Regards sur…

Le Big data

« Big data » : les données comme matière première ?

Un phénomène massif d'une rapidité extrême est intervenu ces dernières années : alors que 2 % des données étaient stockées sous format numérique en 1982, ce sont 98 % d'entre elles qui sont numérisées maintenant. Parallèlement, la croissance des données produites est d'ordre exponentiel. Il est ainsi estimé que les données numérisées produites par l'humanité en 2013 sont égales à toutes les données produites avant 2010.

Naturellement, si le terme « données » recouvre des choses de qualités et de valeurs très différentes, il n'empêche que l'existence de cet univers numérique en expansion très rapide change et va changer très profondément la manière de faire de la science, et l'ensemble des processus de décision.

Les « données » du problème

D'un certain côté, la révolution du « Big data » s'est imposée à nous plus qu'elle n'a été décidée en réponse à un problème. C'est pour cela que l'on s'interroge de toute part sur sa signification et sur les besoins qu'elle pourrait satisfaire.

Le « Big data » est la résultante de trois facteurs :

› **La numérisation de données de toutes sortes :** bases de données classiques, mais aussi textes, photos, vidéos, musiques… ce qui permet des traitements croisés sur tous les types de données. En 30 ans en effet, le coût de stockage des données numérisées a été divisé par 100 000, et les capacités de calcul ont doublé approximativement tous les 18 mois selon la fameuse loi de Moore qui date de 1965 ! De plus, les capteurs de tous ordres : téléphones mobiles, mais demain aussi tous nos appareils ménagers, nos maisons, nos voitures, nos montres, nos vêtements, produisent désormais des données sous format numérique, tout en se miniaturisant à l'extrême, tandis que leur coût diminue de façon prodigieuse.

› **Le développement des réseaux,** et d'Internet en particulier, là aussi à un degré totalement imprévu, permet l'échange des données et leur traitement distribué pour un coût extrêmement modéré. Les réseaux sociaux sont devenus une partie intégrante de nos vies privées et professionnelles. L'échange automatique de données entre « objets connectés » va encore accroître la numérisation du monde et la capacité de réactivité à toute nouvelle information ou donnée.

› **Le développement de nouveaux algorithmes** d'analyse de données allié à des capacités de calcul extraordinairement amplifiées et largement accessibles, par exemple grâce au « cloud ».

Incroyablement, alors que la production de données est devenue phénoménale et qu'elle se fait bien souvent en réaction de plus en plus rapide à d'autres données, une grande partie de cette « écume numérique du monde » est stockée, ce qui ouvre des possibilités complètement nouvelles d'analyse, et provoque un débat entre droit à l'oubli et droit à l'histoire.

Il faut bien comprendre que les technologies et approches classiques de gestion et de traitement de données ne sont plus à même de nous permettre de faire face aux « Big data » et ses nouvelles caractéristiques. Il est ainsi devenu classique de mettre en avant au moins quatre problèmes avec les défis qui les accompagnent :

› **Le volume.** Comme nous l'avons vu, ce volume explose. Le mégaoctet a longtemps été l'unité de mesure de la taille des mémoires des ordinateurs, puis le gigaoctet a témoigné de l'arrivée de la numérisation de l'image animée, le téraoctet (10^{12} octets) désigne la

puissance de stockage désormais accessible à chacun d'entre nous, permettant en théorie de conserver l'équivalent de fonds de grandes bibliothèques nationales. Le petaoctet (10^15 octets) correspond aux masses de données entreposées dans les « fermes de données », et l'exacotet (10^18 octets) est tutoyé dans certains domaines (physique des particules, astronomie). Le stade de fichiers Excel que chacun pouvait examiner sur son ordinateur personnel est complètement dépassé. On dit souvent que le « Big data » commence quand on ne peut plus stocker les données concernées dans la mémoire centrale de son ordinateur, et donc qu'il faut recourir à des traitements sophistiqués pour rendre les calculs réalisables, c'est-à-dire faisables en un temps raisonnable.

› **La vélocité.** Les données modernes sont maintenant produites en flux. Elles incluent les millions de tweets échangés chaque heure, les centaines d'heures de média déposées sur YouTube chaque minute, les données communiquées et produites par nos smartphones, les séquences de clics et de transactions enregistrées sur les sites web, etc. Même les images satellitaires de télé-détection vont maintenant être disponibles toutes les cinq heures pour chaque zone géographique au lieu d'une fois tous les 2 mois (cf. la mise en place du réseau de satellites Sentinelles par l'Europe, sans compter les micro-satellites que des start-up américaines envoient désormais par dizaines dans l'espace). Il faut donc être capable de traiter une grande partie de ces données « à la volée ». De plus la « fraicheur » des données devient un critère qu'il importe de prendre en compte.

› **La variété.** Les données ne sont plus issues de processus bien définis de recueil dans un format établi, mais elles sont désormais stockées au mieux dans des entrepôts de données, au pire dans des fichiers d'origines diverses, avec des formats variés, impliquant possiblement des données multi-média audio et vidéo, du texte brut ou dans des formats plus ou moins propriétaires, des transactions financières, des méta-données, etc. La question de la mise en relation de tous ces types de données très hétérogènes devient ainsi cruciale.

› **La véracité.** Les données étant issues de capteurs ou de sources humaines très diverses, leur degré de précision et surtout de fidélité ainsi que le niveau de confiance qu'on peut leur accorder est très varié. Il faut donc savoir combiner les sources et raisonner en tenant compte de ces indices de précision, des biais éventuellement connus ou identifiés et du niveau de confiance.

Cette disponibilité quasi infinie de données et les nouvelles possibilités de traitements massifs désormais accessibles à bas prix, grâce en particulier au « cloud computing », bouleversent l'approche scientifique du monde.

Avant, la démarche était de réfléchir à une question, par exemple l'existence ou non d'une corrélation entre deux variables (voire entre quelques variables peu nombreuses), d'établir avec soin un « plan d'expériences », de récolter l'échantillon de données aussi limité et aussi propre que possible pour satisfaire les contraintes de significativité statistique, et de mesurer la corrélation faisant l'objet de notre attention, avant de conclure ou non à son existence, par exemple en comparant à une p-value.

Désormais, la démarche est de demander aux machines de découvrir toutes les corrélations multi-variables existantes dans un énorme volume de données souvent bruitées, puis seulement ensuite, d'examiner ce qui peut présenter un intérêt dans cette masse de liens potentiels. De manière alternative, on peut demander aux machines de détecter ce qui émerge comme étant la norme et, à partir de là, d'identifier des « signaux faibles », c'est-à-dire des phénomènes étranges, hors norme, qu'il peut être intéressant d'examiner.

De même, avant, on était centré sur l'ajustement des modèles statistiques aux données (prédire le passé), tandis que l'on cherche à présent des capacités prédictives par la généralisation et l'extrapolation des régularités découvertes.

De plus, les corrélations ainsi découvertes peuvent à leur tour servir d'entrées pour d'autres mécanismes de « data mining », participant ainsi à un processus d'enrichissement (ou de pollution) cumulatif et potentiellement exponentiel.

« Cette disponibilité quasi infinie de données et les nouvelles possibilités de traitements massifs bouleversent l'approche scientifique du monde »

Une nouvelle ère scientifique s'ouvre

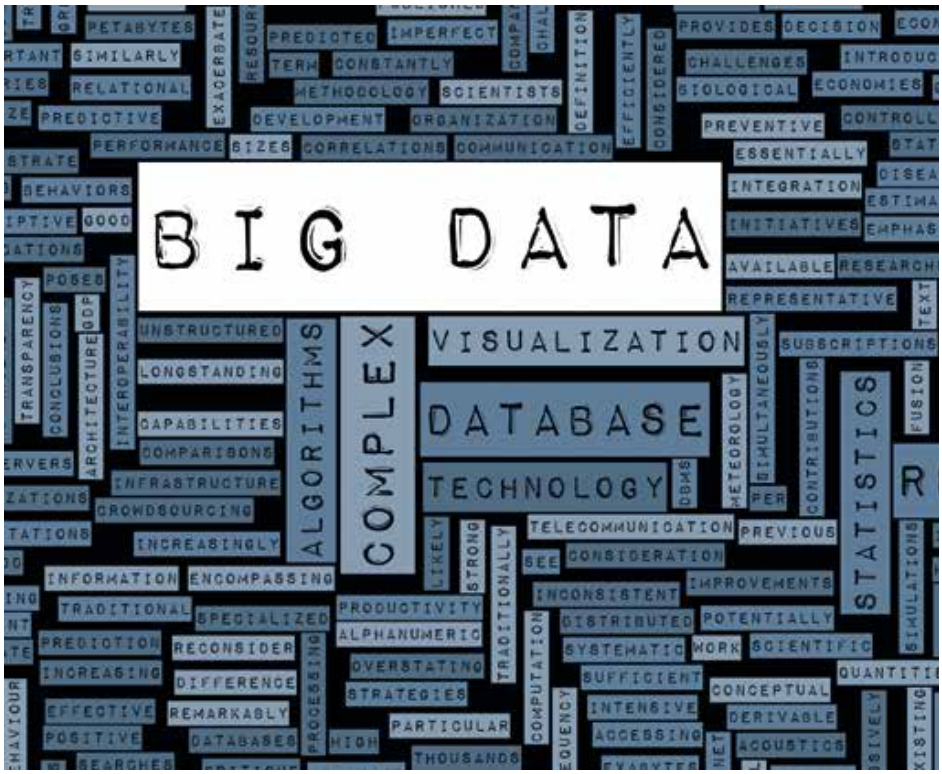
À côté de ces caractérisations techniques, un autre regard sur le « Big data » fait ressortir la nouvelle ère scientifique ouverte grâce à lui. Un découpage en quatre grandes ères scientifiques et en quatre approches est ainsi formulé :

› **Approche empirique.** Elle correspondrait à une première étape de la démarche scientifique, qui consiste à répertorier et à classer les objets, êtres vivants et phénomènes naturels.

› **Approche théorique.** Inaugurée magistralement par Galilée et Newton, elle est associée à la modélisation du monde et à sa mise en équations. Cependant, elle trouve des limites dans son application car toutes les équations, de loin s'en faut, n'ont pas de solutions analytiques.

› **Approche par la simulation.** Heureusement, l'informatique, apparue dans les années 1940, a offert le moyen de résoudre numériquement les équations et modèles mathématiques du monde, et d'en étendre ainsi le champ bien au-delà des systèmes assez simples et simplifiés de la physique du XIX^e^ siècle. Ainsi ces simulations numériques ont permis à la physique quantique, la physique des solides et la relativité générale de faire des prédictions vérifiables. Elles contribuent aussi de manière essentielle au développement des sciences du vivant et de l'environnement et, généralement, des sciences des systèmes complexes naturels ou artificiels.

Il est en tous les cas essentiel de bien réaliser que d'un questionnement orienté et raisonné, on passe avec le « Big data » à une exploration tous azimuts de corrélations ou de signaux faibles ou de tendances, pour ensuite les filtrer, les recouper, et alimenter l'univers numérique. Il n'y a plus en pratique de question de taille d'échantillon, et, de plus, les données ne servent plus seulement à répondre à une question pour laquelle elles ont été récoltées, mais elles sont ré-utilisables à l'infini en fonction de nouveaux traitements que n'importe quel « data scientist » qui en dispose peut imaginer.



› **Approche par exploration des données.** Finalement, nous serions entrés dans l'ère de nouvelles découvertes rendues possibles par l'exploitation des énormes masses de données acquises sur le monde grâce à toutes les nouvelles technologies du « Big data ».
Il est indéniable que des champs scientifiques tels que la sociologie ou les sciences de l'environnement sont en profonde mutation grâce au « Big data ». De même que dans des domaines plus « traditionnels », tel que celui de la physique des particules, les nouvelles découvertes (e.g. boson de Higgs) seraient impossibles sans cette nouvelle capacité à traiter des données hyper-massives.

Une matière première et de nouvelles opportunités

Personne sans doute n'est encore capable de prédire avec précision quelles seront les applications du « Big data » et les (r)évolutions à en attendre. Très généralement, les possibilités suivantes, qui sont neuves, font miroiter tout un ensemble de nouvelles opportunités :

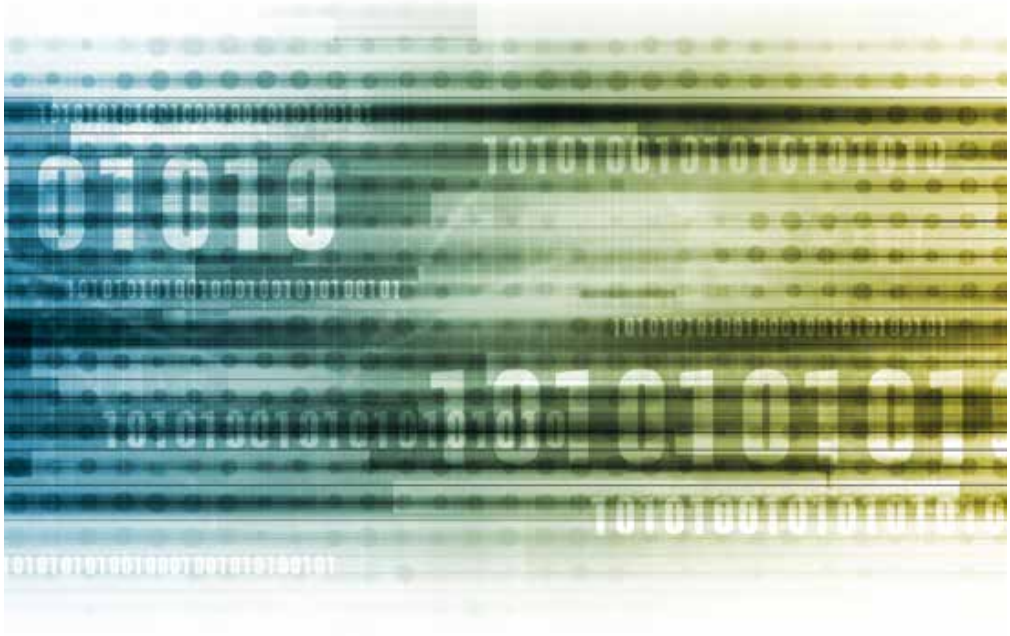
› **Nouvelles possibilités pour comprendre le monde.** La science s'appuie désormais autant sur l'analyse de données que sur la modélisation mathématique ou la simulation. Certaines sciences connaissent grâce au « Big data » des développements considérables : la génomique, la climatologie, la physique des particules, l'astronomie. D'autres sont carrément bouleversées, comme les sciences humaines et la sociologie qui deviennent des sciences quantitatives, grâce à l'analyse des réseaux sociaux et à l'usage massif des Smartphones et autres objets connectés (voir aussi les villes intelligentes basées sur des mesures massives de comportement : exemple le projet Living Lab à Trente en Italie). On parle désormais de « physique sociale ». La médecine se renouvelle profondément grâce aux nouvelles possibilités d'analyse du génome, (voir par exemple : l'entreprise 23andMe qui offre des services basés sur l'analyse du génome de ses clients), et aux outils du « quantified self » en particulier par l'usage de montres connectées, etc. On peut aussi mentionner le « crowd computing » qui permet de faire appel au public, via des interfaces et des réseaux dédiés, pour aider à résoudre des questions scientifiques ou autres : par exemple, étudier des configurations de protéines, ou bien déchiffrer pour une numérisation ultérieure des manuscrits écrits en vieux français par exemple. Finalement, le fait que chacun puisse a priori facilement poser des questions très variées via l'analyse des données rendues publiques ou de ses propres données ouvre la perspective de découvertes et de services inattendus.

› **De nouvelles possibilités d'optimiser le fonctionnement de la société.** On parle ainsi de « villes intelligentes ». Les réseaux de transport pourront être reconfigurés en temps réel pour répondre aux mesures sur les flux de personnes, la distribution de l'énergie et les heures de consommation seront optimisées grâce aux compteurs « Linky » intelligents et à des mesures en temps réel de la météo. La sécurité des lieux publics et privés sera de même révolutionnée par la disponibilité de données multi-sources : caméras de surveillance, objets connectés portés par les individus, traces d'ADN que l'on peut désormais détecter dans l'atmosphère d'une pièce plusieurs jours après le départ de ses occupants…

› **Le développement de panoplies de services très ciblés**, individualisés par exemple pour une médecine personnalisée, des conseils de consommation (livres, films…) et la vie en général (ex : comment optimiser son sommeil en fonction des événements de la journée et de l'agenda du lendemain). Ces mêmes technologies permettent aussi la mise aux enchères en quelques micro-secondes d'espaces publicitaires à introduire dans les pages qui s'affichent durant les recherches Internet d'un utilisateur. Généralement, le marketing va devenir une science avec en particulier une mesure de l'impact en temps réel des messages, et un suivi très fin des comportements, dans les magasins ou sur les sites marchands.

› **L'open data,** c'est-à-dire l'accès libre et gratuit aux données, en particulier gouvernementales et des collectivités locales, avec l'espoir d'une démocratie participative et directe.
Pour se focaliser davantage sur les sciences du vivant et de l'environnement, on peut attendre des impacts importants sur les secteurs suivants :

› **Les sciences de l'environnement** qui vont bénéficier de la possibilité d'intégrer et de combiner des données de capteurs très variés, très multi-échelles (des satellites aux drones et aux capteurs dans les tracteurs et dans les champs) et avec suivi des évolutions. De fait, le changement climatique ne serait peut-être pas encore perçu, et en tous les cas ne serait pas apprécié pleinement, sans une capacité



d'analyse multi-source et à grande échelle des données.

› **La logistique** et les chaînes de distribution, en particulier les chaînes d'approvisionnement en produits frais, vont voir leur fonctionnement très optimisé, avec à la clé beaucoup moins de déchets, et des dates de péremption beaucoup plus précises et attachées à chaque produit par l'analyse de son « histoire » grâce à l'arrivée des multi-capteurs sur les produits eux-mêmes.

› **L'agriculture** se prépare également à une révolution. Pour donner un exemple, John Deere et AGCO (marques de machines agricoles d'occasion) ont ainsi entrepris de relier les machines agricoles entre elles, mais aussi les systèmes d'irrigation, des mesures sur les sols et sur les intrants, via éventuellement des drones, tout cela en plus d'informations relatives à la météo locale à court et moyen terme et de données sur les cours de bourse des produits récoltés et des matières premières, le tout afin d'optimiser les performances d'une exploitation agricole dans son ensemble.

Ce tour d'horizon extrêmement rapide et nullement exhaustif souligne l'importance et le large spectre des mutations attendues. Ce qui est clair c'est que l'on est en train d'assister à un transfert massif de pouvoir des acteurs économiques qui maîtrisent les techniques et les procédés de fabrication ou de services vers ceux qui maîtrisent l'information, c'est-à-dire qui détiennent les données et savent en tirer des régularités exploitables et des prédictions. C'est pourquoi, pour essayer de comprendre l'avenir, tant d'analystes se focalisent sur les GAFA (Google, Amazon, Facebook et Apple) et leurs stratégies, c'est-à-dire sur ces entreprises (et d'autres) très jeunes qui ont détrôné les acteurs traditionnels grâce à leur récolte unique et massive de données sur les utilisateurs, leurs machines, leurs comportements, et peuvent ainsi devenir les vrais donneurs d'ordre reléguant les autres entreprises à de la sous-traitance.

Bien sûr ce nouvel Eldorado annoncé s'accompagne de risques.

Les risques et les défis

Les risques concernent en premier lieu la vie politique. En vrac :

› Risque de surveillance généralisée, détaillée, en temps réel et à une échelle planétaire.

› Tentation de prédiction de comportements « déviants » avant le passage à l'acte.

› Croisement illégal et illicite de données.

› Cycles de décision raccourcis à l'extrême, en raison en particulier de l'utilisation de systèmes de décision automatiques, au détriment du temps de la réflexion et de la consultation.

› Cacophonie sur la décision politique si des experts auto-proclamés de l'analyse de données affirment n'importe quoi et s'appuient sur une pseudo objectivité pour dicter des sentences et des ordonnances.

› Découverte de corrélations stupides par manque de recul et de réflexion, mais avec le sceau de l'objectivité de l'« algorithme ».

› Recul de la solidarité par segmentation ultrafine des usagers. C'est certainement l'une des tentations dans le domaine de l'assurance, qui va tendre à privilégier des offres hyper-segmentées en fonction du profil mesuré des clients, au détriment de la mutualisation des risques.

Pour aller plus loin, il peut être intéressant de se reporter à une journée spéciale « Big data : adieu à la vie privée ? » organisée le 28 mars à la Cité des Sciences en partenariat avec AgroParisTech alumni. Cette journée visait à avertir les citoyens que nous sommes du phénomène du « Big data » avec ses promesses et ses risques, en particulier liés à l'usage, plus ou moins conscient et plus ou moins accepté, de nos données personnelles par tout un ensemble d'acteurs du Web. Lors de cette journée ont été notamment abordés les promesses et les risques dans le domaine de la santé, ainsi que la réalité et les conséquences possibles d'une surveillance généralisée. Les débats et interventions des spécialistes seront bientôt disponibles sur le site de la Cité des Sciences.

Les défis sont d'ordre technique, mais ils sont surtout humains.

Les défis technologiques sont liés aux quatre «V » évoqués dans l'introduction : Volume, Vélocité, Variété, Véracité. De ces quatre V, les deux premiers sont les plus exigeants en termes d'infrastructures. Il faut des capacités de stockage, d'interrogation et de visualisation des données performantes. De même qu'il faut être capable de traiter de gros volumes de données, ce qui peut impliquer de manière routinière du swapping en mémoire centrale, le recours à des clusters de calcul ou à du cloud computing. Certaines applications sur des flux de données demandent un traitement « à la volée » qui impose ses propres contraintes, en particulier sur les systèmes de requêtes et sur les traitements possibles.

Cependant, ce sont les défis en termes de compétences qui sont prééminents et vont conditionner l'avenir du « Big data ». La connaissance des nouveaux outils de stockage et de traitement des données est nécessaire, mais c'est surtout la compréhension des problèmes liés à l'exploitation de données multi-sources, très hétérogènes dans leurs formats et leur fiabilité, et la maîtrise des techniques d'intégration

de données et de ré-expression qui vont être déterminantes. De même, explorer massivement les corrélations potentielles et les signaux faibles demande des experts qui savent utiliser les méthodes adéquates, mais aussi, de manière tout aussi cruciale, qui sont conscients des risques de découverte de régularités fortuites sans signification réelle et savent comment s'en prémunir. La recherche de « causalités » au milieu des corrélations est encore un problème de recherche, mais devra également faire bientôt partie des compétences à maîtriser pour être un expert du « Big data », ce que l'on appelle maintenant souvent un « data scientist ». Pour finir, ces spécialistes devront aussi être informés de tous les aspects juridiques et éthiques, ainsi que des problèmes de sécurité, liés à l'exploitation de données massives comprenant souvent des données individuelles, voire intimes.

« Il est urgent de former des ingénieurs aux métiers liés aux big data »

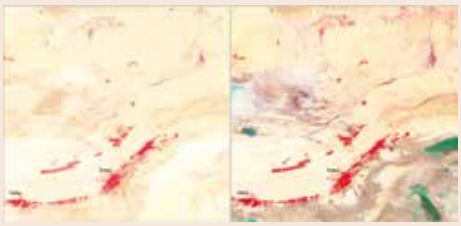
Les estimations sur les besoins en spécialistes de ce type sont faramineuses, se chiffrant par exemple à plus de 100 000 en France dans les 6 prochaines années et à plus d'un million aux Etats-Unis. Même si ces chiffres sont peut-être surestimés, ils donnent une idée de l'urgence de former des ingénieurs en nombre très significatif sur ces métiers, ainsi sans doute que d'organiser une formation continue adaptée. Des questions urgentes se posent telles que : Quelles devraient être les compétences minimales d'un ingénieur sur le « Big data » ? Que faut-il prévoir dans la formation de base ? Et que doit être une formation spécialisée ? À quel type de public peut-elle s'adresser ? Que peut-on attendre d'une formation de trois mois, six mois ou un an ? Combien de docteurs en sciences des données va-t-il falloir former pour irriguer les institutions publiques, les organismes de recherche et les entreprises privées sous peine d'être dépossédé de notre souveraineté sur la compréhension et la maîtrise du monde ? Des initiatives multiples, variées et désordonnées se mettent en place pour répondre à ce défi de formation. AgroParisTech et les écoles du secteur des sciences du vivant et de l'environnement doivent offrir une réponse forte, raisonnée et exemplaire à cette demande.

Il est sans doute opportun de terminer cette section en insistant sur l'importance du fait que, au delà des spécialistes des données et de leur analyse, chacun, en tant que citoyen, ait connaissance des risques liés à une numérisation socialisante du monde et de la vie, les comprenne, et devienne ainsi un acteur éclairé et vigilant dans la définition des politiques publiques sur ces questions. La formation des jeunes, et moins jeunes, peut et doit contribuer à cette prise de conscience.

Pour résumer, il est clair que le « Big data » n'est pas seulement un buzzword, un concept à la mode qui s'évanouira aussi vite qu'il est apparu et qui ne concerne que quelques « geeks ». Il correspond à une mutation profonde de notre rapport au monde et de nos processus de décision.

Ce numéro spécial dédié au « Big data » témoigne de l'importance de la révolution en cours pour notre secteur. L'article d'Isabelle Mougenot et Éric Delaître sur les « *données massives : analyse d'images de télédétection pour suivre les variations de l'environnement* » montre comment les nouvelles techniques d'observation de la Terre et de croisement de données multiples permettent des suivis en temps réel beaucoup plus fins des agrosystèmes partout dans le monde, ainsi que l'analyse de catastrophes naturelles rapides comme des inondations ou à évolution plus lente comme les sécheresses. L'article de David Makowski « *Méta-analyse : pourquoi et comment synthétiser des données d'origine diverses ?* » décrit la technique de méta-analyse qui consiste à rassembler des sources de données multiples pour répondre à des questions qui peuvent aller de l'évaluation de l'effet de certains médicaments ou de recommandations médicales à l'estimation de la responsabilité de certaines pratiques agricoles sur les émissions de gaz à effet de serre. L'article de Céline Lévy-Leduc et de Stéphane Robin « *Les nouveaux défis de la biologie moléculaire* » présente les révolutions successives en cours en biologie par l'introduction de technologies d'analyse de l'activité des gènes : puces à ADN et puces de nouvelle génération. Ici aussi, la science qu'est la biologie ne pourrait connaître son développement incroyable sans techniques puissantes de stockage et d'analyse de données. Finalement, l'article de Dominique Cagnon « *L'Internet des Objets : un Big data puissance 2 arrive* » fait prendre conscience de l'émergence d'une nouvelle étape majeure dans l'ère du « Big data », celle des objets connectés, c'est-à-dire d'objets capables d'effectuer en permanence un ensemble énorme de mesures sur notre environnement de vie, et de non seulement communiquer ces données avec les utilisateurs mais aussi et surtout entre eux grâce à l'Internet des Objets.

Sommaire du dossier



Antoine Cornuéjols (AgroParisTech)