa vertu de trace. Cette production de sens se réalise d'ailleurs dès son recueil : nos instruments ne captent pas accidentellement des variations d'états, ils le font dans le cadre d'un projet. Construites dès l'origine, les données sont transformées, reconstruites au cours du processus, jusqu'au moment d'être distribuées. Elles ne sont jamais tout à fait ce qu'elles représentent car elles sont déjà une transformation de ce qui est et de ce qu'elles sont.

Au cœur de cette question, c'est celle de la valeur des données qui est posée. À l'évidence, cette valeur n'est pas intrinsèque, même si un marché des données se développe depuis longtemps notamment par les pratiques *marketing*. Elle se révèle dans des conditions particulières et concerne, non pas la donnée dans sa généralité, mais des cas particuliers : une adresse pour envoyer un message, un signe d'intérêt pour déclencher une proposition. Il apparaît très souvent que la valeur réside en fait dans sa confrontation à d'autres, sa combinaison à d'autres sources, et on découvrira rapidement qu'il faut la penser non pas comme un stock mais un ensemble de flux entrants et sortants de l'organisation. On pourra défendre finalement l'idée selon laquelle la valeur réside essentiellement dans sa transformation : détecter un marché ou une tendance avant les autres.

Le fait de distinguer les flux des ressources conduit à attirer l'attention sur les processus de transformation des données en indicateurs par le biais de modèles arithmétiques, statistiques ou stochastiques. Ce qui conduit naturellement à s'interroger sur les conséquences en termes de médiatisation. Le big data ne serait finalement que l'apparition d'un nouveau média et d'un nouveau type d'auteur. Un auteur industriel.

Fernando Pessoa
café A Brasileira,
Lisbonne, 7 mars 2010



Ce sont ces données transformées qui sont généralement introduites ensuite (mais pas obligatoirement) dans les processus de modélisation qui peuvent être destinés à des finalités très différentes. Certains visent à donner du sens, d'autres sont purement opérationnels. Les processus destinés à donner du sens peuvent être purement descriptifs : c'est le cas des vieilles techniques de classification, de la cartographie ou encore de l'analyse descriptive des réseaux. Ou bien ces processus peuvent être plus compréhensifs quand des méthodes inspirées de l'analyse économétrique sont mises en œuvre. On se contentera de remarquer ici qu'une tâche peu automatisable intervient et risque d'être un goulot d'étranglement. Pour donner du sens il faut des cadres d'analyse, une connaissance des modèles, avoir développé des catégories, les avoir négociées avec ceux qui fournissent la donnée (les usagers), ceux qui en donnent la mesure (les référentiels), ceux qui en définissent l'ordre (la puissance publique). La statistique n'est pas que du calcul comme le montrent notamment les travaux publiés de manière posthume d'Alain Desrosière (2014). Le nombre a aussi un caractère performatif.

Quant aux méthodes plus opérationnelles, qui se confondent en grande partie avec les méthodes dites prédictives, et auxquelles les techniques de *machine learning* contribuent largement, elles se réduisent généralement à probabiliser l'état d'une variable simple et à construire des « classificateurs ». Un bon exemple de ces méthodes peut être donné avec les moteurs de recommandations pour lesquels Netflix a lancé un concours richement doté en 2009 et arrêté en 2011. De cette expérience on apprend d'une part que les gains en précision ont été faibles en dépit de la participation de dizaines d'équipes et, d'autre part, que la méthodologie employée est à la fois itérative et lente : la construction d'un modèle prend plusieurs mois, et l'évolution constante des données et des comportements conduit à réapprendre aux machines de manière opiniâtre.

Ce dernier exemple montre que l'enjeu n'est pas dans la technique (bien maîtrisée par les spécialistes) mais dans l'imagination et l'assujettissement d'un effort important à la résolution d'un problème extrêmement précis. De ce point de vue, notons que l'immense réussite d'un Google et de son modèle d'affaires, réside dans une donnée simple : les mots que l'on tape dans les moteurs de recherche portent une information précieuse, nos intentions. De même, le succès d'une entreprise comme Criteo (firme française de ciblage publicitaire personnalisé sur Internet créée en 2005) et la performance de ses services, tiennent à ce qu'on considère le clic comme cette même intention. Rien de très complexe, le nombre faisant que les calculs de probabilité deviennent possibles à travers des millions d'éventualités. Prenons un service simple : la suggestion de mots dans le moteur de recherche. Elle se forme simplement dans un calcul de probabilité conditionnelle : quelle est la probabilité de saisir le mot X sachant que les lettres xyz viennent d'être saisies ? La réussite tient dans l'échelle qui élève les faits les plus rares à une régularité probabilisable. Dans de grandes masses, les *serial killers* deviennent des régularités. Le big data n'est pas qu'affaire de moyenne mais l'opportunité d'observer des événements rares. Le singulier rejoint alors le général.

C'est à ce moment que le second aspect du big data se découvre : ses modes de distribution. C'est un aspect finalement moins développé que le premier mais qui s'avère déterminant. Il prend corps aussi bien dans les techniques de visualisation que dans la production de tableaux de bord. Il pose une question essentielle, celle de la manière dont la multitude s'en empare. Cet aspect des modes de distribution instaure

le fait que, dans les organisations et la société dans son ensemble, une réflexivité (dont on peut s'inquiéter dans la mesure où elle sélectionne des indicateurs) réifie certains phénomènes sociaux et conditionne les décisions, encore que reste à identifier la manière dont ces informations sont prises en compte par les acteurs.

Cette distribution prend deux caractères : l'étendue et le degré de spécificité. L'étendue vient d'être analysée. Le degré de spécificité rejoint cette vieille idée d'implicite/explicite au moins dans le sens d'abstrait/concret. Elle pose la question de la nature de la connaissance : à fort contexte comme l'est le renseignement policier pour lequel l'information prend sens dans un contexte très spécifique, ou à faible contexte quand elle se présente comme connaissance générale, telle que pourrait l'être une carte démographique.

Le big data, sans accroître la profondeur ou l'étendue de la connaissance, peut au moins donner une autre échelle à sa diffusion et à son opérationnalisation. Il généralise le renseignement et organise la vision du marché.

Du contrôle/commande³ aux organisations en plates-formes

Nous avons eu longtemps une vision agrégative. Les données parcellaires, désagrégées, en s'accumulant forment une image qui en montre les variations dans l'espace et le temps et dont un œil averti peut décrypter les variations et les causes. Dans ce domaine, la magie du nombre réside dans la finesse de l'image, dans sa granularité accrue. Les cartes sont plus fines, mieux circonstanciées, les séries temporelles prennent toutes les fréquences, de la seconde au siècle. Cependant, ce modèle d'usage ne diffère pas du modèle traditionnel des études. Il donne à un petit nombre d'experts et d'analystes des éléments précieux pour élaborer des stratégies qui seront mises en œuvre ultérieurement par les opérationnels. Dans ce domaine, le gain procuré par le très grand nombre de données et leur diversité restera à la marge : plus de finesse et plus de précision, parfois une incursion dans des domaines d'observation autrefois difficiles à obtenir. Mais aucune révolution en vue.

Le problème central de cette perspective fait réapparaître un très vieux débat, celui d'une statistique descriptive qui s'oppose à une statistique inférentielle. La première trouve dans les structures mathématiques le moyen de faire apparaître les structures réelles, de dévoiler une réalité rendue confuse par les variations aléatoires. La seconde affirme qu'on ne rend compte du réel que si l'on en possède un modèle, l'épreuve empirique visant à tester le modèle, et éventuellement à le paramétrer. Les techniques de *machine learning* rejoignent de ce point de vue les approches de Benzécri (1973) en reprenant d'ailleurs les mêmes techniques : calcul de distances et décomposition de la valeur singulière des matrices, dans un contexte technique où le nombre de lignes et de colonnes concernées relève de plusieurs milliers, et où les vides prennent une proportion considérable. De l'autre côté, les techniques multi-agents.

À mi-chemin entre les deux postures auront émergé, mais bien avant que l'expression big data ne s'impose, les méthodes prédictives, qui, ignorant la spécification conceptuelle des modèles, tentent de prédire l'état d'une variable au travers d'un grand nombre d'autres critères. De l'analyse discriminante aux réseaux neuronaux, du modèle de régression logistique à la régression lasso, au cours des dernières années nombreuses sont les méthodes proposées qui fonctionnent avec une sorte de boîte noire et sont évaluées essentiellement en termes d'ajustement prédictif. Le *marketing* digital est devenu gourmand de ces techniques.

3. Système de contrôle d'un procédé industriel présentant une interface homme-machine permettant la supervision, et un réseau de communication numérique.

Mais, au fond, les méthodes ne sont pas en cause, c'est bien le mode d'usage qui doit être interrogé, et le fait de s'appuyer sur les premières masque sans doute un renoncement : celui d'une compréhension des phénomènes dont on tente de rendre compte.

Les gains espérés dans ce modèle ne seront certainement pas aussi grands que prévu, et dépendront largement des secteurs. Ils peuvent même être négatifs si l'on considère le coût des investissements. Mais le plus important est que ce modèle, en réalité, renforce les organisations existantes sur le mode du contrôle/commande, encourage les systèmes de management par la performance, participe à une centralisation et ne requiert au fond que le développement d'une sorte de salle de contrôle ou de marché. Les organisations du *yield management* (les compagnies aériennes notamment) en sont un exemple remarquable.

Dans ce modèle, les *data scientists* sont une nouvelle forme d'analystes auxquels on demande moins une capacité d'innovation qu'une aptitude à travailler vite et à accroître leur productivité. Les solutions techniques proposées vont dans ce sens : la mise à disposition de vastes morceaux de données et de moyens de calculs permettrait ainsi de travailler un plus grand nombre de segments en conservant un niveau de centralisation élevé. Ce modèle connaît cependant une limite : ses gains sont incrémentaux puisqu'il améliore les processus existants sans exiger une réforme de structure. La décision reste lente même si elle est plus précise. Le big data reste

alors cantonné à un département même s'il prend la forme d'une *start-up* interne. Ce modèle permet de vendre des données transformées, affinées, mais encore trop statiques.

Un deuxième modèle est celui de l'automatisation. L'information construite, transformée, plutôt que d'être livrée à des agents humains qui sont responsables de son évaluation, peut se transformer automatiquement en décision et en action. Le calcul d'un taux d'intérêt en fonction d'un score de risque, l'affectation d'une publicité sur une page en fonction d'un algorithme de bandits manchots, l'affichage d'un prix sur une étiquette électronique piloté par un modèle économétrique de prix. Le big data est aussi une machine à produire de la micro-décision, des millions de micro-décisions. L'exemple des prix est éloquent : dans une chaîne de 500 hypermarchés, en comptant 20 000 références par point de vente, il y a 10 millions de calculs de prix à réaliser de manière journalière. Cela donne une idée de l'ampleur de l'opération.



Pessoa,
Julio Pomar

Sur un plan pratique c'est ce modèle dit de RTB (*Real Time Bidding*) qui semble percer dans le marché publicitaire digital et qui prend jusqu'à 30% de parts de marché. Le RTB consiste à allouer une impression publicitaire à un annonceur et à en fixer le prix en fonction de caractéristiques telles que la taille de la bannière, le contexte de la page *Web* ou encore l'heure et le lieu où elle est visualisée, le tout en temps réel. On se rendra compte rapidement que ces nouvelles entreprises qui ont des tailles de quelques centaines de personnes emploient pour moitié des ingénieurs dévoués à la gestion des Systèmes d'Information (SI), au *data mining* et au développement d'applications. C'est un modèle monocanal car il exige une grande concentration technique dont les rendements s'expriment de manière idéale dans un univers technologique bien

défini, un choix de système, de langage, de méthodes qui peuvent être partagés. Le cas d'Amazon est exemplaire, on en trouvera bien d'autres.

Une troisième voie laisse une part à l'humain. L'information produite peut alors être restituée à la diversité et à l'étendue de l'audience (vendeurs, clients, fournisseurs) sous des formes variées qui tendent cependant toutes à rendre compte des activités que ces acteurs infèrent. En rendant compte de l'activité de tous, on peut espérer que les différents acteurs prennent les meilleures décisions pour eux et pour la collectivité. Ce modèle d'usage est celui de l'*empowerment*. Le nombre des destinataires définit la variété. Le grand nombre, qui rend possible une granularité très fine, permet de donner à chacun ses données les plus personnelles et locales et d'alimenter des tableaux de bords individuels : consommation, effort, bien-être, épargne... La vitesse de traitement et de mise à jour permet à chacun dans l'organisation d'avoir des compteurs de son activité et, par conséquent, de s'auto-réguler.

Le sens du big data est de fournir chacun en information personnalisée – le jugement de crédibilité d'un consommateur à l'égard d'une note, l'interprétation donnée par un chef de rayon à la variation journalière des ventes, l'analyse par un élu des statistiques communales – destinée à éclairer les plus petites décisions. Les marchés de réputation semblent diriger leur économie. Quand une plate-forme telle que booking.com génère sans doute des dizaines de millions d'évaluations, ce n'est pas un simple moyen pour chaque hôtel d'améliorer son service, c'est aussi une façon d'offrir en priorité aux consommateurs les établissements les mieux notés. Ici une information simple oriente le comportement de dizaines de millions de clients et de dizaines de milliers d'établissements.

Ceci pose deux problèmes importants : celui de l'apprentissage (savoir jouer des indicateurs pour agir) et celui du *sensemaking* (la capacité à se donner plus qu'un objectif, un projet).. En termes d'organisation, ce modèle pose une question de motivation. Le risque est d'enclencher un système de gratification fait uniquement de stimulants extrinsèques (des commissions et des médailles) qui diminuent le degré d'intégration, une perte des activités extra-rôles, pro-sociales, gratuites, sans lesquelles la force de l'organisation ne se fait pas sentir : perte de civilité organisationnelle, réduction du *stewardship*, évanouissement de toute bienveillance. Cet *empowerment* passe par une distribution raisonnée de la transparence. Quand cet équilibre est atteint, il redéfinit la frontière de l'organisation qui va au-delà des employés, couvrant la clientèle en les engageant. Dans sa forme exacerbée, c'est celui du *Quantified Self*. Pour le commun, cela prend les formes du *self-care*.

Dans sa forme simple, qui laisse une part importante du travail de prescription à des agents humains, l'essentiel est que l'information tienne plus au renseignement qu'à la courbe de performance. Quand les tableaux de bord servent les cadres, l'information qu'ils doivent fournir à leurs agents, ceux qui sont au front, ne consiste pas tant à retracer l'information qu'à livrer un *prospect* à un vendeur, à signaler un risque, à avertir d'une action d'un concurrent, à préparer un rendez-vous. Bref, à agir mieux qu'ils ne pourraient le faire sans cette assistance.

On peut aller encore plus loin et envisager que les données puissent être transformées en services. Il ne s'agit plus seulement de fournir tableaux de bord et dispositifs d'actions, mais d'offrir les solutions, de prendre les devants : l'application Google Now, sans qu'on ne lui demande rien, donne les renseignements basiques de la vie quotidienne (la météo, la carte, les vols à prendre, les rendez-vous, des options de

restaurants et de visites). La connaissance de la localisation, du calendrier, quelques index de sources importantes, un peu de *search* et de hasard, font que le consommateur n'a même plus à chercher, il trouve. Le service est une anticipation littérale du besoin et il faut y répondre avant qu'il ne s'exprime.

C'est sans doute la voie la plus innovante. L'exemple émergent est celui du *coaching*. Les bracelets de *fitness* qui se diffusent aujourd'hui de manière massive, même si Nike vient d'abandonner la production au profit d'un investissement accru dans le *software*. Il y aurait 5 millions d'unités livrées en 2013 et près de 300 millions prévues en 2018. Dans ce modèle, on devine un *coaching* digital et une multitude de questions qui se posent.

Le cas de la consommation collaborative est particulièrement intéressant. Ce modèle économique se construit dans la capacité à mobiliser des actifs sous-utilisés – ce qui permet de se passer de capital, tout en assurant l'appariement d'offres et de demandes très hétérogènes. Il joue sur une économie de diversité, et se déploie dans des marchés à plusieurs versants. Ce qui permet de jouer sur ces caractéristiques est la maîtrise des données au service de la mise en relation. Assurer la confiance, ajuster les offres et les demandes, coordonner à une échelle de dizaines voire de centaines de millions d'individus, résulte d'un processus de traitement de données continu dont les temps de réponse sont de moins de 100 millisecondes.

Dans ce modèle, le cœur du problème est d'assurer que, dans cette diversité, une valeur supplémentaire soit apportée par l'ensemble, une sorte de bien commun, et que les comportements induits par le système d'incitation ne puissent détruire cette valeur. Un exemple simple est celui des sites de rencontres qui nécessitent un équilibre et où il s'agira de limiter les comportements excessifs des hommes pour mieux se conformer aux attentes des femmes. Un autre exemple est celui des plates-formes de co-voiturage où il s'agira de contenir les offres de transports « professionnalisés » pour garder un esprit de partage. Le big data sera employé ici à développer les systèmes de réputation, les signaux de confiance, les dispositifs de *nudges* (d'incitations) et l'ensemble des mécanismes de prescription. Dans les plates-formes, le *marketing* devient une véritable police dans un sens double : à la fois celui d'un maintien de l'ordre et celui du règlement intérieur, qui est une traduction incomplète et plus générale de l'idée des « *policies* ». Le big data se confond alors avec l'organisation, il est ce qui régule, incite, anime. Il est le moyen par lequel la règle s'exerce mais aussi ce qui fait la règle quand celle-ci est issue des algorithmes et de leur politique.

C'est certainement le modèle le plus ambitieux, le plus intégré, un modèle politique dans lequel justement ce que certains dénomment la « politique de l'algorithme » joue à plein régime. Les compétences requises ne sont pas simplement informatiques et statistiques, elles sont aussi juridiques et politiques et sont mises au service d'une véritable architecture de la décision dans le plein sens de Thaler et Sunstein (2010).

Un retournement des données ?

On comprendra que, dans ces quatre idéaux-types, ce ne sont ni les mêmes compétences ni les mêmes ressources qui sont requises. Les buts sont aussi différents et les enjeux ne portent pas sur les mêmes objets. Le seul fait de les mettre en évidence suffit à démontrer qu'il est vain d'attendre du phénomène une sorte de nouveau « *one best way* » de la gestion.

Ces modèles posent aussi naturellement une question de vie privée et d'intimité à des degrés divers. Ce point est suffisamment développé pour que nous puissions être rapides. Rappelons simplement qu'une violation grossière du sentiment que le public a de ses droits pourrait rompre la confiance minimale qu'on accorde aux entreprises qui collectent ces données, même si, en la matière, un paradoxe de la vie privée est identifié depuis longtemps. Du côté des entrepreneurs, on peut donc s'attendre à un usage raisonné de ces données, moins du côté des États.

On souligne plus rarement que le paysage politique et juridique définit aussi l'espace de ces technologies : le droit limite les rapprochements de données, les degrés d'anonymisation, façonnant la structure même de ces données tant dans leurs lignes (leurs index) que leurs colonnes (leurs attributs) ; les différences de droit d'un pays à l'autre, d'une région à l'autre peuvent ainsi favoriser de manière différentielle certains modèles et orienter l'innovation dans des directions divergentes.

Le droit, et les politiques qui l'inspirent, ne sont pas la seule force en jeu. Les pratiques exercent aussi une influence certaine. L'exemple de Adblocker est sans doute le plus remarquable à présent, tout autant que l'utilisation désormais massive des *anti-spams*. Plus généralement le mouvement VRM lancé par le Berkman Center de Harvard, même s'il est jusqu'à présent décevant en dehors des deux exemples cités, reste potentiellement un élément clé du paysage, et les pratiques qu'il induit peuvent conduire à des alternatives majeures. On apprend ainsi par des exemples limités, comme celui de la restitution de données personnelles (MesInfos.fing.org), portée par la FING (Fondation Internet Nouvelle Génération), que la production de services s'appuyant sur le croisement de données provenant de sources distinctes peut très bien s'opérer à l'échelle de l'individu. Ces *self-data*, s'appuyant sur des *clouds* personnels, peuvent ouvrir à d'autres formes de calculs, d'autres modèles de données, et d'autres *business models* ■

Références

- Anderson Chris (2008) "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired Magazine*, n° 16-07.
- Benzécri Jean-Paul (1973) *L'Analyse des Données. Volume II, L'Analyse des Correspondances*, Paris, Dunod.
- Ciborra Claudio & Lanzarra Giovan F. (1999) "Hospitality and IT", in Ljungberg Fredrik [ed] *Informatics in the Next Millennium*, Lund (Sweden), Studentlitteratur, pp. 161-176.
- Desrosières Alain (2014) *Prouver et gouverner. Une analyse politique des statistiques publiques*, Paris, La Découverte (coll. Sciences humaines).
- Diebold Francis X. (2012) "A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline", PIER Working Paper 13-003 Second Version, <http://ssrn.com/abstract=2202843>.
- Goes Paulo B. (2014) "Big Data and IS Research" Editor's comment, *MIS Quarterly*, vol. 38, n° 3, pp. iii-viii.
- Gomes Lee (2014) "Feature RoboticsArtificial Intelligence. Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts", posted 20 October, <http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts>.
- McAfee Andrew & Brynjolfsson Erik (2012) "Big Data: The Management Revolution", *Harvard Business Review*, vol. 90, n° 10, pp. 60-68.
- Simondon Gilbert (1958) *Du mode d'existence des objets techniques*, Paris, Aubier.

- Stiegler Bernard [ed] (2014) *Digital studies: organologie des savoirs et technologies de la connaissance*, Limoges, FYP Editions.
- Thaler Richard H., Sunstein Cass R. & Balz John P. (2010) "Choice Architecture", Working Paper April 2, <http://ssrn.com/abstract=1583509>.
- Tufekci Zeynep (2014) "Engineering the public: Big Data, surveillance and computational politics", *First Monday*, vol. 19, n° 7, 7 July.
- Weinberg Bruce D., Davis Lenita & Berger Paul D. (2012) "Perspectives on Big Data", *Journal of Marketing Analytics*, vol. 1, n° 4, pp. 187-201.



*Reconstitution du navire de Magellan,
Barcelone, 16 novembre 2014*