Les 5V du *big data*

Thomas Bourany
RCE

Bien qu'omniprésentes dans notre actualité, les données massives et l'intelligence artificielle constituent des phénomènes nouveaux et parfois difficiles à définir. Le terme « *big data* » a été popularisé par John Mashey, informaticien chez Silicon Graphics dans les années 1990. Il fait référence aux bases de données trop grandes et complexes pour être étudiées avec les méthodes statistiques traditionnelles – et, par extension, à tous les nouveaux outils d'analyse de ces données. En 2001, Douglas Laney a analysé cette nouvelle tendance à travers une liste très simple de trois « V », ensuite élargie à cinq « V » :

– Volume : la grande quantité d'information contenue dans ces bases de données.

– Vélocité : la vitesse de leur création, collecte, transmission et analyse.

– Variété : les différences de natures, formats et structures.

– Valeur : la capacité de ces données à générer du profit.

– Véracité : leur validité, *i.e.* qualité et précision ainsi que leur fiabilité.

Les cinq « V », un changement de nature des données à l'ère digitale

L'accroissement du **volume** de données est peut-être l'évolution la plus spectaculaire. La masse des

données générée chaque seconde dans le monde ne se mesure plus en gigabits ni en térabits – unités de mesure usuelles des disques durs d'ordinateur – mais en zettabits ou en brontobits – qui sont respectivement un milliard et mille milliards de fois plus grand que le térabit. De nos jours, chaque minute voit la création d'un volume de données aussi important que l'ensemble des données du monde – livres, œuvres d'art, musique – depuis l'invention de l'écriture jusqu'en 2008. La croissance exponentielle de la taille de ces données implique d'innover dans de nouvelles méthodes de stockage (le *cloud* et les « systèmes distribués ») et d'analyse statistique (via l'apprentissage supervisé ou non-supervisé), qui donnent une dimension nouvelle à ce domaine.

La **vitesse** de création, de collecte et de transmission de ces données n'échappe pas non plus à cette tendance exponentielle. En marketing, le choix des publicités sur internet est par exemple décidé via des enchères – appelée *Real Time Biding* – réalisées pendant le chargement de la page, en moins de 10 millisecondes. Pendant ce court laps de temps, les données d'utilisateurs sont transférées et analysées par les algorithmes pour un nombre moyen de 10 millions de prédictions publicitaires par seconde. Un deuxième exemple est celui des transactions financières à haute fréquence – le *High Frequency Trading* – où l'échelle commune est devenue celle de la nanoseconde (soit un milliardième de seconde). C'est à cette fréquence extrêmement rapide que la plupart des actifs sont achetés et vendus sur les marchés financiers du monde entier. Cette vélocité croissante implique d'analyser ces données en temps réel – parfois en même temps qu'elles sont générées – pour éviter leur stockage.

Au vu de la **variété** des domaines d'applications, les données prennent différentes formes. Elles peuvent être structurées : tableaux de nombres ou données qualitatives

avec différentes variables. Cependant, aujourd'hui, 80 % des données sont non-structurées (textes bruts, images, vidéos, ou encore données de capteurs et d'ADN par exemple) présentant de nouveaux défis. En effet, si les méthodes statistiques « traditionnelles » sont efficaces pour analyser de « simples » données numériques, elles se révèlent incapables de traiter les images pixélisées ou les séquences textuelles, même converties au format numérique. Les nouvelles méthodes d'apprentissage statistique, elles, permettent ce type d'analyse.

La **valeur** potentielle de ces données et de leur analyse incite les entreprises à développer et adopter les procédés du *big data*. Alors que l'ensemble des projets liés au *big data* ne présentait qu'un revenu de 100 millions d'euros en 2009, ces revenus – et profits – ont considérablement augmenté depuis 2012, générant près de 42 milliards d'euros de revenu mondial de marché en 2018, et certaines prédictions – notamment émises par le portail Statista – évaluent à plus de 100 milliards ces revenus à l'horizon 2030.

Enfin, la question de la **véracité** de ces données a pris une dimension particulière depuis les fréquentes polémiques autour des « infox » (les *fake news*), notamment dans le cas des réseaux sociaux. De manière plus large, la validité, la qualité et la confiance que l'on accorde aux données influencent leur utilisation. Ces questions se posent en particulier quand il y a un manque – paradoxal ! – d'information sur les sources et la méthodologie de collecte de ces données.

Le big data *comme changement de paradigme technologique et scientifique*

L'augmentation croissante du volume de données créées ces dernières années a été en partie maîtrisée par la croissance de la puissance de calcul – prédite par la loi

de Moore – et des capacités de stockage. Cependant, pour les entreprises utilisant les données à gros volume, il s'est avéré nécessaire de ne plus stocker ou analyser les bases de données « sur machine » mais sur des serveurs hébergés à distance : c'est ce qu'on appelle le *cloud computing*. De plus, pour accélérer les calculs des algorithmes, la parallélisation des tâches et l'innovation des systèmes dits « distribués » constituent également des avancées informatiques majeures dans différents domaines scientifiques.

Mais quelles sont les méthodes du *big data* et quels sont leurs atouts ? La plupart des algorithmes d'apprentissage statistique permettent de résoudre des tâches de classification. Par exemple, un algorithme de triage va « classifier » certains mails comme « spams » ou « emails indésirables » en fonction des caractéristiques du message. Dans la recherche médicale, un algorithme servira à classifier si une tumeur est bénigne ou cancéreuse en fonction de variables comme la taille, la couleur, etc. Ces deux algorithmes ont en commun d'analyser les données pour reconnaître des motifs de ressemblance : ils « apprendront » au cours du temps et peuvent être très performants sur des types de données très variés et de grandes dimensions.

Pour conclure, le *big data* n'est pas seulement l'apanage de certains secteurs médiatisés comme les réseaux sociaux, mais concerne aussi de nombreux domaines de la recherche scientifique : les images des sondes spatiales ou des accélérateurs de particules, les données médicales et génomiques constituent d'immenses volumes.

Pour proposer des services à toujours plus grande échelle pour les scientifiques, l'industrie et les particuliers, les

entreprises de la haute technologie – que ce soit les GAFAM[1] ou la myriade de start-ups de ce secteur – poussent toujours plus loin la frontière de l'innovation. Enfin, le caractère transdisciplinaire de ce domaine – à l'intersection entre l'informatique, les statistiques et leurs applications – demande de repenser complètement le paradigme technologique, scientifique et économique de de nos sociétés.

Bibliographie

Colombus L. (2018), « 10 Charts That Will Change Your Perspective Of Big Data's Growth », *Forbes*.

Laney D. (2001), « 3D Data Management Controlling Data Volume Velocity and Variety », *META Group*.

Lohr S. (2013), « The Origins of 'Big Data': An Etymological Detective Story », *Bits Blog*.

Marr B. (2015), *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*, Wiley.

Marr B. (2018), « How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read », *Forbes*.

« IBM, Microsoft and Google Make Little Headway Against Amazon's IaaS/PaaS Dominance », *Synergy Research Group*.

« The Leading Cloud Providers Increase Their Market Share Again in the Third Quarter », Synergy Research Group.

« Qu'est-ce que le cloud computing ? Guide du débutant », *Microsoft Azure*.

1 Acronyme des géants du Web — Google, Apple, Facebook, Amazon et Microsoft.