Rencontre Capitoul - Big Data

Joseph Saint Pierre, Université Toulouse

Jeudi 18 Février 2016

Pour débuter cet exposé une référence à un tout petit article du Canard Enchaîné du mercredi 3 février 2016, numéro 4971, en page 8.

Cet article cite un client du site « Amazon »qui se voit recommander le livre de Nicolas Sarkozy « La France pour la vie »en raison de son activité sur le site. Or ce client a commandé trois articles, des lingettes, du dégrippant, des figurines pour décorer les sapins. L'article traite de manière ironique et fort drôle la connexion entre les objets commandés et le personnage de Nicolas Sarkozy. Il y a aussi dans l'article une présentation ironique des techniques employées, les analyses, les recoupements par l'entreprise « Amazon ». Sans que le mot soit employé il y a une critique des mégadonnées, ou Big Data. Il est assez fréquent d'avoir des recommandations non adaptées par les grandes entreprises de l'Internet qui utilisent et ont promu le Big Data, « Google », « Amazon », « Facebook »etc...

On peut mettre à côté de cet exemple de recommandation non adaptée ce qui est écrit dans l'horoscope du journal 20minutes du vendredi 5 février 2016, numéro 2944, page 20, pour le signe du bélier : « Vos activités ne vous passionnent pas. Surtout à cause de la mauvaise ambiance qui règne autour de vous ».

Il arrive parfois que ce qui est écrit dans les horoscopes semble plus adapté que des recommandations fondées sur des analyses de type Big Data, or les horoscopes sont fondés sur des principes irrationnels et non scientifiques, les techniques du Big Data sont parfaitement rationnelles. L'explication de cet apparent paradoxe réside dans le rôle du hasard.

Cet exposé est particulier car il s'adresse essentiellement à des informaticiens administrateurs de systèmes et de réseaux en étant préparé par un mathématicien. C'est à dire un statisticien qui perçoit les mégadonnées ou Big Data comme le prolongement de ce que l'an appelait, dans les années 1990 et encore dans les années 2000, la fouille des données ou Data Mining. Et même de ce que l'on appelait dans les années 1970 l'analyse des données avec, toujours l'idée d'un changement très important par rapport à l'époque précédente et donc des statistiques classiques.

Voici l'extrait d'une de mes interventions lors d'une rencontre sur le Big Data en santé, le 3 novembre 2015, dans l'université de Toulouse.

« Je tiens à souligner que nous ne sommes pas passés directement de la statistique au Big Data, il y a eu des étapes intermédiaires. Le Data mining était déjà une évolution par rapport aux statistiques individuelles. Quand on remonte bien avant on peut dire que l'analyse des données était déjà une remise en cause du modèle causal. Les Big Data ne sont pas vraiment une nouveauté car la taille des données augmente depuis longtemps »

Dans l'histoire longue les statistiques et avant le calcul ont toujours été tributaires, des données disponibles, des moyens techniques, des méthodes.

Les abaques sont très anciens, chez les grecs, les chinois, les indiens. Les moyens les plus anciens sont fondés sur l'alignement de cailloux d'où l'étymologie latine du mot calcul, qui désigne aussi en français des cailloux dans les reins.

Le Moyen-Âge européen a vu l'arrivée des chiffres indiens et du papier chinois avec aussi un intérêt pour les méthodes de calcul arabes, c'est à dire l'algèbre. C'est aussi pour des raisons historiques compliquées qu'apparaît la volonté de mathématiser le hasard. Ces éléments techniques et théoriques ont permis d'arriver à une compréhension empirique de ce que l'on appelé la la loi des grands nombres et l'émergence à la Renaissance de la théorie des probabilités. Il y a un héritage qui n'est pas fondé uniquement sur le mot « grand »de cette période dans les traitements de données.

La Renaissance a été essentielle dans le développement de l'astronomie et dans la mise en cause du modèle géocentrique. Les lunettes et télescopes ont permis une très forte augmentation des observations et donc des données.

L'astronome danois Tycho Brahe (1546-1601) a eu un rôle important dans l'évolution de l'observation, voici ce qui est écrit sur la page wikipedia en français sur le rapport de Brahe aux données

`https://fr.wikipedia.org/wiki/Tycho_Brahe`

À une époque où prévaut encore le respect de la tradition et des anciens, il donne la priorité à l'observation, avec le souci constant de valider ses hypothèses au regard de celles-ci. Il prend grand soin de la fabrication et de la mise au point de ses instruments qui lui permettent de recueillir un nombre considérable de données. Bien qu'effectuées à l'oeil nu, ces mesures sont, à leur meilleur, au moins 10 fois plus précises que celles de ses prédécesseurs en Europe.

C'est à partir des très nombreuses données d'observations astronomiques de l'observatoire de Tycho Brahe que l'astronome allemand Johannes Kepler (1571-1630) découvrit les trois lois portant son nom régissant le mouvement des planètes.

L'observatoire de Tycho Brahe lui appartenait, par la suite de très grands observatoires appartenant à des états ont été créés comme celui de Paris en 1667 et celui de Greenwich à Londres en 1675, des observations en quantités très importantes ont été prises par de nombreux personnels et archivées.

Du 17ème au 19ème siècle des grands scientifiques ont été à la fois mathématiciens et astronomes et ont contribué en l'enrichissement des probabilités

et donc plus ou moins directement aux statistiques, Galilée, Huygens, Newton, Leibnitz, Euler, Laplace, Gauss etc... Gauss a inventé la méthode des moindres carrés, essentielle pour les statistiques et sans doute, implicitement au moins, pour le traitement de données massives. Adolphe Quetelet (1796-1874) en plus d'être mathématicien et astronome, fondateur de l'observatoire royal de Belgique, utilisa les statistiques pour la démographie et l'étude des populations, il créa le concept de « l'homme moyen »

La démographie, les recensements, pour les impôts, la conscription, les élections ont eu une grande importance dans le développement des statistiques, l'origine du mot est lié à l'état. On peut retenir le personnage d'Herman Hollerith (1860-1929) qui a été recruté comme statisticien au bureau du recensement des Etats-Unis et qui a participé au dépouillement du recensement de 1880. Hollerith crée par la suite une machine statistique avec un système de cartes perforées. Cette machine permettra d'accéler le traitement du recensement de 1890, passant de 9 ans pour 1880 à 3 ans Hollerith crée en 1896 la société Tabulating Machine Co. qui deviendra en 1917 International Business Machines Corporation, c'est à dire IBM Cela bien avant l'invention de l'ordinateur.

Les recensements organisés par l'Institut National des Statistiques et des Études Économiques (INSEE) ont changé depuis 2002, afin d'avoir une estimation plus continue de la population, cela rappelle un petit peu les principes du Big Data. Depuis 2010 l'INSEE a commencé à utiliser Internet pour recenser la population.

En 2008 les élections présidentielles des États-Unis ont mis en avant un statisticien spécialiste de la prédiction des résultats des matchs de baseball et résultats des élections, Nate Silver qui, comme l'a écrit un journaliste du Monde a été bombardé « pape »du Big Data. Dans ses interventions Nate Silver minimise fortement la valeur prédictive du Big Data et appelle à la prudence. Nate Silver est beaucoup sollicité pour les élections de cette année 2016.

Les élections présidentielles des États-Unis de 2008 ont été aussi caractérisées par l'importance du rôle de tweeter et cela a contribué indirectement à la croissance du Big Data.

Les données deviendront encore plus massives et continues avec le développement actuel de l'internet des objets.

les statistiques produites pour les recensements sont assez simples mais concernent de très grands échantillons. En principe les statistiques assez fines sont faites sur des sous ensembles, soit au niveau global d'un pays soit sur la population d'une partie géographique (commune, département, état, etc...).

La collecte des données pour les recensements, comme pour les élections se fait de manière décentralisée, l'ensemble des supports de données brutes, c'est à dire les fiches pour les rencensements, les bulletins pour les élections ne sont pas centralisés. Dans le cas des élections, tous les bureaux de vote établissent les décomptes et les calculs de pourcentages au niveau de bureau et ces données « traitées »qui sont transmises à des niveaux plus globaux, locaux, régionaux, nationaux, suivant les types d'élection.

Lors d'élections nationales, des estimations très précises, depuis les années 1970 sont connues dès la fermeture des derniers bureaux de votes, en général à 20h en s'appuyant sur des dépouillements et des calculs effectués sur des bureaux de vote fermant à 18h. Depuis quelques années des estimations circulent sur Internet peu de temps après 18h. Cette évolution a un rapport avec celle qui mène des statistiques au Big Data, où la rapidité des traitements devient un critère important.

L'idée que les données obervées servent à faire des prédictions reste présente et même primordiale dans le Big Data et il serait illusoire de croire que l'augmentation massive du nombre des données traitées rend le monde plus prévisible. Cette illusion est en partie fondée sur l'oubli de la variabilité des données. Les données de plus en plus nombreuses ne permettent d'avoir une lois des grands nombres bien plus « forte ». Le monde reste profondément incertain et le Big Data ne permettra pas vraiment de faire l'économie des apports du calcul des probabilités.

Comme dans les statistiques « classiques », il est délicat d'extrapoler à partir d'observations même massives.

Le Big Data est souvent présenté comme la fin de la théorie, la fin de la causalité et l'émergence de techniques fondées sur les corrélations.

Lors d'un exposé précédent, j'ai abordé le sujet de la confusion des variables. Cela correspond assez bien à la citation attribuée à David Spiegelhalter :

« There are a lot of small data problems that occur in big data, they don't disappear because you've got lots of the stuff. They get worse »

Un exemple amusant de confusion de variables concerne la liaison entre le fait de dormir avec des chaussures et le fait de se réveiller avec un mal à la tête, cette liaison pouvant être explicable par une troisième « variable », la consommation d'alcool. La consommation d'alcool pouvant amener, assez rapidement, à certains oublis comme celui d'enlever ses chaussures et pouvant aussi amener plusieurs heures après une veisalgie, c'est à dire ce que l'on appelle communément une gueule de bois.

Il existe des méthodes statistiques multivariées fondées sur l'étude des covariances ou des corrélations qui sont assez anciennes, suivant les sources l'analyse en composantes principales date de 1901 par Karl Pearson ou par Harold Hotelling dans les années 1930.

Un des changements les plus importants apportés par le Big Data et sûrement un des moins faciles à percevoir concerne les sciences sociales et il n'est pas vraiment technique.Le développement d'Internet a créé des forums, des pages web, puis des blogs, des pages sur Facebook, des comptes tweeter et cela créé énormément de textes assez facilement disponibles par des outils informatiques. Les institutions et même certaines entreprises mettent des données disponibles à tous et le nombre de données ouvertes augmentent et concernent des entités de taille diverses, des états, de très grosse entreprises mais aussi des petites communes. La mise en connexion des données disponibles devraient permettre de donner des descriptions de la société. Il existe déjà ce que l'on appelle le journalisme des données ou data journalisme :
`https://fr.wikipedia.org/wiki/Journalisme_de_donn\%C3\%A9es`

Il pourrait, de manière analogue, exister une science sociale des données, cela nécessiterait d'autoriser des écarts aux méthodes plus traditionnelles fondées sur des enquêtes, des questionnaires, des entretiens. Cela semble plus difficile que l'apprentissage de nouvelles techniques, de plus de science. Les données sociales « ouvertes »existent depuis l'antiquité, mais la collecte et l'analyse de tous les textes libres anciens est fort difficile, les équivalents modernes sur des supports informatiques sont beaucoup plus facilement récupérables et avec des traitements simples il est possible d'obtenir des informations.

Le passage de SQL à NOSQL peut sembler n'être qu'une modification de l'organisation des bases de données entraînant des changements de logiciels, de langages, d'organisation des données, mais cela correspond aussi assez bien à des questions ouvertes et fluctuantes. Les sciences sociales ne sont pas nécessairement prêtes à changer rapidement leurs méthodes. Les grandes sociétés qui ont popularisé le Big Data comme Google, Amazon, Facebook interrogent pratiquement jamais par questionnaires mais elles étudient, par les méthodes de Big Data qu'elles utilisent et promeuvent, les comportements des utilisateurs qui passent par leurs services. Les internautes, la très grande majorité, sont directement concernés par le Big Data car leurs utilisations créent les données qui sont analysées par ces sociétés. Il est absolument indispensable qu'il y ait une connaissance minimale, la plus large possible, de ce que représente le Big Data, et que cela ne reste pas un domaine réservé aux scientifiques spécialistes des méthodes, des logiciels, des systèmes. Cela pour éviter que cela paraisse plus relever de la magie que des sciences.