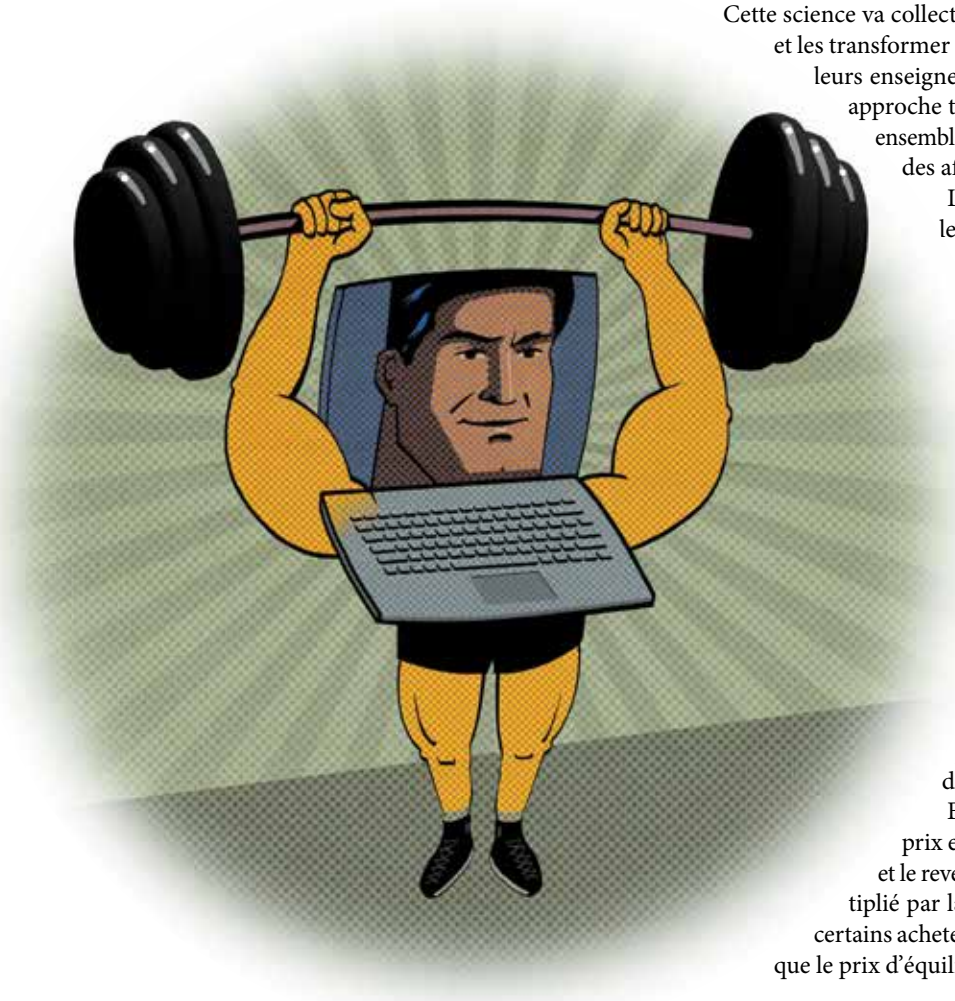
Mégadonnées, **mégamuscles**

La puissance informatique stimule l'apprentissage automatique et révolutionne les affaires et la finance

Sanjiv Ranjan Das

LA MASSE de données disponible aujourd'hui aurait été inimaginable il y a encore dix ans. Les données s'accumulent trop vite pour être classées ou analysées; il reste donc aux entreprises à trouver une solution pour exploiter cette manne afin de mieux informer leurs décisions et rehausser leurs performances.

La «science des données» est une nouvelle discipline qui permet de produire des connaissances exploitables à partir de grosses quantités de données, les mégadonnées, en les analysant pour dégager des structures, des tendances et des associations. Cette science va collecter, organiser, analyser les données et les transformer en connaissance, avant de traduire leurs enseignements en actions concrètes. Cette approche touche l'activité humaine dans son ensemble : l'économie, la finance et le monde des affaires n'y échapperont pas.



La science des données applique les outils de l'apprentissage automatique — une intelligence artificielle qui permet aux ordinateurs d'apprendre sans programmation explicite (Samuel, 1959) — aux quantités gigantesques de données que nous produisons. Une voie riche en promesses, qui pourrait révolutionner la gestion des entreprises et l'analyse des politiques économiques.

Profilage des consommateurs

Sans surprise, compte tenu de ses arguments économiques de choc, la science des données remporte un succès fou dans le monde des affaires.

En situation de libre concurrence, le prix est le même pour tous les acheteurs et le revenu du vendeur est égal au prix multiplié par la quantité vendue. Or, on sait que certains acheteurs seraient prêts à payer plus cher que le prix d'équilibre. Ils détiennent un «surplus du

consommateur» que les mégadonnées permettent d'exploiter grâce à un profilage des acheteurs.

En faisant payer à chaque consommateur un prix défini en fonction de son profil, le vendeur obtient le montant le plus élevé que l'acheteur est désireux et capable de payer pour un même produit. Cette segmentation tarifaire peut rapporter gros. Déjà établie chez les compagnies aériennes, la pratique s'étend désormais à d'autres secteurs.

En outre, les vendeurs peuvent aussi pratiquer des prix plus bas pour des clients qui ne pourraient pas payer le prix d'équilibre, ce qui va simultanément accroître leur revenu, étendre leur clientèle et éventuellement améliorer le bien-être social. L'utilisation des mégadonnées en profilage explique en grande partie les valorisations boursières stratosphériques d'acteurs comme Facebook, Google et Acxiom, qui proposent des produits et des services sur la base des données de leurs clients.

Si les mégadonnées peuvent être utilisées pour exploiter le consommateur, elles modifient toutefois les pratiques commerciales en sa faveur. À l'aide des données générées par les interactions sur les médias sociaux, certaines entreprises parviennent à mieux comprendre leurs habitudes de crédit. En rapprochant les historiques d'emprunteur et l'activité sur les médias sociaux, on obtient un meilleur système de notation des débiteurs. Ainsi, certains consommateurs, qui, autrement, n'auraient pas eu accès au crédit, peuvent obtenir un prêt.

En particulier, les mégadonnées éliminent les préjugés qui apparaissent lorsque des décisions sont prises sur la base d'informations limitées. Le manque de données fines avait donné naissance dans les années 30 aux listes noires, pratique selon laquelle des banques délimitaient les quartiers auxquels ils n'accorderaient pas de prêt en se basant sur la race ou l'ethnie de la population. C'est ainsi que des pans entiers de la société n'ont pas eu accès au crédit.

Avec les mégadonnées, le stéréotype ne fait plus la loi. Des données fines et individualisées remplacent les conceptions subjectives simplistes. Au lieu de se limiter aux éléments démographiques habituels comme le revenu, l'âge et la localisation, les agences de notation peuvent exploiter l'hétérogénéité que révèlent les médias sociaux, les SMS, les microblogs, les profils d'utilisation de cartes de crédit et les données de profilage (Wei *et al.,* 2014). Les données granulaires facilitent la classification des individus par qualité d'emprunteur.

Prévisions et analyse du risque

La science des données a transformé l'art de la prévision économique. Certaines statistiques économiques clé comme le PIB trimestriel ne sont d'habitude disponibles qu'avec un gros retard. La science des données contourne ces retards : elle utilise des chiffres relevés plus fréquemment (chômage, commandes industrielles ou même climat général) pour prédire les variables plus espacées dans le temps.

Cette démarche, surnommée *nowcasting* ou prédiction du présent, peut être assimilée à de la prévision en temps réel (voir «La reine des chiffres», *F&D* mars 2014).

Autre application de la science des données, l'analyse du risque financier systémique. Dans un monde de plus en plus interconnecté, mesurer ces connexions est riche en promesses pour la prise de décisions économiques.

Penser le risque systémique en termes de réseaux est riche d'enseignements. Les scientifiques de données utilisent des masses de données pour construire des représentations de l'écheveau des relations entre les banques, les assurances, les différents intermédiaires, *etc.* Il est clairement utile de savoir quelles banques sont plus connectées que d'autres et lesquelles ont le plus d'influence en utilisant une méthode fondée sur les *eigenvalues*. Une fois les réseaux construits, on peut mesurer le degré de risque d'un système financier, ainsi que la contribution d'un établissement donné au risque d'ensemble. Pour le régulateur, c'est une nouvelle méthode pour analyser et gérer le risque systémique. Voir Espinosa-Vega et Solé (2010); FMI (2010); Burdick *et al.* (2011); Das (2016).

Ces méthodes au carrefour de plusieurs disciplines universitaires doivent beaucoup aux mathématiques des réseaux issus de la sociologie et sont appliquées à des réseaux très étendus grâce à des modèles informatiques sophistiqués.

Pas que des mots

L'analyse textuelle est une branche très porteuse de la science des données qui vient compléter utilement les données quantitatives en finance et en économie (voir «Les deux visages du changement», dans ce numéro). Une foule d'applications exploitent la fouille de textes : iSentium évalue le climat économique à court et à long terme grâce aux tweets; StockTwits produit ses indicateurs de «l'air du temps» avec une application sur mobile.

On peut maintenant attribuer un classement à une société en fonction de son bénéfice trimestriel à partir du formulaire 10K,

Les mégadonnées éliminent les préjugés qui apparaissent lorsque des décisions sont prises sur la base d'informations limitées.

rapport annuel que chaque société cotée remet à la Securities and Exchange Commission (SEC). Un décompte des mots liés au risque dans les rapports trimestriels permet de prévoir précisément le classement des entreprises par bénéfices. Les entreprises dont les rapports sont difficiles à lire tendent à être moins bénéficiaires — peut-être l'obscurité des propos vise-t-elle à masquer des réalités peu flatteuses (voir Loughran et McDonald, 2014)? Il est très facile de noter les rapports financiers selon une mesure de lisibilité très connue, l'indice Gunning Fog. Un organisme de régulation comme le Consumer Financial Protection Bureau songe à établir des normes de lisibilité.

D'après certaines études, la longueur du rapport trimestriel permettrait à elle seule de détecter de mauvaises nouvelles (un rapport long augure de résultats en baisse) car il est considéré que les propos obscurs sont verbeux; si l'on suit ce raisonnement jusqu'au bout, la taille du fichier que les entreprises envoient sur le site de la SEC pourrait nous renseigner sur son bénéfice trimestriel. Ce champ d'étude en plein essor recèle bien des promesses.

Dans le même registre, «l'analyse des nouvelles» puise des données dans les actualités. On y trouve de plus en plus d'acteurs comme RavenPack, qui fournit des scores de climat économique, utilise l'analyse prédictive des transactions et produit des prévisions macroéconomiques. D'énormes masses de données non structurées issues de la presse et des médias sociaux sont converties en données et indicateurs granulaires, au service d'acteurs de la finance — gestionnaires d'actifs, teneurs de marché, gestionnaires du risque et spécialistes de la conformité.

D'après certaines études, la longueur du rapport trimestriel permettrait à elle seule de détecter de mauvaises nouvelles.

L'analyse des nouvelles en flux est particulièrement intéressante. Les fonds spéculatifs traitent chaque jour plusieurs milliers de fils d'actualité pour en extraire les principaux sujets et suivent l'évolution proportionnelle de leur nombre d'occurrences pour détecter les tendances des marchés. Une telle analyse serait utile à des acteurs publics et à des régulateurs comme les banques centrales. Ainsi, lorsqu'un changement brutal est détecté dans l'équilibre des thèmes traités dans la presse (inflation, taux d'intérêt, croissance) peut-être est-il temps d'infléchir la politique des taux d'intérêt.

Pour analyser les thèmes, on commence par construire un gigantesque tableau de fréquence des mots, «la matrice des termes» qui mouline des milliers d'articles. Chaque mot occupe une ligne, et chaque article une colonne de la matrice. Cela permet de faire apparaître des thèmes, par l'analyse mathématique des corrélations entre les mots et entre les documents. Les grappes de mots sont indexées et les thèmes sont détectés par apprentissage automatique, par exemple par analyse sémantique latente ou par allocation de Dirichlet latente (LDA). L'analyse LDA permet d'obtenir une série de thèmes et des listes de mots afférents à ces thèmes.

Sans entrer dans trop de détails, disons qu'il s'agit en fait de techniques statistiques pour repérer les principales associations de mots présentes dans une collection de documents (par exemple un fil d'actualité). Ces signaux linguistiques seront très utiles aux responsables économiques et dans la prise de décision — par exemple pour redéfinir les messages des campagnes électorales.

L'intelligence artificielle et l'avenir

Avec leur puissance de calcul toujours plus grande, les ordinateurs peuvent traiter d'énormes quantités de données, ce qui a permis les avancées de l'intelligence artificielle. Une nouvelle classe d'algorithmes, dits «d'apprentissage profond», inspirés des réseaux neuronaux biologiques, émule le fonctionnement du cerveau avec une puissance considérable, permettant de nombreux exemples réussis d'intelligence artificielle.

L'apprentissage profond est une méthode statistique qui utilise des réseaux neuronaux artificiels pour mettre en correspondance une multitude de données d'entrée et de sortie — c'est-à-dire qui identifie des schémas. L'information est disséquée par un réseau de neurones mi-physiques mi-logiciels. Les connexions entre ces neurones se voient renforcées par l'arrivée de nouvelles données, exactement comme le cerveau humain qui consolide son expérience. Les progrès de l'apprentissage profond s'expliquent par l'immense quantité de données disponibles et par la croissance exponentielle de la puissance de calcul, grâce au développement de puces spécifiques.

L'apprentissage profond est le moteur qui fait tourner beaucoup de technologies que nous tenons pour acquises, comme la traduction automatique, les voitures autonomes, la reconnaissance et l'indexation d'images. Cette technologie est peut-être en passe de révolutionner l'économie et la gouvernance. Les agences de notation s'en servent déjà pour produire des rapports sans intervention humaine. De grands réseaux neuronaux d'apprentissage profond vont bientôt nous donner des prévisions et repérer des liens entre variables économiques mieux que les méthodes statistiques classiques.

Ne nous hasardons pas à prédire quels domaines de cette «science lugubre» qu'est l'économie profiteront le plus de l'apprentissage automatique. Une chose est sûre, selon les mots de l'écrivain de science-fiction William Gibson, «L'avenir est là; il n'est pas très bien réparti, voilà tout». ■

Sanjiv Ranjan Das est Professeur à la Leavey School of Business à l'Université de Santa Clara.

Bibliographie :

Billio, Monica, Mila Getmansky, Andrew W. Lo, and Loriana Pelizzon, 2012, "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors," Journal of Financial Economics, *Vol. 104, No. 3, pp. 535–59.*

Burdick, Douglas, Mauricio A. Hernandez, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana Stanoi, Shivakumar Vaithyanathan, and Sanjiv Das, 2011, "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study," IEEE Data Engineering Bulletin, *Vol. 34, No. 3, pp. 60–7.*

Das, Sanjiv, 2016, "Matrix Metrics: Network-Based Systemic Risk Scoring," Journal of Alternative Investments, *Vol. 18, No. 4, pp. 33–51.*

Espinosa-Vega, Marco A., and Juan Solé, 2010, "Cross-Border Financial Surveillance: A Network Perspective," IMF Working Paper 10/105 (Washington: International Monetary Fund).

International Monetary Fund (IMF), 2010, "Systemic Risk and the Redesign of Financial Regulation," Global Financial Stability Report, *Chapter 2 (Washington, April).*

Lin, Mingfen, Nagpurnanand Prabhala, and Siva Viswanathan, 2013, "Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending," Management Science, *Vol. 59, No. 1, pp. 17–35.*

Loughran, Tim, and Bill McDonald, 2014, "Measuring Readability in Financial Disclosures," Journal of Finance, *Vol. 69, No. 4, pp. 1643–71.*

Samuel, A.L., 1959, "Some Studies in Machine Learning Using the Game of Checkers," IBM Journal of Research and Development, *Vol. 3, No. 3, pp. 210–29.*

Wei, Yanhao, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas, 2015, "Credit Scoring with Social Data," Marketing Science, *Vol. 352, pp. 234–58.*