Les données massives: baguette magique et matière première du XXIe siècle

Secteur économique en rapide expansion, la branche de l'information et de la communication ouvre de vertigineuses perspectives en termes de valeur ajoutée. L'exploitation électronique de données massives («big data») jusqu'à présent dissimulées comprend un véritable potentiel économique qui sera de plus en plus compris comme une nouvelle révolution technologique. La Suisse pourrait se profiler, à cette occasion, comme une source majeure de données, devenir un pionnier des données ouvertes («open data») et un haut lieu des technologies de l'information en Europe.

Les données massives ouvrent des possibilités entièrement nouvelles d'optimisation des processus, de corrélations et d'aide décisionnelle. Elles s'accompagnent aussi de nouveaux défis. Photo: Keystone

Le portail de messagerie sociale *WhatsApp* avec ses 450 millions d'utilisateurs a été récemment acheté par Facebook pour 19 milliards d'USD, soit près d'un demi-milliard par collaborateur. Les données massives sont en train de changer le monde. Ce terme, forgé il y a plus de quinze ans, s'applique à des volumes d'informations tellement importants qu'ils ne peuvent plus être traités par des procédés informatiques classiques. Les données massives sont très souvent désignées comme le «pétrole du XXIe siècle». Pour en tirer profit, nous devons apprendre à les «extraire» et à les «raffiner», c'est-à-dire à les transformer en informations et connaissances utiles. En un an, le monde produit aujourd'hui autant de données que l'humanité dans toute son histoire et ce volume double tous les quatorze mois et demi environ.

Ces masses de données sont le fruit de quatre innovations technologiques:

- *Internet,* notre outil de communication planétaire;
- le *World Wide Web,* réseau des sites Internet accessibles de partout grâce à la découverte au Cern du *protocole* de *transfert hypertexte (http);*
- les *réseaux sociaux* tels Facebook, Google+, WhatsApp ou Twitter, qui ont tissé des toiles de communication entre particuliers;
- l'*Internet des objets,* qui met aussi en contact électronique des réseaux d'objets et de mesures fournies par des capteurs. D'ici peu, il y aura davantage de machines exploitantes que d'usagers humains sur Internet.

Des stocks de données qui représentent plusieurs fois les plus grandes bibliothèques

Les volumes de données que brassent des entreprises comme Ebay, Walmart ou Facebook sont de l'ordre du pétaoctet (un trillion d'octets), soit cent fois le contenu de la bibliothèque du Congrès américain, la plus grande du monde. Ces données massives ouvrent des possibilités entièrement nouvelles d'optimisation des processus, de corrélations et d'aide décisionnelle. Elles s'accompagnent toutefois de nouveaux défis, résumés par quatre mots clés:

- les *volumes:* les quantités de données à traiter sont énormes;

Pr Dirk Helbing
Professeur de sociologie (modélisation et simulation), EPF Zurich

– la *vitesse:* souvent, un traitement en temps réel est indispensable;
– la *variété:* les données sont généralement très diverses et non structurées;
– la *(non-)fiabilité:* les données peuvent être incomplètes, non représentatives, voire erronées ou manipulées.

Il est donc nécessaire de développer des algorithmes – donc des méthodes de calcul – tout à fait nouveaux. Puisqu'il n'est pas efficient de stocker toutes les données pertinentes dans une seule mémoire partagée, leur traitement doit être assuré de manière décentralisée, le cas échéant sur des milliers d'ordinateurs. Cela passe par des procédures de calcul parallèles, tels que *MapReduce* ou *Hadoop*. Les algorithmes liés à ces données massives identifient en outre des corrélations intéressantes pouvant se prêter ici ou là à une exploitation commerciale: entre les conditions atmosphériques et le comportement des acheteurs, par exemple, ou entre les conditions de vie et les risques sanitaires ou de crédit. De même, la lutte contre la criminalité et le terrorisme s'appuie aujourd'hui sur l'analyse de grandes quantités de données comportementales.

Quelques exemples d'applications

Les applications en matière de données massives se répandent à toute vitesse. Elles débouchent sur des offres, des services et des produits personnalisés. L'un de leurs plus grands succès est la compréhension et le traitement automatiques de la parole. L'application Siri d'Apple comprend ce que dit l'usager qui cherche un restaurant proche, et Google Maps l'y conduit. Google Translate traduit à partir de comparaisons effectuées dans une gigantesque collection de textes traduits. Le système IBM Watson comprend une instruction en langage naturel et ne bat pas seulement des joueurs expérimentés (dans un jeu télévisé), mais conseille déjà des clients sur des permanences téléphoniques, souvent mieux que des préposés humains. IBM a décidé d'investir un milliard d'USD dans le développement et la commercialisation du système.

Les données massives jouent aussi un rôle important dans le secteur financier. Quelque 70% des transactions sur ce marché sont commandées par des algorithmes automatiques, ce qui correspond à peu près chaque jour à la masse totale d'argent existant dans le monde. De telles sommes attirent aussi le crime organisé. Les transactions financières sont donc soumises à des analyses algorithmiques destinées à détecter les procédés douteux. À l'aide d'un logiciel analogue nommé Aladdin, l'entreprise Blackrock spécule sur 15 000 milliards d'USD de fonds de clientèle, soit plus de 30 fois le produit intérieur brut de la Suisse.

Des potentiels considérables…

McKinsey évalue entre 3000 et 5000 milliards d'USD par an le potentiel économique mondial supplémentaire lié aux seules données ouvertes à tous[1]. Un potentiel présent dans quasiment tous les secteurs de la société. Le compteur intelligent («smart meter»), par exemple, permet de mieux ajuster l'une à l'autre la production et la consommation d'énergie; cela contribue à éviter les pics d'énergie, à gérer plus efficacement les ressources et à ménager l'environnement. Les risques peuvent être mieux identifiés et évités, les conséquences indésirables de certaines décisions atténuées et de nouvelles possibilités mises à profit, alors qu'on les laissait précédemment échapper. Les données ouvertes permettent, en outre, à la médecine de s'ajuster plus finement aux patients et à la prévention de prendre le pas sur le traitement des maladies.

… mais des risques bien réels aussi

Comme toute technologie, les données massives ne sont pas sans danger. La sécurité des communications numériques est fragile. La cybercriminalité – détournements financiers, vols de données et d'identités, etc. – prend toujours plus d'importance. Les infrastructures sensibles – comme l'approvisionnement énergétique ou les systèmes financiers et de communication – ne sont pas à l'abri de cyberattaques qui peuvent les mettre temporairement hors service.

De plus, les algorithmes généralement liés aux données massives peuvent certes être optimisés, mais la plupart des corrélations trouvées ne sont pas fiables et ne constituent pas des rapports de causalité. Une utilisation «naïve» des algorithmes peut donc aboutir à des conclusions inexactes. Des erreurs de classification (par exemple pour distinguer les bons des mauvais risques) ne sont pas rares. On peut aussi être amené à choisir sans le vouloir une procédure inadéquate. Il faut donc prendre garde à certains problèmes, tels que les mauvaises décisions, les discriminations et le favoritisme. Il convient de développer des procédures efficaces de contrôle qualité. À cet égard, les universités ont un rôle important à jouer. De même, des mécanismes efficaces doivent être trouvés pour protéger la sphère privée et le droit à l'autodétermination informationnelle, par exemple sous la forme d'un portefeuille de données personnelles (voir *encadré 2).*

Encadré 1

Les investissements massifs des géants du Web

Pour avoir un aperçu des tendances dans les technologies de l'information, considérons Google et ses (plus de) cinquante plateformes logicielles. L'entreprise dépense près de 6 milliards d'USD par an pour la recherche et le développement. Sur une seule année, Google a présenté des voitures sans chauffeur, investi massivement dans la robotique et lancé un projet visant à doter Internet de l'intelligence artificielle. Il a aussi investi dans l'Internet des objets en rachetant Nest Labs pour 3,2 milliards.

1 McKinsey Global Institute, *Open Data: Unlocking Innovation and Performance with Liquid Information,* octobre, 2013. L'analyse quantifie la valeur potentielle de l'application dans sept domaines: formation, transports, produits de consommation, électricité, pétrole et gaz naturel, santé, produits financiers.
2 Consommateurs qui jouent un rôle actif et croissant dans le processus de production, par exemple en créant leurs propres produits pour les vendre sur Internet ou en les fabriquant chez eux sur une imprimante 3D.

Réagir très vite face à la révolution technologique

Les technologies de l'information et de la communication révolutionnent la plupart de nos institutions: système éducatif (apprentissage personnalisé), science (étude des données), mobilité (voitures automotrices), transport de marchandises (drones), consommation (Amazon et Ebay), production (imprimantes 3D), système de santé (médecine personnalisée), politique (transparence accrue) et économie globale (prosommateurs[2]). Les banques doivent céder toujours plus de terrain au négoce algorithmique, aux bitcoins, à Paypal et à Google Wallet. Les assurances voient une grande partie de leurs affaires se dérouler sous forme de produits financiers tels que les swaps sur défaillance. Selon toute vraisemblance, ce processus de transformation économique et sociale vers une société numérique s'opérera sur une période de vingt ans, voire moins, ce qui est très court si l'on pense que la planification et la construction d'une route, par exemple, prend souvent trente ans ou plus. es

Encadré 2

Les infrastructures d'une ère numérique

Quelle société et quelle économie sortiront de la révolution numérique? Comment en faire bénéficier le plus grand nombre et maîtriser les risques? Pour mieux saisir le processus, souvenons-nous des nombreux facteurs qui ont assuré le succès de l'ère automobile: invention du moteur, de la voiture et de la production de masse; construction de voies publiques, de stations-service et de parkings; création d'auto-écoles et du permis de conduire; enfin, règles de circulation, panneaux de signalisation, contrôles de vitesse et police de la route.

Quelles sont les infrastructures techniques et les institutions juridiques, économiques et sociales requises pour que la société numérique soit, elle aussi, une pleine réussite? Globalement, l'ère numérique a besoin de systèmes informatiques fiables, transparents, ouverts et participatifs. Dans le cadre du Forum économique mondial (WEF), des représentants économiques, politiques et scientifiques ont à cette fin élaboré un document consensuel pour un «New Deal on Data»[a]. Ce texte énonce trois grands principes:

- pour les technologies liées aux données massives dignes de confiance et acceptées par la société, il faut un meilleur équilibre entre les intérêts de l'économie, de l'État et du citoyen ou consommateur;
- les individus doivent pouvoir reprendre le contrôle de leurs données personnelles, comme l'exige le droit à l'autodétermination informationnelle;
- les individus doivent toucher une participation équitable aux bénéfices réalisés avec leurs données personnelles.

Comment développer des technologies de l'information compatibles avec les valeurs de notre société? Nous pourrions, par exemple, faire participer les citoyens au développement du réseau pour l'Internet des objets en devenir. Celui-ci permettrait de mesurer en temps réel la situation de notre monde (système nerveux numérique) et de mettre à disposition un moteur de recherche européen. Pour la protection de la sphère privée, toutes les données collectées y afférentes seraient classées dans un portefeuille personnel, où l'individu choisit les applications pour lesquelles ses données peuvent être utilisées. À travers un système de paiement des microtransactions, les fournisseurs de données, titulaires de droits et innovateurs recevraient une rémunération équitable pour leurs services. Un système pluraliste et contrôlé par les utilisateurs, basé sur la réputation, encouragerait un comportement responsable dans le monde tant réel que virtuel.

Le recours à une plateforme participative permettrait à tout un chacun de «téléverser» des données, algorithmes mis en œuvre par ordinateur et notations ou d'utiliser (gratuitement ou contre paiement) les contributions de tiers. Il en résulterait un écosystème d'information et d'innovation à forte croissance, qui exploiterait le potentiel des données pour l'économie, la politique, la science et les citoyens (catalyseur de l'innovation). Des filtres d'information à configuration individuelle et des réseaux sociaux spécifiques assisteraient en outre l'intelligence collective au même titre que la création de capital social (par exemple confiance), ce qui aurait aussi toute son importance pour les fonctionnalités des marchés financiers. Enfin, une plateforme de l'emploi et de projets créerait les conditions d'un marché de l'emploi 2.0 flexible.

a Voir *Personal Data: The Emergence of a New Asset Class,* WEF.

Il est donc urgent d'agir sur les plans technologique, législatif et socioéconomique. Les États-Unis ont lancé il y a plusieurs années une initiative concernant la recherche dans les données massives, dotée d'un budget de 200 millions d'USD. Celle-ci s'est accompagnée de différents programmes de large envergure. Le projet *FuturICT (www.futurict.eu)* de l'UE doit développer des concepts permettant à l'Europe de relever les défis de la société numérique. D'autres pays l'appliquent déjà. Le Japon a ainsi lancé récemment, au *Tokyo Institute of Technology,* un vaste projet décennal évalué à 100 millions d'USD. Il existe de nombreux autres projets encore, avec des enveloppes financières souvent bien plus importantes, notamment dans le domaine militaire et de la sécurité.

La Suisse, un moteur d'innovation européen dans le domaine numérique

En Suisse, les conditions pour bénéficier de l'ère numérique sont bonnes, mais il ne suffit pas de recopier les technologies existantes. Notre pays doit procéder à de nouvelles découvertes pour influencer l'ère numérique. N'oublions pas que le World Wide Web a vu le jour à Genève, au Cern. Ce dernier reste d'ailleurs la première compétence du monde civil en matière de données massives. Les États-Unis et l'Asie sont encore «leaders» dans leur exploitation commerciale, mais le scandale de la NSA, le développement des capteurs de mesure à communication sans fil et l'Internet des objets vont changer la donne.

En ciblant le soutien fourni aux hautes écoles dans le domaine des sciences informatiques, la Suisse pourrait jouer un rôle de précurseur dans la R&D en Europe. Elle assume déjà la direction académique de trois des six initiatives phares lancées par l'UE. Pourtant, les seuls grands thèmes qui bénéficient actuellement d'un soutien sont la modélisation numérique du cerveau humain et la robotique. Les EPF prévoient d'investir davantage dès 2017 dans la *science des données,* un domaine d'étude en pleine éclosion, qui s'occupe d'analyser scientifiquement les données. Or, le monde informatique évolue rapidement, son potentiel économique est important de même que le pouvoir de transformation de sa technologie. Il est donc très urgent et dans l'intérêt national d'accélérer la recherche, d'en élargir la portée et de lui donner une substance.

Avec ses valeurs démocratiques, son cadre juridique et ses entreprises informatiques, la Suisse est bien placée pour devenir le moteur d'innovation de l'Europe en vue de l'ère numérique.