2. *L'analyse de mégadonnées :* big data, data analytics

L'analyse de **corrélations** statistiques sur des données massives est une des caractéristiques du phénomène des « mégadonnées », les *big data*. Si on parle très souvent ici de la masse de données (stockées dans d'impressionnantes fermes de données, les *data center*), ce qui est important d'un point de vue théorique ce sont surtout les possibilités de nouveaux traitements statistiques, basés sur de nouvelles approches de la corrélation (on parle ici de *big data analytics* ou de *business analytics*) :

– On peut d'abord traiter aujourd'hui des données qui sont complexes et **hétérogènes**, à la fois dans leur nature (données structurées et données non structurées : textes, images, vidéos...), dans leur origine (interne, externe, géo-démographique, *open data* publiques...) et dans leur saisie (données des capteurs, données des mobiles, multi-canal, données en temps réel par *data streaming*, données « historisées » dans leur évolution...).

– On peut surtout travailler aujourd'hui sur le **croisement** de ces données complexes et hétérogènes, avec de nouvelles méthodes statistiques de recherche de régularité (réduction de variables, régressions, réseaux neuronaux...) et de recherche de singularité (signaux faibles, sévérité d'événements aléatoires, analyse inductive de situations-contextes...) : voir ci-dessous la section 3 sur l'intelligence artificielle et l'apprentissage automatique.

Les applications de ces analyses de corrélations-proximités sont impressionnantes, notamment dans le marketing. Le site Amazon a jeté depuis longtemps les bases de la recommandation sur le Web (si vous aimez ceci, vous aimerez cela). Mais, si le marketing avait toujours cherché à expliquer un comportement d'achat en fonction de quelques grandes caractéristiques psychologiques (par exemple celles du test BFI, *Big Five Inventory* [Goldberg, 1990] : ouverture à l'expérience, autodiscipline, extraversion, empathie, non-stabilité émotionnelle), ces caractéristiques peuvent aujourd'hui être **directement prédites** en fonction du simple usage d'un téléphone portable (avec seulement 36 indicateurs sur localisation, régularité, diversité des contacts, temps mis à répondre à un texto...).

Pour être réalisés en temps réel, ces calculs peuvent nécessiter une énorme puissance de calcul (Yahoo a révélé avoir connecté un *cluster* de 4 000 serveurs, qui fonctionnent en calcul parallèle) : en finance de marché, si les salles d'ordinateurs de la Bourse de Paris sont en fait à Londres, c'est que 160 km représentent déjà 1/1 000e de seconde de temps gagné pour des algorithmes qui travaillent au centième de dollars.

D'un point de vue théorique, le débat qui oppose modélisation et corrélation a été lancé par Chris Anderson (2008) : « La corrélation n'est pas causalité, mais cela n'a plus d'importance... laissons les algorithmes trouver les modèles que la science n'arrivait pas à trouver... analysons les données sans faire d'hypothèses sur ce qu'elles vont produire... avec suffisamment de données, les chiffres parlent d'eux-mêmes. » Le risque est ici de passer d'un océan de données à un véritable **océan de corrélations** (et donc peut-être à un océan de faux-savoirs). Le rôle de la démarche scientifique est-il seulement de prédire ou bien d'expliquer ? Peut-on construire quelque chose qui « fonctionne », mais que nous ne comprenons pas ?

D'un point de vue éthique, cette ruée sur les corrélations entre mégadonnées renouvelle des questions importantes. Nous entrons ainsi dans un monde de la prédiction et de l'évaluation systématique, mais pour assister les décisions de qui : des consommateurs, des médecins, de la justice... ou bien plutôt des banques, des assurances, de la publicité, de la police prédictive ? La maîtrise des données et de la prédiction ne renforce-t-elle pas les inégalités de pouvoir et les risques de discrimination ?

Mégadonnées, *big data* : une révolution pour la prise de décision ?

Vis-à-vis de la décision, l'analyse aujourd'hui possible de données massives (*big data*, mégadonnées) pose au moins deux grandes questions :

a) D'un point de vue technologique, l'analyse des mégadonnées représente-t-elle vraiment une « révolution » ?

Par rapport à l'informatique décisionnelle (*cf. Business Intelligence* au chapitre 4), l'analyse des mégadonnées est plutôt une conjonction de plusieurs évolutions (informatiques et statistiques), mais qui finissent en effet par créer ensemble une vraie rupture, en termes de vitesse, de volume et de variété.

Trois évolutions du côté de l'informatique, augmentant considérablement le volume et la vitesse :

– dans la phase de saisie d'abord : une **collecte massive** de données et de métadonnées, depuis les réseaux (transactions, réseaux financiers, Internet, médias sociaux...), depuis les téléphones (connexions, géolocalisation...), depuis les objets connectés (domotique, étiquettes RFID, Web des données...) et depuis tous les capteurs (climatiques, scientifiques, caméras...) ;

– dans la phase de stockage ensuite : des capacités presque sans limites (dans les *data centers*) et surtout une **nouvelle organisation des bases de données**. Dans ces bases NoSQL (*Not only SQL*, car elles ne sont plus orientées « tables » mais orientées colonnes, graphes, documents...), les données ne transitent plus sur un réseau : elles sont distribuées en milliers de nœuds (*Hadoop* ou *General Parallel File System* d'IBM) et les programmes s'exécutent sur chacun de ces nœuds ;

– dans la phase de traitement enfin : les **traitements en parallèle** peuvent s'exécuter beaucoup plus rapidement sur des supercalculateurs en grappes (les *clusters*) et les résultats sont ensuite agrégés (*Map-Reduce*, initié par Google).

Deux évolutions du côté des statistiques, augmentant considérablement la variété :

– on peut travailler aujourd'hui sur des **données complexes et hétérogènes**, à la fois à cause de leur nature multimédia, de leur origine (internes, externes, publiques...) et de leur saisie (multi-canal, en temps réel, « historisées »...) ;

– on peut travailler aujourd'hui sur le **croisement** de ces données complexes, avec de nouvelles méthodes statistiques de recherche de régularité (réseaux neuronaux, forêts d'arbres de décisions...) et de recherche de singularité (signaux faibles, sévérité d'événements aléatoires...).

b) D'un point de vue sociétal l'analyse des mégadonnées représente-elle vraiment un « danger » ?

On peut répondre qu'il y a en effet des discours dangereux :

– « Les chiffres parlent d'eux-mêmes ! » Tout calcul donne des réponses, certes, mais qui pose les questions ? La culture de la quantification et du classement risque d'organiser une société de la surveillance et du contrôle (*cf.* la « Gouvernementalité » algorithmique, Rouvroy *et al.* 2013)

– « Corrélation n'est pas causalité... mais cela n'a plus d'importance ! » Il suffirait de laisser les algorithmes trouver les modèles que la science n'arrivait pas à trouver... Mais ces « corrélations massives » peuvent très facilement justifier de pseudo-savoirs : astrologie, anticancéreux, homéopathie, créationnisme... mais aussi en finance ou en marketing.

– « À défaut du pourquoi, on peut savoir comment ça marche ! » Comme les théories sont certes toujours incomplètes et que les variables causales sont certes infinies (de la Cosmologie... aux Sciences humaines), il suffirait d'Observer, Calculer et Prédire : plus besoin de Théorisation. Mais le monde n'est pas un jeu d'échecs, et par définition ce qui est imprévisible ne peut pas être prédit. Faut-il seulement « prédire » ou bien penser pour « expliquer » ?

– « La masse de données peut compenser la qualité ! » Mais *big data* ne signifie pas *whole data*, il y a toujours des choix à faire. Une donnée a toujours un contexte historique et les mégadonnées sont donc toujours choisies et « cuisinées ». Même les mégadonnées sont toujours perçues puis représentées, avant de donner lieu à une information ou une connaissance (A. Einstein : « C'est la théorie qui décide de ce que nous sommes en mesure d'observer. »)