

grand public et pour l'instant, ce qui est le plus apparent du *Big Data*, c'est l'utilisation de données personnelles par de grandes entreprises commerciales, principalement pour des publicités ciblées. C'est le cas de *Google*, qui analyse les requêtes et les courriels des internautes pour mieux cibler sa publicité, ou d'*Amazon*, qui suggère aux usagers des livres à acheter.

Des armes nouvelles aux mains des dictatures... et des citoyens

Récemment, en particulier dans le cadre de l'affaire Edward Snowden aux États-Unis, la presse a souligné que divers gouvernements utilisent l'analyse de données privées à des niveaux surprenants. La principale raison invoquée pour ce type d'utilisations est la lutte contre le terrorisme. Mais le système PRISM de la NSA (*National Security Agency*, l'agence américaine de sécurité) et les systèmes équivalents des autres États sont aussi utilisés pour l'espionnage, notamment industriel. Ils peuvent l'être pour surveiller les opposants politiques. Le contrôle et l'analyse des données sont certainement les nouvelles armes parmi les plus inquiétantes des dictatures et des gouvernements totalitaires.

A contrario, la généralisation dans les pays démocratiques de l'*open data* (l'accès libre aux données du secteur public) devrait permettre aux «journalistes de données», et plus généralement aux citoyens concernés, de contrôler les actions de leurs gouvernants ainsi que celles des grandes entreprises. Les mêmes technologies rendent possibles la surveillance des personnes par l'État et celle de l'État par les citoyens. Les gouvernements de pays tels que les États-Unis, le Royaume-Uni, ou plus récemment la France se sont engagés dans le mouvement de l'ouverture des données. Cela ouvre la voie à de nouveaux services et permet aussi de mieux contrôler les actions des gouvernements et des grandes entreprises. On peut espérer que cela conduise à donner plus de responsabilité au citoyen et à refonder la démocratie.

Les technologies du *Big Data* semblent particulièrement adaptées pour prévoir des crises sanitaires, des problèmes environnementaux, des catastrophes naturelles, et pour y réagir. Elles devraient aider à résoudre les problèmes de santé, de transport, d'écologie, à lutter contre la pauvreté.

Hadoop pour gérer les données massives

Hadoop, logiciel libre de la fondation *Apache*, a été conçu en 2004 par l'Américain Doug Cutting. Ce logiciel est adapté à l'analyse de grandes masses de données. Il est fondé sur la technique *MapReduce*, que *Google* a utilisée pour son moteur de recherche.

Dans un calcul *MapReduce*, on commence par découper le problème en de nombreux sous-problèmes indépendants (étape *Map*) que l'on confie à

des ordinateurs distincts. Ces machines résolvent les sous-problèmes et envoient leurs résultats à d'autres machines qui ont pour tâche de combiner les résultats (étape *Reduce*). L'objectif est de pouvoir travailler sur d'énormes volumes de données en parallélisant les calculs sur des grappes d'ordinateurs. *MapReduce* ne permet de traiter que des problèmes qui peuvent se décomposer en de multiples tâches parallèles.

Le logiciel de *MapReduce* le plus populaire est *Hadoop*, à la base des centres de données de géants du Web tels que *Amazon* et *Facebook*, et on le retrouve de plus en plus dans des offres de « Cloud computing ». Cependant, malgré ses atouts, *Hadoop* est un logiciel encore jeune et imparfait : il n'utilise pas un langage standardisé, ce qui rend sa gestion complexe, et ne réalise qu'un traitement différé, et non en temps réel, des données.

Dans nombre de ces situations, il faut combiner des analyses de grandes masses de données stockées et des analyses en ligne sur un flux de données obtenues en temps réel. De telles combinaisons se retrouvent ainsi, par exemple, dans le suivi personnalisé de personnes en grande difficulté, de personnes très âgées, d'élèves en échec scolaire, etc. Il faut être capable d'analyser des évolutions sur de longues périodes, de prévenir si possible les problèmes, mais aussi savoir détecter l'urgence et réagir en temps réel à des crises.

Prenons l'exemple de la santé. Les données en rapport avec la santé d'un individu croissent sans cesse. Le cœur en est constitué par des informations telles que les examens médicaux, les diagnostics, les soins et les prises de médicaments. Mais on voit aussi se généraliser les données génomiques – la société californienne de biotechnologie *23andMe* propose ainsi aux individus le séquençage d'une partie importante de leur génome pour 99 dollars (moins de 80 euros). Il faut aussi prendre en compte les données sur la vie quotidienne de l'individu, son alimentation, son activité physique, son exposition à des pollutions particulières, etc. De plus en plus, de telles informations deviennent disponibles *via* des équipements de type téléphone intelligent ou *via* les réseaux sociaux.

Parallèlement, les chercheurs dans le domaine médical font un emploi croissant de l'analyse de données. Par exemple, ils espèrent découvrir des corrélations entre la prise de certaines combinaisons de médicaments et des pathologies particulières. Les patients, surtout, pourraient en profiter. Il s'agit d'abord de personnaliser les soins en adaptant les médicaments à chacun, en

contrôlant les quantités administrées. On peut imaginer agir de manière préventive en proposant à chacun une hygiène de vie adaptée à ses risques, en accompagnant chaque personne dans sa nutrition, ses activités physiques, sa vie quotidienne. Mais on peut aussi espérer détecter les crises et aider à les gérer, par exemple en régulant la prise de médicaments.

Un problème : l'accès aux données personnelles

Le fait de disposer de toutes les données relatives à une personne ouvre ainsi considérablement le champ des possibles. Mais cela soulève le problème de l'accès aux données personnelles. Un assureur ou un employeur doit-il avoir accès à tout ou partie de l'information constituée par les données médicales et génomiques d'un client, ses données fiscales, ses achats, sa géolocalisation, ses courriels, ses échanges dans les réseaux sociaux ?

Ce sont les données personnelles du client. À ce titre, elles lui appartiennent, et il devrait être seul à pouvoir décider qui y a accès et comment elles sont utilisées. Mais ce n'est pas aussi simple. Par exemple, pour faire progresser la médecine, on devrait pouvoir faire des analyses sur les données médicales de tous. Il semble raisonnable que les résultats de ces statistiques soient publics. Mais il semble tout aussi clair que les données brutes, par exemple celles d'un hôpital, ne puissent pas elles-mêmes faire partie des données ouvertes, même anonymisées, puisqu'il serait impossible de garantir leur confidentialité. Il reste beaucoup à faire pour concilier ces différents aspects. ■