A paraître dans :
Livre Blanc Tome 2 - "**Contribution des outils numériques à la transformation des organisations de santé**", *pour la* Commission des Affaires Sociales, Assemblée Nationale, 2019.

Corrélations artificielles *vs* intelligence des causes

Giuseppe Longo

Centre Cavaillès, République des Savoirs,
CNRS et Ecole Normale Supérieure, Paris
School of Medicine, Tufts University, Boston
http://www.di.ens.fr/users/longo/

De la méthode scientifique : les données suffisent-elles ?

Selon H. A. Simon (1977), l'un des fondateurs de l'Intelligence Artificielle, un ordinateur aurait pu découvrir les lois de Kepler sur la base des données de Tycho Brahe. Sommes-nous sur le point de remplacer l'intelligence scientifique par l'intelligence artificielle ? Revenons aux sources ...

A l'âge d'or de l'astronomie arabe (IX – XIII^{e} siècles), les scientifiques avaient déjà un nombre énorme de données sur les dynamiques planétaires visibles. Parmi eux, Ibn Yunus (Egypte, fin X^{e}) avait calculé les positions de toutes les planètes connues en produisant d'immenses quantités de données astronomiques, utilisées plus tard dans les tables Alphonsines (Espagne catholique, 1483). Le cadre théorique de l'époque était géo-centré (Ptolémée), ce qui, du point de vue mathématique, est parfaitement justifiable ; en effet, *tout ensemble fini de points sur une ellipse autour du Soleil peut être interpolé par suffisamment d'épicycles centrés sur la Terre*, les uns circulant sur les autres. En termes modernes, il est question de somme de séries convenablement centrées. Les prévisions ? Mathématiquement, c'est l'enfer, c'est même impossible ... sauf pour la Lune, évidemment ; on ne peut se baser que sur le passé, et ce seulement si on a de la chance. L'astrologie aussi faisait partie des compétences des astronomes de l'époque, (Blake 2016) et la certitude des prévisions des destinés des hommes donnait lieu à un débat très animé, (Livingston 1971). Que dire des ordinateurs ? Préfèrent-ils des épicycles ou des ellipses ? Osent-ils faire de l'astrologie ? On y réfléchira ...

Des décennies après Copernic (1473-1543), Tycho Brahe (1546-1601) a certainement contribué, par ses données, à la naissance de la nouvelle astronomie, mais ces données ne sont pas suffisantes, disions nous, pour changer de perspective : la théorie de Ptolémée les organise très bien en épicycles. En revanche, les allers et retours des planètes ("mouvements rétrogrades") sont incompatibles avec un principe fondamental de la révolution en physique, le principe d'inertie de Galilée (1564-1642). Copernic connaissait peu l'astronomie des arabes et n'avait pas ce principe – une source d'inspiration importante pour lui a plutôt été la *perspective* inventée par la peinture italienne, construction géométrique capable d'expliciter et modifier le point de vue de l'observateur (van Frassen 1970). La fin du géo-centrisme est tout d'abord un changement métaphysique et théorique, un changement de ''perspective'', dont nombreux ont payé le prix. C'est ainsi que la conjonction du point de vue de Copernic, du principe de Galilée, des données de Brahe et des propriétés de Kepler a permis de

comprendre différemment les mouvements planétaires, même les plus familiers. Leur cadre d'intelligibilité a totalement changé : il se base sur des propriétés physiques (le mouvement inertiel) et géométriques (lois de Kepler) incompatibles avec l'approche géo-centrique. Dans le temps, on arrivera à unifier la chute des pommes, les mouvements des corps célestes et leurs causes en termes de courbure de l'espace-temps relativiste ….

Voilà l'unité et la force d'une pensée scientifique : les régularités émergentes des Big Data, si nombreuses déjà chez les astronomes arabes, n'auraient jamais produit ce changement de système. Pour quelle raison devrait-on décider, contre toute doxa, de prendre ‘‘le point de vue du Soleil’’ ? de penser l'inertie, ce mouvement rectiligne et uniforme qui ne peut être observé nulle part, comme ‘‘état par défaut’’ de la matière ? d'inventer des équations sans sens physique (Newton), pour arriver à leur donner un sens grâce à une nouvelle géométrie de l'espace-temps (Einstein) ? Même l'imposition d'un principe d'optimalité ne pourrait suffire sans ces décisions et ces principes.

Seulement l'insensibilité à l'histoire de certains scientifiques, surtout américains reconnaissons-le, peut faire oublier le parcours historique richissime de la construction de la connaissance humaine – il faut ajouter à cela l'ignorance de la puissance interpolatrice des mathématiques qui peuvent projeter partout des épicycles, voire des régularités, même dans le hasard.

Le hasard des corrélations fallacieuses dans les grandes bases de données

L'idée que l'on puisse remplacer la construction de connaissance scientifique par une quantité suffisante de données est allée plus loin ces dernières années. « Correlation supersedes causation … with enough data, the numbers speak for themselves … No semantic or causal analysis is required » (Anderson 2008) : plus on a de données, plus les corrélations, que seule une machine peut trouver, permettront d'agir – point besoin de comprendre. Non, ça ne marche pas ….

D'une part, et les statisticiens le savent très bien, « si on *torture suffisamment* les données, elles finiront bien par confesser » ; de l'autre, « si on *a suffisamment* de données on peut y trouver n'importe quelle corrélation » - précisons cela. La ‘‘torture’’ la plus fréquente consiste dans le forçage d'un biais dans le choix des observables, de la mesure … un regard et une mesure biaisés permettent de lire à leur guise tout phénomène : on choisit de mesurer (comment ?) certains observables et pas d'autres (la couleur de la peau par exemple dans la ‘‘justice prédictive’’ (Garapon Lassègue 2018)). Quant aux corrélations, les mathématiques nous disent que, justement, quand on a ‘‘beaucoup’’ de nombres, on en trouve, *toujours*. Bref, dans (Calude Longo 2017), des résultats classiques de Théorie Combinatoire des Nombres permettent de démontrer que *quelle que soit* la corrélation entre nombres que l'on se donne, on peut calculer un nombre, *m* disons, tel que *toute base de données* ayant au moins *m* éléments contient la corrélation donnée. Ce n'est donc qu'une question de taille, car *toute* "base de données" (ensemble de nombres) *suffisamment* grande (avec au moins *m* éléments*)*, même produite au hasard (des lancements de dés), contient des régularités avec les caractéristiques demandées – dans énormément de nombres on trouve donc ‘‘n'importe quoi’’, mieux : ‘‘ce que vous vouliez, *a priori*’’. Le nombre *m* est très souvent immense, toutefois les résultats de la théorie de Ramsey utilisés dans (Calude Longo 2017) en produisent de l'ordre de grandeur des Big Data stockés dans nos gigantesques cluster d'ordinateurs.

Malgré ces résultats négatifs, la foi dans les Big Data contamine aussi certains secteurs de la médecine (Issa et al. 2014). Par exemple, contrairement à ce que préconisait l'étiologie du cancer basé sur le Dogme Central de la Biologie Moléculaire (le ‘‘phénotype cancer’’ a pour causes primaires des mutations somatiques), on constate la diversité imprédictible des « myriad mutations afflicting individual cancer cell genomes » (Weinberg 2014) tout comme « tumors without mutations » (Versteg, 2014), voire « cancer cells [that] display … a mutational burden similar to and perhaps even lower than that of adjacent normal cells » (Gatenby, 2017). Par conséquent, on ne peut proposer diagnoses, prognoses ni thérapies basées sur l'ADN des cellules cancéreuses, contrairement à ce qui avait été promis, et dans l'immédiat, lors du décodage du génome humain (2001 !). Mais, puisque « generating

large data sets became an almost-addictive undertaking », on collectionne les données de tous les « ‘‘omics’’ ... genomes, transcriptomes, proteomes, epigenomes, methylomes, glycomes ... » (Weinberg 2014). L'IA permettra des détecter des régularités et de proposer des diagnoses, prognoses et thérapies sans en savoir plus sur l'étiologie de cette maladie dont l'incidence a doublé en 40 ans. Quand les ‘‘omics’’ concernent dix millions de patients, un projet en cours, on frôle le nombre de données demandées par la théorie de Ramsey plus haut et … on en déduit ‘‘n'importe quoi’’ (Longo 2018).

Est-ce que, au moins, les prévisions automatisées des comportements, des consommateurs aux justiciables, peuvent s'avérer ? Oui, si elles arrivent, comme c'était parfois le cas dans l'astrologie des anciens, à *canaliser* et *uniformiser* ces comportements : quand on conseille les livres les plus achetés ou que l'on incrimine en priorité les individus des groupes sociaux les plus risqués, on calcule facilement quels livres ou criminels sont à prévoir. Le mythe d'une IA tout puissante dévoile alors son vrai visage : une vision et une praxis ‘‘instructive’’ de la vie et de l'histoire. Il faut que les hommes suivent la ‘‘règle formelle’’, sans sens : alors l'interaction homme-machine et les prévisions automatisées seront pleinement efficaces.

C'est n'est qu'en s'extrayant de cette vision qui subordonne l'homme à ce qui est mécanisable que l'on pourra utiliser au mieux ces formidables machines digitales qui sont en train de changer le monde : à nous de les insérer dans l'histoire afin d'enrichir notre socialité, au lieu d'en réduire la diversité.

Références (voir https://www.di.ens.fr/users/longo/download.html)

Anderson C. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*.

Blake, S P. 2016. *Astronomy and Astrology in the Islamic World*, Edinburgh.

Calude, C., Longo, G. 2017. The deluge of Spurious Correlations in Big Data. In *Foundations of Science* Vol. 22, 3, pp 595–612.

van Frassen C. 1970. *An introduction to the Philosophy of Space and Time*, Random House, NewYork.

Garapon, A., Lassègue J. 2018. *Justice digitale*. Presse Univ. France, Paris.

Gatenby, RA 2017. Is the Genetic Paradigm of Cancer Complete? *Radiology*, 284:1–3.

Issa NT, Byers SW, Dakshanamurthy S. 2014. Big data: the next frontier for innovation in therapeutics and healthcare. *Expert Rev Clin Pharmacol.*, 7, 293-298.

Livingston, John W. 1971. Ibn Qayyim al-Jawziyyah: A Fourteenth Century Defense against Astrological Divination and Alchemical Transmutation, *J. American Oriental Soc.,* 91 (1): 96–103,

Longo, G. 2018. Information and Causality: Mathematical Reflections on Cancer Biology. In *Organisms. Journal of Biological Sciences*, vol 2, n.1.

Simon H.A. 1977. Does Scientific Discovery Have a Logic? In: *Models of Discovery*. Boston Studies in the Philosophy of Science, vol 54. Springer, Dordrecht.

Versteeg, R 2014. Tumors outside the mutation box, *Nature*, vol. 1.

Weinberg, R. 2014. Coming Full Circle - from endless complexity to simplicity and back again. *Cell* 157*,* March 27.