Big Data & "Net" Savoir: Entreposer, Classer, Penser

Jean- Christophe Thalabard
MAP5, UMR CNRS 8145
Université Paris Descartes & APHP

18/07/2013

Table des matières

1 Introduction

1.1 Pourquoi intervenir ?

Je connais la réputation de l'Ecole d'Eté de Corte depuis de nombreuses années et les efforts inlassables de Mario Fieschi pour promouvoir un lieu de réflexion, prospective et échanges autour de l'Information médicale au sens large. Ma trajectoire personnelle ne me rendait guère qualifié pour intervenir dans cette enceinte et les circonstances ne m'ont guère permis au cours des années passées de venir écouter et m'instruire. Je ne suis pas sur que la situation ait profondément changé, mais Mario a su vaincre, par son insistance amicale, mes réticences naturelles et cette année je me retrouve devant vous. Je n'ai pas de compétence dans le domaine des Big Data sous ses aspects techniques aussi je voudrais partager avec vous quelques réflexions sur l'évolution du concept de données et leur accessibilité, telle que je peux la percevoir maintenant au terme, ou presque, d'une vie professionnelle dans le champ de la santé.

1.2 A propos du titre

Mon propos fait référence dès son titre à deux monuments de la littérature du XXième siècle, JL Borges et Georges Perec, auprès desquels nous nous permettons de faire de larges emprunts, dans une tentative timide de montrer que le phénomène des Big Data est autant si ce n'est plus un phénomène social que scientifique proprement dit, renvoyant au fonctionnement même de l'esprit humain qui tente de saisir le monde qui l'entoure, et dont il est bien souvent incapable de prévoir les évolutions.

1.3 *Le Cygne noir, la puissance de l'imprévisible* Partie II "Les prévisions sont tout bonnement impossibles" (Taleb, 2007)

Quand je demande aux gens de citer trois technologies récentes qui exercent un impact considérable sur notre monde, ils pensent généralement à l'informatique, à l'Internet et au laser. Toutes trois n'avaient été ni planifiées, ni prévues ; elles furent décriées lorsqu'elles apparurent et continuèrent à l'être bien après qu'on eut commencer à les utiliser. Elles ont eu des conséquences majeures. C'était des Cygnes Noirs...

Les années 60- 90 : Un monde qui se digitalise

Je me souviens... qu'en 1969, une année donc après mai 1968, je passais quelques concours nationaux et qu'un des sujets de composition française et de réflexion portait sur la rotation autour de la lune d'un équipage américain et sur les espoirs que cela pouvait susciter dans la tête de jeunes esprits scientifiques. Le bouclier de la capsule de retour trouva comme application le revêtement de la poêle Tefal. Comment aurions- nous pu le deviner ?

Je me souviens aussi que l'ENS- Cachan, seule parmi toutes les ENS et Ecoles d'ingénieurs, informait les candidats admis qu'une nouvelle filière était proposée intitulée Informatique, que nous considérions avec mépris dans notre sotte ignorance. Le cursus que j'eus la chance de suivre comportait une initiation obligatoire de 3 semaines à un langage informatique (FORTRAN), source d'énervement devant ces nouvelles contraintes, qui nous laissaient bien perplexes. Pourtant, dans un univers volontiers porté vers l'abstraction, le cours d'analyse numérique de Jacques Louis Lions nous démontrait que ces nouvelles capacités de calcul rendaient la théorie opérante.

Je me souviens alors des premiers exemples de reconnaissance automatique de l'écriture par les ingénieurs du centre de recherche de ce qui était encore les Postes et Télécommunications. Conscient de la nécessité de connaître mieux ces domaines, je décidais, par la suite, de suivre l'option informatique qui venait de s'ouvrir à l'Ecole nationale des Ponts & Chaussées. Merveilleux domaine que l'analyse numérique qui conjuguait le plaisir de belles théories mathématiques avec la nécessité de l'implémentation pratique. Il fallait ruser avec les matrices creuses, les maillages en grande dimension par rapport à la taille des mémoires des appareils de l'époque pour espèrer arriver à quelques résultats. L'ambiance y était particulière puisque le nombre d'enseignants dépassait celui des étudiants. Fortran, Cobol, Assembleur, PL1 étaient au menu des milliers de cartes que nous apportions religieusement au centre de calcul, chambre opératoire stérile, où ne rentraient que quelques happy few, merveilleux refuge en été pour qui souffrait de la canicule. La sanction apparaissait quelques heures, voire quelques nuits ou jours plus tard, sous forme de listings, plus ou moins épais, reflets de notre ignorance, qui devait se débrouiller dans la jungle des adresses hexadécimales. Le temps était autre qui forçait à penser sur le papier avant de tester.

Je me souviens d'avoir rencontré en 1974 le Pr Gilles de Gennes, à l'imagination si créatrice, alors qu'il prédisait l'essor des cristaux liquides, sans que sachions en saisir la portée. Pourtant les cartes disparaissaient, remplacées par un écran, d'abord à une ligne, puis plusieurs. Quelques outils de calcul apparurent dans les laboratoires, mais la logique restait centralisée. D'autres langages scientifiques apparaissaient que nous nous efforcions d'apprendre si ce n'est maîtriser (Pascal, Basic), déjà bien conscients de leurs caractères plus ou moins fugaces et leurs fragilités dans le temps.

Je me souviens de l'accouchement... d'une discipline, l'informatique médicale ... avec un acteur majeur en France, le **Pr. François Grémy**, dont nous sommes tous, à des degrès divers, des élèves. Le scepticisme des milieux médicaux était encore grand, appréhension d'une menace sur leurs savoirs et donc sur leur autorité ? Les années 80 virent apparaître les premiers ordinateurs portables, formidable délivrance d'un monde assez hiérarchisé du perforateur- vérificateur au chef de centre en passant par les pupitreurs et les analystes. Posséder le stockage était déjà un élément de pouvoir. Des supports nouveaux apparaissaient qui supplantaient les précédents avant de perdre, à leur tour, au fur et à mesure leur hégémonie, formidable perte glissante de données. C'est à cette période (1982-1983) qu'IBM décida de commercialiser les premiers ordinateurs individuels, gages d'autonomie, dont la facilité d'acquisition et mise en oeuvre s'opposaient au plan Mitra 15, avec sa tentation encore centralisatrice. L'intelligence artificielle apparaissait riche de promesses pour la médecine (projet Medicine de Carnegie Mellon ; Centre Mondiale Informatique de JJ Servan- Schreiber et Jean- Louis Funck-Brentano). Les conséquences étaient la multiplication des développeurs, l'émergence de multiples petites bases de données autonomes grâce à des outils comme DBASE,au prix d'une grande instabilité des langages, des supports et des hommes.

Je me souviens, aussi, ...d'un échange avec un directeur d'IBM...qui soulignait combien il serait difficile pour le patient de ne plus voir le soignant consulté écrire de sa main dans un dossier et le voir, plutôt, taper sur un clavier, le nez collé sur un écran ;

Je me souviens...des balbutiements des réseaux locaux/ distants Dès le milieu des années 1980, quelques académiques fureteurs eurent vent du projet américain de réseau civil de recherche EARN, piloté par IBM, qui permettait, via modem et ligne téléphonique ordinaire, aux physiciens du CERN à Genève de communiquer rapidement avec leurs collègues

FIG. 1 – L'hôpital, un lieu de tradition d'adaptation difficile à la modernité

aux USA. Ce réseau original s'était ouvert aux chercheurs intéressés, extérieurs à ces domaines. Le souvenir de cette période reste fort avec cette impression étrange, quasi magique, lorsque le message envoyé sur l'écran indiquait une cartographie géopolitique subtile entre quelques villes - noeuds de communications européennes (Montpellier, Genève, Darmstatt,..) et leurs homologues nord américaines. Quelques années plus tard et la diffusion par Apple des icônes et de la souris, l'ergonomie avait changé radicalement. Le Monde du 10/07 dernier rendait compte de la disparition de Douglas Engelbart (Stanford Research Institute), qui conçut le premier prototype de souris **dès 1964**, avec une vision extraordinaire des développements potentiels futurs (fenêtres multiples, vidéoconférence, courrier électronique, hypertexte). Windows apparaissait. Internet prenait une forme plus conviviale s'imposait pour étendre sa toile toujours plus loin.

C'est également dans cette décennie que le **minitel** permit à AJ Valleron et son équipe de développer le premier réseau français de connaissance sur les maladies grippales saisonnières, renouvelant la technique de cartographie sanitaire, qui permit à Charles Nicolle de comprendre la propagation du typhus, en développant des outils nouveaux (krigeage, distances temporelles...).

Je me souviens du CITI2 alors dirigé par P Lebeux et son idée de banque d'information génomique qui ne connut guère de succès

Je me souviens des engouements pour le projet HEGP- Newton

Et l'hôpital public dans tout cela : une mutation...bien lente

Un lieu de fabrique de savoirs dans la tradition

L'hôpital restait un lieu de formation au contact des maîtres (compagnonnage) avec toute la force des écoles. Le livre dominait plus que les journaux. La langue anglaise restait peu diffusée. Le savoir circulair de manière très hiérarchique (*Currents contents*), limité aux abonnements intra- service, d'accès soigneusement contrôlés, avec des possibilités de photocopie limitées, imposant un fastidieux travail de demandes de tirés à part et d'archivage, mais où le regard pouvait favoriser les débordements de son propre domaine d'intérêt.

Les années 90- 2010

Le décor se transforme

L'écran entre dans le décor des secrétariats puis des bureaux médicaux. L'abandon de la machine à écrire est difficile, marquant la perte d'une maîtrise lentement acquise et des privilèges associés, avec la transition par les machines électriques (IBM, Olivetti). L'espace

FIG. 2 – Le mouvement EBM

sonore se transforme, des maux sont induits. Les questions d'archivage, de compatibilités matérielles, de services d'assistance sont omniprésentes. La dotation aux personnels soignants par ordre hiérarchique du patron au CCA et enfin l'interne en passant par le cadre infirmier remet en cause quelques compétences d'adaptation à un monde nouveau. Les connections sont contrôlées, les accessoires "ludiques" proscrits (lecteurs CD, clés USB), jusqu'à ce que des raisons économiques externes et des contraintes réglementaires (codage des actes) imposent un équipement généralisé jusque dans les cabinets de consultation et les postes de soin et fassent disparaître bon nombre des contraintes antérieures. Les pratiques professionnelles ont profondément changé. L'avalanche d'information remplace le temps de réflexion. Les secrétariats se vident et/ ou se disqualifent tandis que les boîtes aux lettres des experts se remplissent. Le temps a perdu de son épaisseur.

La connection internet & les services éclatés

La rapidité de la diffusion dans le grand public d'outils simples de connexion internet doublée des promesses d'économie de fonctionnement a permis le déploiement généralisé des connexions au sein des services hospitaliers, même si ce déploiement en interne a du faire face à l'inadaptation de locaux, au respect d'une certaine attribution hiérarchique et la mise en place de verrous et filtres plus ou moins adaptés. Pour le soignant, un des aboutissements le plus marquant a été la disparition des micro- bibliothèques locales pour un accès en ligne aux revues scientifiques et médicales, véritable changement dans l'exercice quotidien, dès lors que le soignant pouvait sur son lieu même d'exercice convoquer un savoir devenu accessible à partir de quelques mots clés. La vision prophétique de Archie Cochrane et Ian Chalmers devenait réalité : la médecine pouvait se fonder sur des preuves objectives mises à jour et accessibles à chaque instant

La vision Bernardienne se renouvelle : un nouveau laboratoire apparait au sein même du lieu de soin à travers des nouvelles pratiques que sont l'analyse bibliométrique, la lecture critique, la recherche sur les recherches (méta- analyse), les recherches *in silico*.

De nouvelles richesses sont créées au sein de l'hôpital comme les banques de tissus biologiques, les collections : une nouvelle *économie* s'organise et se structure à travers des consortia, des cohortes géantes (Oakland, Gazelle, NOWAC, E3N, Eden...) générant un activité nouvelle (croisements de bases de données) et des métiers nouveaux ("réducteurs" de dimensionnalité)

L'informatisation du lieu de soin a fait rentrer l'écran dans la relation soignant- soigné.

Cela n'a pas été sans réticence, l'hôpital ayant été le dernier à rentrer dans ce mouvement. Ces systèmes, conçus initialement pour la gestion de l'activité et ses liens avec l'assurance maladie, ont progressivement intégré une dimension médicale. Quelques initiatives stimu-

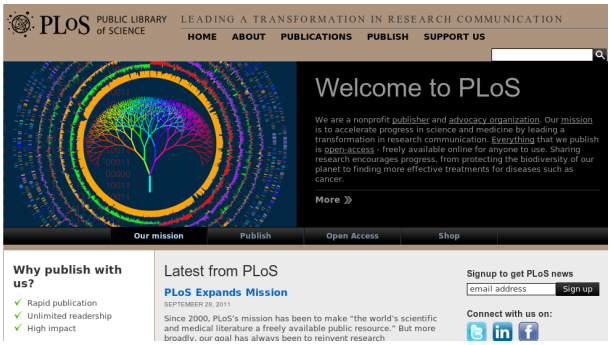


FIG. 3 – L'essor d'une diffusion libre des savoirs

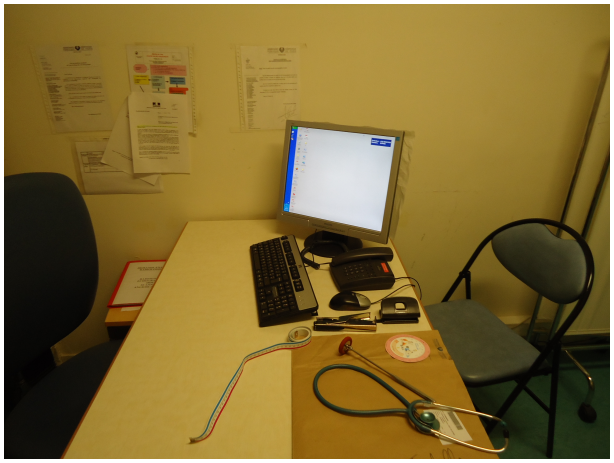


FIG. 4 – L'écran, la rencontre soignant- patient en présence d'un nouvel acteur, l'écran

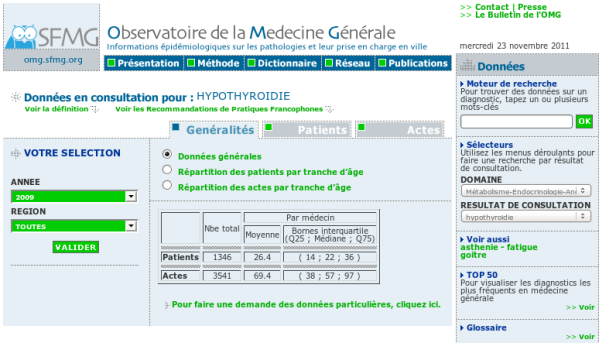


FIG. 5 – L'union fait la force : l'exemple de la SFMG

FIG. 6 – Apprentissage médical renouvelé : le patient virtuel

lantes, comme l'observatoire de la Médecine Générale développé par la SFMG, montrent tout l'intérêt de ce regard médical dès la conception de ces produits

Le formidable développement d'une informatique stimulée par les besoins d'animation (cinéma, jeux vidéo) trouve des applications en recherche, notamment dans l'éducation et la santé, comme l'illustre les possibilités d'apprentissages virtuels (Serious games, patients virtuels)

ou le jeu Foldit initié par l'université de Seattle (Firas et al., 2011, fol, 2008))

Le paysage a donc profondèment changé au cours de ces 30 dernières années, grâce aux possibilités nouvelles offertes de stocker, organiser et penser les connaissances sur lequel nous voudrions revenir maintenant

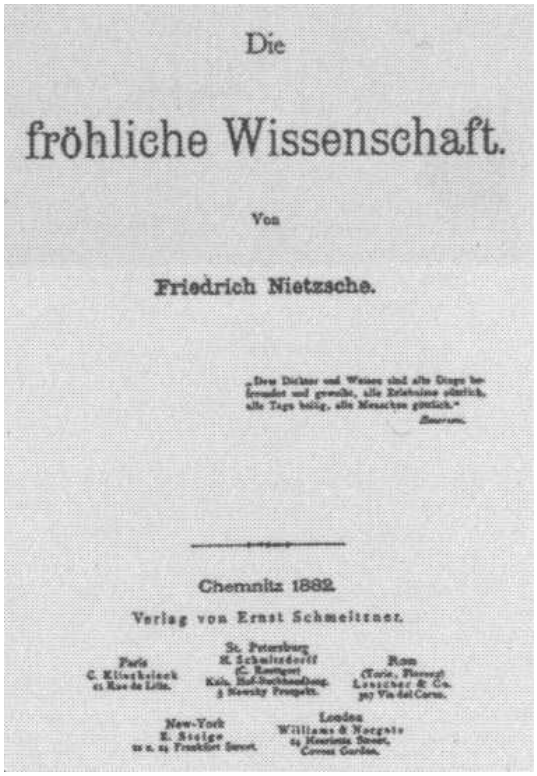
2 Le trépied d'une connaissance

Big Data : où les regards de Georges Perec et Jorge Luis Borges sont en écho

L'écrivain, le novelliste observent le réel et le traduisent dans leur monde. Saisir ce réel est complexe : une énumération complète, en fait quasi- impossible, ne peut être que terriblement ennuyeuse, alors qu'une simple sensation fugace peut réveiller toute une scène. Incomparable Borges qui fut capable de nous raconter le monde à travers sa propre cécité.

2.1 Collecter

Le souci d'une information aussi exhaustive que possible se retrouve dans deux nouvelles empruntées à ces deux auteurs


Die

fröhliche Wissenschaft.

Von

Friedrich Nietzsche.

Chemnitz 1882.

Verlag von Ernst Schmeitzner.

FIG. 7 – Le Gai savoir

- Chez Borges, il s'agit de la réalisation d'une carte parfaite d'un royaume imaginaire pour son roi, renvoyant à la LA RIGUEUR DE LA SCIENCE ("De la rigueur de la science", 1999)
  "En cet empire, l'art de la cartographie fut poussé à une telle perfection que la carte d'une seule province occupait toute une ville et la carte de l'empire toute une province. Avec le temps, ces cartes démesurées cessèrent de donner satisfaction et les Collèges de Cartographes levèrent une carte de l'empire, qui avait le format de l'empire et qui coïncidait avec lui, point par point. Moins passionnées pour l'étude de la Cartographie, les générations suivantes réfléchirent que cette carte dilatée était inutile et, non sans impiété, elle l'abandonnèrent à l'inclémence du soleil et des hivers. Dans les déserts de l'ouest, subsistent des ruines très abimées de la Carte. Des animaux et des mendiants les habitent. Dans tout le pays, il n'y a plus d'autre trace des disciplines géographiques (Suarez Miranda, Viajes de Varones Prudentes, Livre IV, Chapitre XIV, Lérida, 1658)"
- Chez Perec, il s'agit d'une tentative d'épuisement d'un lieu parisien "Il y a beaucoup de choses place Saint-Sulpice, par exemple : une mairie, un hôtel des finances, un commissariat de police, trois cafés dont un fait tabac, un cinéma, une église à laquelle ont travaillé Le Vau, Gittard, Oppenord, Servandoni et Chalgrin et qui est dédiée à un aumônier de Clotaire Il qui fut évêque de Bourges de 624 à 644 et que l'on fête le 17 janvier, un éditeur, une entreprise de pompes funèbres, une agence de voyages, un arrêt d'autobus, un tailleur, un hôtel, une fontaine que décorent les statues des quatre grands orateurs chrétiens (Bossuet, Fénelon, Fléchier et Massillon), un kiosque à journaux, un marchand d'objets de piété, un parking, un institut de beauté, et bien d'autres choses encore. Un grand nombre, sinon la plupart, de ces choses ont été décrites inventoriées, photographiées, racontées ou recensées. Mon propos dans les pages qui suivent a plutôt été de décrire le reste : ce que l'on ne note généralement pas, ce qui ne se remarque pas, ce qui n'a pas d'importance : ce qui se passe quand il ne se passe rien, sinon du

temps, des gens, des voitures et des nuages."

2.2 Organiser

Du réel à sa compréhension : la projection comme voie vers l'opératoire

Dans une nouvelle intitulée "Le langage analytique" de John Wilkins, JL Borges se référe à une encyclopédie Chinoise appelée "L'empire céleste de la connaissance bénévole" dont il attribue la découverte à un dénommé Franz Kuhn, dans laquelle les animaux sont classés selon les traits suivants

- Ceux qui appartiennent à l'empereur
- Les embaumés
- Ceux qui sont dressés
- Les cochons têteurs
- Les sirènes
- Un être fabuleux
- Des chiens errants
- Ceux inclus dans cette classification
- Ceux qui tremblent comme s'ils étaient fous
- Les non dénombreéss
- Ceux qui ont été extraits avec une brosse en poils de chameau.
- etc,etc,...
- Ceux qui viennent juste de casser le pot de fleurs
- Ceux qui de loin ressemblent à des mouches

Tandis que Perec dans "Penser/ Classer" propose une autre classification

- Animaux sur lesquels on parie
- Animaux dont la chasse est interdite entre le 01/04 et le 15/09
- Baleines échouées
- Animaux soumis à une quarantaine aux frontières
- Animaux en possession partagée
- Animaux empaillés
- Etcetera
- Animaux susceptibles de transmettre la lèpre
- Chiens d'aveugle
- Animaux reus comme dons dans le cadre du décès d'un oncle paternel
- Animaux qui peuvent être transportés en cabine
- Chiens errants sans collier
- Anes
- Juments avec un poulain

2.3 Penser

Classer/ Penser G Perec (Perec, 1985) "Que me demande-t-on, au juste ? Si je pense avant de classer ? Si je classe avant de penser ? Comment je classe ce que je pense ? Comment je pense quand je veux classer ? (...) Tellement tentant de vouloir distribuer le monde entier selon un code unique ; une loi universelle régirait l'ensemble des phénomènes : deux hémisphères, cinq continents, masculin et féminin, animal et végétal, singulier pluriel, droite gauche (...) Malheureusement ça ne marche pas, ça n'a même jamais commencé à marcher (...) N'empêche que l'on continuera encore longtemps à catégoriser tel ou tel animal selon qu'il a un nombre impair de doigts ou des cornes creuses"

Le problème des bibliothèques est un problème double : un problème d'espace d'abord, et ensuite un problème d'ordre (Perec, 1985)

Nous ne voudrions pas terminer cette digression vers les sciences humaines sans citer les fondements de la sociologie : ainsi, les travaux fondateurs d'Emile Durkheim inscrivent la sociologie dans le courant Comtien de l'époque, introduisant les régularités statistiques (exemple des taux de suicides) dans les outils sous- tendant des lois au sein de groupes humains. Nous pourrions citer, par opposition, les travaux d'un Erwin Goffman qui ne participaient pas de l'engouement pour le quantitatif mais mettaient en avant l'individu au sein de son environnement local avec les potentiels de stigmatisation locale, dans une approche de micro- dissection des rapports humains. Cette tension entre le local et le général et ces questions d'échelle de compréhension des phénomènes a bien sur pris une ampleur particulière avec l'avènements des réseaux sociaux, qui perdent, peu ou prou, leur ancrage géographique local pour créer de nouveaux environnements virtuels. Le professionnel de santé ne peut qu'être interpellé par l'engouement massif pour cette notion de "médecine personnalisée", slogan attirant pour celui qui expose ses symptômes, et sans doute simple reflet d'une séméiologie qui s'affine grâce à l'accès à des nouveaux marqueurs cellulaires et sub- cellulaires. Les cadres nosologiques changent mais la méthode classificatoire demeure.

3 Big Data : une question de terminologie

Big Data renvoit à deux notions : collection massive de données, d'une part, croisements de données, d'autre part, avec en filigrane l'exploitation de données collectées à d'autres fins. Elles contiennent en elles- mêmes l'idée que de l'information est présente au sein de jeux de données, au delà de leurs raisons initiales, scientifiques, administratives, etc... Les Big Data représentent en fait des ensembles mouvants de données hétérogènes produites en temps réel à partir d'une multitude de sources publiques ou privées, locales ou mondiales. Elles s'opposent ainsi aux sources classiques de données (registres, cohortes, etc...) dans leurs finalités et leurs dynamiques, par les éléments suivants :

- Les 3 V **V**olume, **V**itesse, **V**ariété (avec le 4ième V **V**éracité- pour IBM)
- Unité : petabyte (PB) $1.10^15 = 1000 terabytes$ (Doctorow, 2008)
- Système à feedback permanent : Algorithmes analytiques ⇒ Résultats ⇒ Décisions ⇒ Evolution systèmes

Dans "**The end of theory ?**" Chris Anderson(Wired, 2008) avance que le pilotage d'un tel système complexe repose sur une démarche empiriste (**Hume**) et non sur une démarche d'ingénieur avec mise en équations (**Descartes**)

4 Big Data : stocker

Big Data en biologie et santé (Viari, 2012) : un phénomène global

Le stockage des masses de données n'est ni récent, ni circonscrit aux sciences du vivant. Si depuis 2001 et la première cartographie du génome humain, suivis des progrès considérables des techniques de sequençages, une réflexion a été rendue nécessaire sur la gestion et l'analyse de telles masses de données, dans le domaine de la physique des particules, cette question du stockage de larges quantités de données est présente depuis longtemps (LHC du Cern) avec des flux de données similaires à ceux de Facebook ou Youtube. Ces masses de données posent des challenges particuliers. Laurent Alexandre, au parcours singulier et original exprimait dans une de ses chroniques du Monde, supplément Sciences, qu'"il devenait urgent que les spécialistes du cancer observent comment les astrophysiciens gèrent des exabytes (milliards

de milliards) de données produites. Sinon, le pouvoir risque de changer de mains" (Alexandre, 2012). Ils soulevaient les points suivants :

- Ils heurtent une culture de stockage et d'analyse des données encore assez artisanaux
- Ils nécessitent des réseaux spécifiques
- Les données sont de natures très variées allant de l'information moléculaire à l'information tissulaire, corps entier (imagerie), cliniques, environnementaux.
- L'expérimentation avec les Big Data prend une signification particulière distincte des schémas habituels de pensée

Si les premier calculateurs mécaniques, formes modernisées du boulier, n'avaient pour but que de rendre automatique des calculs répétitifs, ils n'avaient pas de vocation de stockage d'information qui continuait à être gardée sous trace papier dans des tableaux ou registres. Tout au plus des systèmes de microfiches permettaient-ils de réduire l'espace nécessaire, avec toute une gymnastique manuelle pour aller et venir entre ces différentes échelles. La possibilité de stocker, récupérer, transformer de l'information sous forme électronique, via un codage, marque, de fait, l'entrée dans l'ère informatique. Nous ne cessons de suivre la densification et l'optimisation énergétique qui permet de transformer avec soi sa bibliothèque entière ou presque. La notion de grandeur reste toute relative et le domaine des Big Data suit les progrès à la fois de ces possibilités de stockage et de leur accessibilité dans une course permanente où l'obsolescence suit de très près la modernité.

Cette accumulation de données est la conséquence du développement d'une course technologique à la puissance compensant la tendance régulière à la chute des coûts afin de maintenir un potentiel industriel attractif, plus ou moins irrigué par le système public et les fondations.

Réseau, stockage, puissance de calcul s'inscrivent dans une logique de développements technologiques, où les questions de la pérennité de l'information jouent un rôle central mais où également des questions sur leurs coûts énergétiques se posent.

Un vocabulaire propre

Le vocabulaire associé aux Big Data renferme quelques particularités puisqu'il y est question d'entreposer, de banques (avec son cortège de nouveaux banquiers, qui prêtent de l'information et la valorisent (!)), de fermes de données (et leurs nouveaux *fermiers généraux*), de containers (et leurs armateurs), de DataMarts (...et la stratégie de Target exposée par Antoine Geissbuhler (Geissbuhler, 2013)), des clouds divers, marquant la fin de repères simples temporo- spatiaux

La pérennité du stockage

Chaque nouveau support est supposé assurer une amélioration de la durée de vie du stockage. Ainsi, les supports de type CD ont remplacé très rapidement les disquettes et bandes magnétiques, qui n'avaient elles- mêmes pas cessé de présenter des améliorations successives, génératrices de modifications constantes de matériel. Pourtant dans le cadre du projet Arnano (`http://www.arnano.fr/`) (Larousserie, 2012), Alain Rey, chercheur CEA, a montré comment, sur 113 DVD enregistrś gravés entre 2004-2008, 12% présentent des signes de vieillissements importants. Leur durée de vie n'est en fait que de quelques dizaines d'années au maximum. De nouvelles idées apparaissent comme la proposition, dans le cadre du projet "Mémoire" þorté par l'ANDRA, d'une gravure particulière, sous forme de dépôt d'une couche de nitrure sur le saphir suivie d'une résine photosensible éclairée par un laser selon le texte à inscrire. Puis la résine est enlevée et une couche de verre est apposée dessus. Il n'y aurait alors plus aucun besoin de conditions particulières de conservation. Ceci ne résoud pas pour autant la question de la densité de l'information.

FIG. 8 – Images AM3 Amsterdam, Cold corridor, confinement aérolique, green computing, green IT (optimisation des infrastructures IT, bâtiment, énergie, exploitation)

Un coût énergétique inquiétant

Dans l'état des technologies actuelles, la course au stockage de l'information pose la difficile question du coût énergétique de ce stockage et sa fragilité. La quantité d'énergie dépensée par les centres de données informatique (qui comptent prés de 80000 servers) est impressionnante : le nouveau AM3 (St Denis) engloutit chaque mois l'équivalent de la consommation d'une ville de 20000 à 50000 habitants. Un chiffre de 2% de la consommation énergétique mondiale a été avancé par Greenpace, soit 104 milliards en 2020 au niveau européen. La réduction des coûts devient un enjeu financier pour venir à bout de la concurrence. Ainsi l'utilisation des eaux profondes pour le refroissement des serveurs et l'utilisation de l'eau chaude générée par les serveurs pour alimenter le chauffage urbain, serres, etc...`http://www.equinix.fr` sont considérées comme des pistes innovantes. La nature physique des composants à l'oeuvre dans le fonctionnement des machines modernes, volontiers déréalisés dans l'imaginaire collectif, ressort pour faire naître d'autres sources de richesses de la pensée humaine en action. Dès 1992, le sommet de Rio sur l'environnement avait attiré 40 entreprises au sein d'un Business Environmental Leadership Council (BELC).

Des nouvelles voies

(Driscoll and Sleator, 2013) "Nous ne pouvons résoudre des problèmes en reproduisant le type de pensée à l'oeuvre pour les créer" (Einstein). L'idée force dans les problèmes concernant le stockage des données est peut- être de penser plus petit et non plus grand ! L'ADN synthétique a une forte capacité de stockage : 455 exabytes/g. 40 ZB de données peuvent être stockées dans 90 g de ssADN. La disponibilité des enzymes de restriction, de la PCR, des ligases permet de juxtaposer les éléments comme des mots. L'ADN est une molécule très stable, comme démontré par la découverte d'ADN d'espèces disparues il y a plus de 10000 ans.

Ainsi plusieurs projets sont en cours autour de l'utilisation de cette capacité de mémoire bactérienne

- Joe Davis' microvenus
- JCV1-syn1.0 : bactérie avec génome synthétique. Marques de séparation formées de bases qui codent pour une adresse internet, le nom des auteurs et le titre (Joyce, Oppenheimer, Feynman)
- Church (Harvard) : petits bouts d'ADN se chevauchant. Codage d'un livre entier avec texte, images après conversion en html puis séquence binaire de 5.27 megabit transformée en 54898 blocs.
- Goldman : amélioration du taux d'erreur de lecture avec utilisation d'un codage par triplet et augmentation de la redondance de chevauchement

FIG. 9 – L'ADN bactérien, le stockage du futur ? (Driscoll and Sleator, 2013)

– Le futur : ADN ré-inscriptible

Toute la difficulté de ces mémoires ROM est la lecture qui nécessite de disposer du code, équivalent de la pierre de rosette ayant servi à décoder les hiéroglyphes.

Au delà de ces verrous technologiques, qui sous- tendent toute l'entreprise Big Data et génèrent une partie de son modèle économique, il reste bien sur la promesse contenue dans ces Big Data de leurs classification aux fins d'analyses

5 Big Data : classifier

Tirer de l'information utilisable nécessite une mise en oeuvre de filtrages successifs, dont il convient de bien contrôler, ou au moins comprendre, les limites par rapport aux objectifs de la classification, à moins que, justement, les promesses de l'exploitation des Big Data ne soient pas précisément dans la génération d'informations peu ou pas orientées. Les "*expériences*" Big Data ne se pensent et ne se gèrent pas comme des expériences traditionnelles. Elles impliquent des filtrages, des approximations, des petits arrangements à chaque étape pour gèrer l'incomplétude des données. Il faut apprendre à mettre ensemble médecins, pharmaciens, biologistes, bio- informaticiens, statisticiens. Les hiérarchies sont bouleversées dans un monde non stabilisé où le véritable "chef" est celui qui domine les avancées technologiques avec l'avance nécessaire et a des objectifs clairs vis à vis du positionnement de sa structure.

Dans le champ biologique et médical, nous pouvons repérer avec (Chen et al., 2013) quatre grandes sous- disciplines

– Bioinformatique qui engrange des données aux niveaux moléculaires et cellulaires
– Informatique pour l'imagerie qui correspond à des données tissulaires ou d'organes, avec des frontières qui s'estompent (ManuKea, Optique non linéaire, polarimétrie,...)

Domaines	Objectif de recherche	Type de données	Outils
Bioinformatique	Séquençage	Information sur séquence	Reconnaissance de forme
	Analyse de structure	Microarrays	Data mining
	Analyse d'expression	Spectromètre	Machine learning
	Arbre phylogénique	SNP	Visualisation
	Visualisation structure spatiale	Haplotypes	Annotation
Imagerie	Identification	DICOM	
	Segmentation	JPEG	
	Reconstruction	TIFF	
	Annotation	PNG	
	Indexation	GIF	
	Visualisation	BMP	
Information clinique	Observation papier	SC	Décision probabiliste
	Observation électronique	SB	Systèmes experts
	Diagnostic	Anamnèse	Vérification/ Validation
	Traitements		Résumé dossier
Niveau Santé Publique	Suivi des maladies infectieuses		
	Suivi des interventions cliniques		
	Suivi des facteurs de risque		

- Informatique clinique phénotypique
- Systèmes de santé publique

Ces grands domaines se retrouvent dans le tableau suivant :

L'avalanche de données : une nouvelle unité le petabyte (PB)

- Exemple de la recherche dans le domaine des RNA non codant (ncRNA) (Liu et al., 2012) Les développements technologiques récents permettent de recueillir des quantités de données importantes concernant des RNA ne codant pas pour des protéines. L'attention s'est initialement focalisée sur un type de ces RNA (ncRNA), à savoir les miRNA de 19- 25 nucléotides, qui semblent exercer un rôle de régulateur sur la traduction ou la répression des RNA codants. Plus récemment, des lncRNA ($\sim$ 200 nucléotides) ont été mis en évidence, qui joueraient sur la modification de la chromatine et la traduction des RNA codants. Il y a en fait actuellement plus de 45 formats pour les séquences avec des formats dominants correspondants aux structures les plus visibles (EMBL, GenBank, SwissProt, PIR). Ce manque de standardisation introduit des difficultés dans les nécessaires validations croisées.
- Le projet 1000 génomes (`http://www.1000genomes.org`)
- Le projet Huma-Num `http://www.huma-num.fr` : Huma-Num est une très grande infrastructure (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales. Pour remplir cette mission, la TGIR Huma-Num est bâti sur une organisation originale consistant à mettre en oeuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs.
  La TGIR Huma-Num favorise ainsi, par l'intermédiaire de consortiums regroupant des acteurs des communautés scientifiques, la coordination de la production raisonnée et collective de corpus de sources (recommandations scientifiques, bonnes pratiques technologiques). Elle développe également un dispositif technologique unique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche. Ce dispositif est composé d'une grille de services dédiés, d'une plateforme d'accès unifié (ISIDORE) et d'une procédure d'archivage à long terme.
  La TGIR Huma-Num propose en outre des guides de bonnes pratiques technologiques généralistes à destination des chercheurs. Elle peut mener ponctuellement des actions d'expertise et de formation. Elle porte la participation de la France dans le projet DARIAH `www.dariah.eu` en coordonnant les contributions nationales.

Cet ensemble d'activités représente en soi- même un domaine soucieux de régulations avec émergence d'une nouvelle discipline, la biocuration, et de nouveaux acteurs, au sens de l'histoire des sciences et des pratiques telle qu'abordée par des auteurs comme Monique Bensaud- Vincent à propos de la chimie et le personnage de Lavoisier.

La biocuration (Howe et al., 2008) `http://www.biocurator.org`

La nécessité de mesures révolutionnaires pour le data management, l'analyse et l'accessibilité.

Définition La biocuration correspond à l'activité d'organiser, représenter et rendre l'information biologique accessible aussi bien pour les personnes que pour les ordinateurs. Cependant, cette activité reste loin derrière en terme de financement, dévelopement et reconnaissance.

Le souhait d'actions urgentes Pour Alain Viari (Viari, 2012), il convient de réaliser que tout dispositif de gestion de données repose sur un modèle conceptuel permettant secondairement d'effectuer des requêtes. Il n'existe pas de modèle universel, que ce soit pour un même type de données ou pour des ensembles de données de natures différentes. Il convient donc de permettre l'**interopérabilité**e de bases de données différentes, tant sur le plan technique que surtout sur le plan **sémantique**. Ceci suppose tout un travail sur les **ontologies**, les **supports d'images**, avec dans ce dernier cas, la question de la compression sans perte d'information. Ainsi se dessine un ensemble d'actions nouvelles délimitant des métiers et organisations au sens de la sociologie des organisations, permetant des actions communes des auteurs, des journaux et des curateurs pour faciliter les échanges de données entre journaux et bases de données, et par voie de conséquence, d'evelopper une structure de reconnaissance, augmenter la visibilité des (bio)curateurs en tant que profession reconnue. Le rôle d'un *(bio)curateur* est multiple

- Extraire l'information des journaux publiés
- Connecter des informations de sources variées d'une manière cohérente et exhaustive
- Inspecter et corriger automatiquement les structures et séquences
- Développer un vocabulaire adapté
- Intégrer des bases de connaissance pour représenter des systèmes complexes
- Corriger inconsistances et erreurs dans la représentation des données
- Aider les utilisateurs pour les rendre plus productifs
- Piloter le plan de resources reposant sur le web
- Interagir avec les recherches

Penser à l'avance les modalités d'analyse permet certainement de mieux structurer la collecte des données et leur archivage, condition d'une biologie numérique fructueuse, résultat d'un véritable dialogue entre professionnels de santé et professionnels des systèmes. dans un idéal d'une curation collective, où la tâche d'annotation est reconnue comme partie normale du travail académique et comme le prolongement naturel de la tâche de publication mais pourrait également s'ouvrir au grand public. Ces activités méritent diffusion en elles- mêmes comme le montrent déjà certaines initiatives : Daphnia Genomics Consortium ; International Glossima Genomics Initiative ; Sloan Digital Sky Survey. Des sites propres ouverts apparaissent (WikiProfessional Life Sciences `http://www.wikiprofessional.org`) ainsi que des journaux dédiés (The Journal of Biological Databases and Curation) `http://www.oxfordjournals.org/our_journals/databa/about.html`

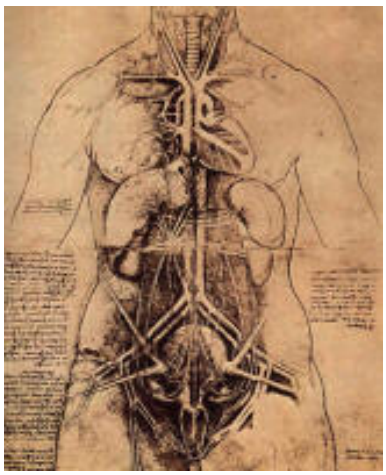


FIG. 10 – Tout regard est interprétation : le canal utéro- mammaire "expliquant" la montée laiteuse en post partum (Leonard de Vinci, planche anatomique)

6 Big Data : Penser

L'analyse de ces vastes jeux de données, hétérogènes et plus ou moins complètes définit des nouveaux champs de recherche méthodologiques, très actifs et rapidement évolutifs, redonnant de l'importante à des techniques anciennes (capture - recapture) (Amoros et al., 2007a,b, 2008), des emprunts à l'économie (méthodes à classes latentes). Les approches fréquentistes (Walter et al., 2012, 2013) sont largement concurrencées par les approches bayésiennes (Joseph et al., 1995, Demissie et al., 1998, Sewitch et al., 2013), affranchies des difficultés calculatoires, permettant d'injecter de la "**croyance**" *a priori* dans l'analyse. Tous ces développements sont facilités par l'importance du monde libre (Open Source, GNU, R, Bioconductor) et des échanges facilités entre développeurs et utilisateurs, comme bien montré dans l'exposé de P Degoulet dans ce même colloque (Degoulet, 2013).

Analyse de données hétérogènes en santé

L'analyse de données hétérogènes représente un grand bouleversement par rapport à tout le mouvement de rationnalisation qui, non sans mal a permis de "moraliser" des pratiques de recherche approximatives et favoriserl'émergence du mouvement EBM, socle de la médecine actuellement enseignée. Elle vise , en théorie, à faire resortir des **signaux faibles** par leurs capacités à travailler en grande dimension. Ainsi l'exploration systèmatique de banques de données génomiques peut laisser espérer repèrer des voies métaboliques communes ou des gènes communs et mieux comprendre la physio- pathologie de maladies rares, les classifier en opérant des rapprochements, et éventuellement proposer des nouvelles voies thérapeutiques. Il convient ici, sans doute, de distinguer Open Data de Big Data, dans la mesure où les premières correspondraient à des données bien définies dans leurs modalités de recueil et leur temporalité et mises à disposition avec un ensemble d'information et de limites, alors que les secondes constituent un ensemble de données non figées, floues, mal contrôlées. Une accumulation de données, aussi larges et hétérog‘enes qu'elles soient, ne constitue pas une connaissance en soi. Ce qui peut en être tiré représente une alchimie subtile entre diverses intentionnalités, celles des collecteurs d'un côté et celles des exploiteurs, de l'autre, sans oublier le rôle des intermédiaires.

Quelle valeur en terme de connaissance potentielle attribuer à ces données stockées ? Un groupement ou regroupement d'informations n'est, nous l'avons déjà dit, jamais neutre

et constitue une forme de pré- connaissance (comme il existe des pro- hormones ?). Ces pré- connaissances vont permettre de construire une connaissance valide selon de critères scientifiques à partir de l'exploitation des données. Cette étape nécessite une interprétation de résultats, constitutive de toute connaissance ("Story telling"). Il reste une dernière étape qui concerne l'application de ces résultats dans la vie courante sociale.

Desrosières, tout récemment disparu, qui fut une grande figure de la pensée statistique et sociologique contemporaine à travers notamment son ouvrage princeps sur la politique des grands nombres et l'histoire de la raison statistique (Desrosières, 1993), s'intéressait, ces derniers temps, à la prise en compte dans la production même des statistiques des effets incitatifs de leur publication sur le public concerné

Les registres

Nous évoquons rapidement les registres. Ils répondent à une définition bien précise avec validation des informations recueillies par recoupement de plusieurs sources. En France, 53 registres qualifiés par les autorités, couvrent environ 15% de la population : 14 tous cancers, 11 cancers spécialisés, 6 maladies cardio- vasculaires, 4 malformations congénitales, 11 maladies rares.

Les données des systèmes de protection sociale

- Les données couplées SNIIRAM- CNAMTS Exemple du dossier des pilules contraceptives, du dépistage, pharmaco- épidémiologie...les liens avec les bases de données du système de santé (données remboursements anti- coagulants en fonction classes d'âges : problème de pilules non remboursées...et habitudes de pharmaco- vigilance des soignants...) Au Danemark, croisement données registres, ventes produits, identifiant unique...premier registre danois des décès en 1875, des jumeaux en 1870 (!), des enfants adoptés (!). Rôle crucial des registres scandinaves (25 millions personnes) et australiens pour la surveillance des dispositifs....nombre limité de produits autorisés (exemple des prothèses de hanche) Pour Camilla Stoltenberg (Institut Norvégien de Santé Publique), il n'est pas éthique qu'elles [les grandes bases de données provenant de l'assurance maladie] soient si peu utilisées.
- Les données d'assurances privées
- Le registre Rhône- Alpin des accidents de la route a posé quelques difficultés d'homologation à ses débuts, du fait d'une définition inhabituelle. Il a permis de rectifier les estimations habituelles des blessés graves en lien avec des accidents de la voie publique (AVP) issues des constats de la gendarmerie (BAAC). Il n'a pas été sans intéresser les compagnies d'assurance (Martin et al., 2004, Laumon, 1998) Au passage, remarquons qu'un laboratoire de l'ISSTAR travaille depuis de nombreuses années sur l'analyse détaillée des AVP, posant la difficile question des conditions de possibilité et d'objectivité de mener une recherche sur un sujet sensible sociétal, qui pénètre largement la vie privée

Les analyses rétrospectives de jeu de données

Nous souhaitons rappeler ici les travaux menés par S Barker (Barker and Osmond, 1988, Barker et al., 1989, 1993) et ses collaborateurs à partir des données systématiquement recueillies par les sages- femmes à la naissance des enfants. La possibilité de faire un lien entre l'état de santé à l'âge adulte grâce au ssystème anglais du NHS et des paramètres simples recueillis à la naissance (terme, genre, poids du placenta, poids et morphotype de l'enfant),

représente un axe de recherche intense maintenant. Une telle découverte est heureuse car aucune méthode dite de haut niveau de preuve est recevable. Dans le même esprit, nous aimerions rappeler l'effet produit par l'ouvrage d'Emmanuel Todd (Todd, 1976) lors de sa parution, tout la finesse du travail de ce dernier ayant consisté à reconstituer tableaux et graphes à partir des statistiques officiellesdes pays du bloc de l'est concernant la natalité et la mortalité, pour annonce le déclin de l'empire soviétique, bel exemple d'information retirée de ce qui précisément voulait être masqué.

Le Data Mining

Le Data mining : suppose une définition de modalités d'extraction (distance, logique, etc...). La fouille de données peut faire resortir des entités nosologiques insoupçonnées via une voie métabolique commune ou un gène commun. Ainsi, dans le domaine des maladies rares, l'exploration systèmatique des banques de données génomiques permet des rapprochements, des identifications voire l'émergence de thérapeutiques.

Open Data

Elles permettent d'explorer quelques hypothèses et préfigurer des programmes d'action. La perception claire de leurs limites en font soit un outils intermédiaire, soit un outil de validation en situation réelle

Big Data

De par leurs fluidités, leurs intemporalités, leurs recompositions permanentes, les Big Data s'opposent aux types de données évoquées ci- dessus et ne visent pas les mêmes fonctions. L'accumulation de données aussi large et hétérog'ene qu'elle soit, ne constitue pas une connaissance en soi. Nous avons déjà évoqué son résultat comme résultante de diverses intentionnalités, celles Ã l'origine des données (à quand le TCL des intentionnalités) et celles de ceux qui les exploitent.

- Réussir à profiler un internaute en fonction de la nature et de la fréquence des sites fréquentés, de ses achats dans une finalité économique n'est pas exactement la même chose que comprendre le fonctionnement d'un organisme. Le moteur économique n'est a priori pas le même que le moteur de la science, même si diverses lectures de le vie de Galilée ont pu nous suggérer quelque moteur commun (Brecht, 1990). La capacité de la fouille des données web pour prédire le résultat d'une élection connait pourtant quelque succès comme illustré par la notoriété d'un Nate Silver, (Silver, 2012) qui s'illustra initialement comme "sabermétricien" (SABR : Society for American Base Ball Research) et son blog `FiveThirtyEight.blogs.nytimes.com/author.nate-silver`, rendant indirectement un rôle social aux nerds (No One Really Dies) par opposition aux geeks.
- Les techniques marketing se mettent à jouer un rôle dans le domaine humanitaire : **Datakind** Jack Porway `http://lesclesdedemain.lemonde.fr/innovation/le-big-data-au-ser a-54-2848.html` rapporte ainsi comment des technologies développées pour recommander des produits ou des films sont maintenant utilisées pour apporter des soins sur mesure à des communautés rurales et pour lutter contre la pauvreté, dans le monde entier. Le secteur des ONG se transforme
  - Les associations caritatives et les fondations n'ayant rien à voir avec la technologie ont maintenant accès à la numérisation de leurs propres informations, qu'elles peuvent utiliser plus efficacement pour accomplir leurs missions. L'association britannique Keyfund, qui fournit des bourses aux étudiants afin de les aider à acquérir

des compétences de bases, utiles dans la vie en société, ont constaté, grâce à l'analyse des données de leur programme de soutien scolaire, qu'elle pourrait économiser et fournir un service aussi efficace en regroupant deux niveaux de son programme. La Croix rouge de Chicago collecte les données incendies de la ville, mais commence seulement maintenant à les recouper avec celles produites par la municipalité, afin de prédire quels quartiers sont plus à risque, avant même que des incendies ne s'y déclare.

  - Vers l'ère du mobile Les programmes utilisant les réseaux mobiles dans le champ de l'humanitaire sont des catalyseurs pour l'adoption de la culture des données dans le secteur caritatif. Avec 90% de la population mondiale couverte par le réseau mobile, les associations se tournent de plus en plus vers ces services, produisant dans le même temps de plus en plus de données. *Mobilizing Health*, qui utilise les téléphones mobiles pour apporter des soins aux populations rurales en Inde, a remarqué, en analysant les données produites par les bénéficiaires du service qu'elle pouvait déterminer le meilleur moment pour envoyer des messages aux médecins, et même prédire quelles seraient leurs réponses, leur permettant de fournir les soins plus rapidement encore. La *Grameen Foundation* fournit des informations aux fermiers des zones rurales par téléphone mobile. En analysant les données, elle est en mesure de mieux comprendre quelles sont leurs besoins ainsi que de déterminer si son programme atteint réellement plus de fermiers. Les Nations-unies et la Banque Mondiale ont disséqué les usages des téléphones mobiles par certaines populations afin de mesurer les niveaux de pauvreté selon les zones.

- Des applications moins directement dans le champ médico- social et de l'évaluation des pratiques émergent. Citons quelques exemples
  - La cartographie des épidémies (Hay et al., 2013) Dans la lignée du courant utilitariste du 19ième siècle, et d'un Charles Nicolle cartographiant le Typhus à Tunis, les moyens informatiques modernes permettant de surveiller les maladies infectieuses, comme dans le cadre du projet HealthMap, qui donne en temps réel la survenue des cas (`http://www.healthmap.org/en/`, `http://www.flunearyou.org`, `http://www.influenzanet.eu`)
  - Des nouveaux outils permettent de nouvelles classifications, donc de nouveaux savoirs conceptuels. Le Pr. Raoult (Marseille) a ainsi identifié la catégorie des virus géants. "Pêcheur collectionneur" enthousiaste et passionné, sa science repose sur ces outils modernes avec des moissons toujours plus grandes et plus belles, proportionnelles à la puissance des séquenceurs : petit pêcheur à la traine, il est devenu pêcheur au chalut dans un monde nouveau. Encore fallait-il reconceptualiser pour penser : l'information ne respecte plus les entités vivantes patiemment répertoriées. Elle doit être prise globalement, communautairement, comme l'illustre la réflexion actuelle sur le biotope intestinal
  - L'oncologie avec le projet IBM Dr Watson (IBM) (Sillig, 2013) Partant de la notion a priori simple qu'un patient de dit pas toujours la vérité ou se comporte différemment selon la personne qui est en face de lui, IBM s'intéresse à la notion même d'interaction et à l'explication des rèponses apportées plus qu'à réponse elle- même. Ainsi, dans le jeu "Jeopardy", la machine se trompe parfois lourdement mais justifie toujours sa réponse. Elle apprend toujours des cas qui lui sont soumis. L'idée est alors de transposer ce principe de jeu à l'observation de l'interaction patient/ soignant pour retracer les étapes du raisonnement du soignant. La difficulté est que l'idéologie de l'ordinateur s'installe.
  - Utiliser les ressources du cloud pour comprendre l'impact de la variabilité génétique sur le fonctionnement du cerveau est actuellement un projet conjoint Inria (MapRe-

duce), Microsoft (Azure) (Antoniu et al., 2012)

L'importance des partenariats

Très peu d'organisations et d'associations sont en mesure de s'offrir des **Data Scientists** (scientifiques des données). Le potentiel de développement n'est ainsi généralement pas exploité au maximum. Selon une étude de l'institut Mc Kinsey, le manque de Data Scientists est estimé à environ 150 000 aux Etats-Unis. Il est donc important de combler ce déficit en montant des partenariats entre les talents de l'industrie d'une part, les finalités des associations d'autre part : c'est la mission que s'est donnée Datakind. Les scientifiques des données interviennent sur les usages possibles des gisements de données non exploitées, mais ont toujours besoin de consulter les spécialistes des associations du champ pour lequel ils effectuent ces recherches, afin de connaître au mieux les besoins réels. Après avoir travaillé avec Datakind, la petite ONG *Action for children* est passée du dessin des cartes à la main à l'utilisation des Systèmes d'information géographiques (**SIG**). Ce type de changements va aider le secteur à se tourner plus franchement vers les Big Data.

Gigascience `http://www.gigasciencejournal.com`

GigaScience vise à révolutionner la dissémination des données, leur organisation, leur compréhension et leur utilisation. Il s'agit d'un journal en ligne, en accès libre, qui publie des études "Big Data" concernant un large spectre des sciences de la vie. Pour atteindre cet objectif, le journal a adopté un nouveau format de publication : à côté d'un manuscrit classique, le journal fournit un lien vers le jeu de données étendu utilisé, les outils d'analyse développés et les ressources nécessaires de calcul sur le cloud. L' objectif n'est pas seulement les données de type "omic" et le champ des données à large flux, actuellement pris en charge par de nombreux entrepots publics mais également le champ croissant de données moins accessibles comme les images, les neurosciences, l'écologie, les données de cohorte, les données de biologie des systèmes et tous les autres types de données partageables.

Big Data : quelle vérité ? L'exemple suivant publié dans la revue Lancet (Bovet and Gedeon, 2013) à propos d'un article de J Salomon (Salomon et al., 2012) sur poids des maladies illustre le risque de se limiter à de larges bases de données sans se donner la peine de vérifier l'existence d'autres sources d'information ; L'étude de J Salomon et al s'intéressait à l'espérance de vie à naissance aux Seychelles entre 1990- 2010 et rapportait une valeur passant de 62.0 à 61.3 ans chez les hommes, passant de 72.4 à 71.8 ans chez les femmes, respectivement, alors qu'il existait un registre local ancien de qualité méconnu des bases OMS qui indiquait des valeurs passant de 63 à 69 ans chez les hommes, et passant de 74 à 79 ans chez les femmes, avec donc des conclusions tout à fait différentes quant à l'évolution de l'espèrance de vie. La standardisation indirecte utilisée par les auteurs (Wang et al., 2013) ne pouvait pallier à l'information manquante.

7 Big Data : vos données m'intéressent

Big data et vie privée : un paradigme nouveau ? (Schadt, 2012)

2011 a marqué le centenaire de la mort de Francis Galton, fondateur avec Karl Pearson du premier laboratoire d'eugénisme, plus tard pudiquement renommé Institut de Génétique, du fait de la connotation péjorative du premier vocable (Rose and Rose, 2011). Nous n'oublions pas son laboratoire d'anthropomètrie public où toutes sortes de mesures individuelles étaient

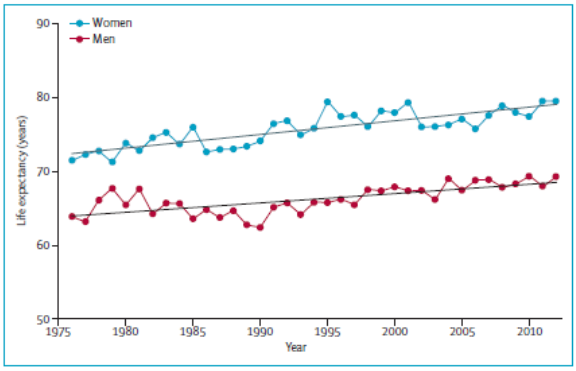


Figure: Life expectancy at birth in Seychelles between 1976 and 2012

FIG. 11 – Evolution de l'espérance de vie aux Seychelles 1990- 2010. Les insuffisances de l'exploitation des grandes bases. (Bovet and Gedeon, 2013)

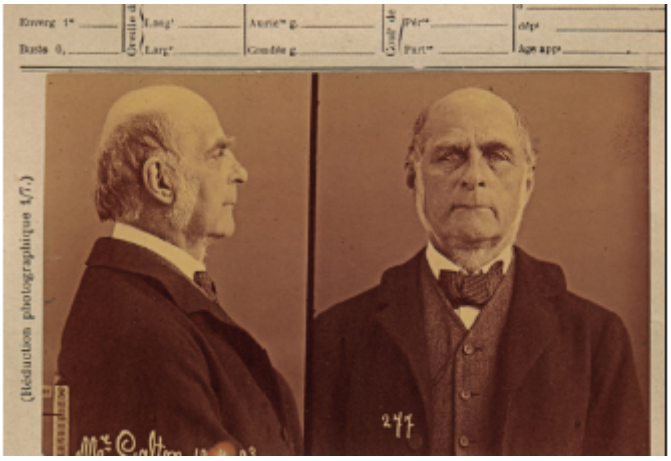


FIG. 12 – Galton et le spectre de la biométrie eugéniste ; Du génie héréditaire et de l'intelligence naturelle

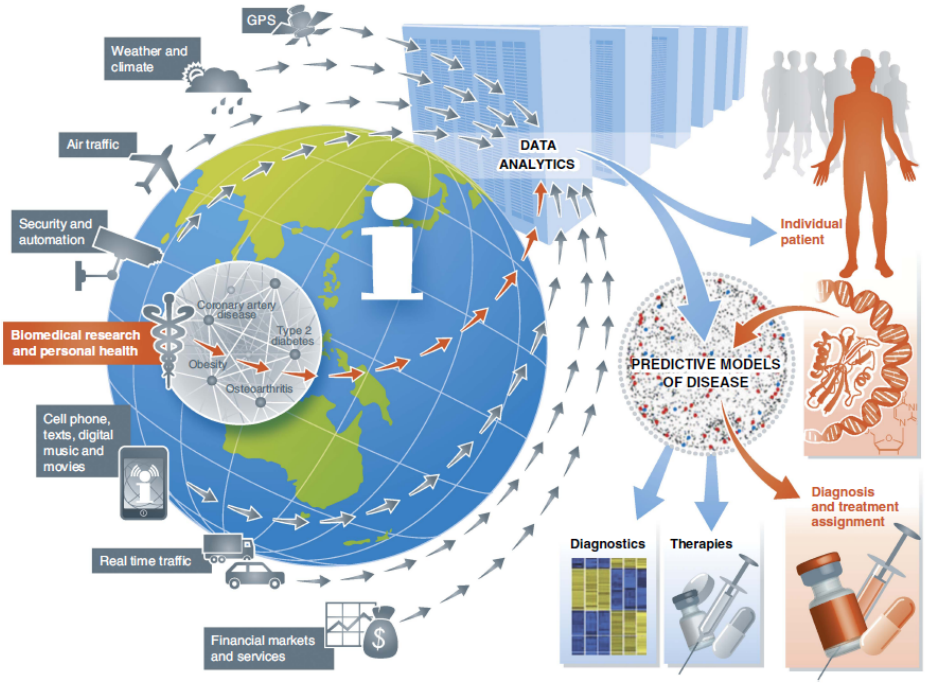


FIG. 13 – Big Data et santé : la circulation de l'information et ses retombées potentielles (Schadt, 2012)

collectées pour l'aider à classifier les capacités intellectuelles et confirmer son hypothèse de races inégales (*Du génie héréditaire et de l'intelligence naturelle*) (Ioannidis, 2013) De plus en plus de voix s'élèvent pour supprimer ou au moins alléger la notion de consentement à l'ère des "Big Data" sous le prétexte de faciliter des découvertes majeures à partir des bases de données disponibles. Ainsi, dans le cas de données non interventionnelles, l'Institut Médical Mondial préconise une telle attitude. Mark Rothstein et Abigail Shoben discutent cette attitude en insistant sur l'importance du biais de consentement et son influence sur les résultats. Le 20ième siècle a été marqué par un certains nombres de "ruses" de l'intelligence, lors de la phase d'intervention (plans expérimentaux) puis lors de la phase d'analyse (analyses multivariées par exemple) pour le prendre en compte, voire le corriger (analyse de sensibilité, pondération par l'inverse de la probabilité). Se passer des techniques d'intervention actives pour basculer dans l'observationnel *a posteriori* peut laisser penser qu'une étude non interventionnelle n'est pas susceptible d'influencer la vie privée et la dignité de la personne. Pourtant faire connaître des informations que la personne aurait préférée taire est tout sauf anodin. L'importance et le sens du biais de consentement ne sont pas toujours facilement appréhendables et de ce fait, sa correction reste délicate, si ce n'est impossible. La question est d'autant plus importante qu'il existe une tendance très forte pour ré- analyser des données déjà collectée pour d'autres études. La question du consentement se pose pourtant de la même façon avec des petites études type cas- témoins ou dans les très vastes jeux de données. Ces dernières ne sont pas de vraies recherches au sens noble du terme. Ionnadis parle d'Oxymoron, c'est à dire de recherches qui ne sont pas des recherches.

L'*open data* : ouvrir pour fournir un cadre à la réutilisation

C'est la thèse développée par Mr Xavier Crouan, responsable technique de la ville de Rennes, qui rapportait l'intérêt d'une appropriation citoyenne des jeux de données publiques

rendus disponibles pour une intelligence commune créative et partagée. Ceci l'amenait à proposer l'ouverture de concours pour des applications intelligentes à partir de ces données. Ce qui peut être vrai dans le cadre d'un projet d'aménagement urbain citoyen, est-il transposable au domaine de la santé ? C'est ce qu'a proposé une petite start-up *Fourmi Santé* en utilisant les données publiées par l'assurance-maladie pour proposer un comparateur des tarifs pratiqués par les médecins ? Le service a séduit le jury du premier concours DataConnexions organisé par Etalab, mais beaucoup moins l'assurance-maladie qui n'a pas apprécié cette réutilisation non-autorisée des informations proposées pourtant sur le site Ameli-Direct. Ce récent conflit vient s'ajouter à une longue liste de confrontations liées aux données "fermées". Comment peut-on les analyser ? L'open data est-il le problème ou bien une partie de la solution ? L'ancien président du Comité National d'Ethique, le Pr. Didier Sicard, signait une tribune dans le Monde en 2013 avec le Pr. Jean de Kervasdoué, ancien Directeur des Hôpitaux et professeur au CNAM, intitulée : "Plus grave que la pilule, l'affaire des données de santé publique" (de Kervasdoué and Sicard, 2013), dénonçant les lourdeurs alléguées par ces deux auteurs pour accéder aux données de l'assurance maladie. En ce sens, ils rejoignaient un propos tenus par un statisticien, le Pr. A Bar- Hen, et un enseignant de Santé Publique, le Pr. A Flahault, dans ce même quotidien (Bar-Hen and Flahault, 2013) intitulé "Donnons aux citoyens accès aux données de santé"

Personne n'a intérêt à la réutilisation non-autorisée des données, même pas le développeur. En procédant hors d'un cadre clair, il doit faire face à une incertitude juridique qui le pénalise. Il prend le risque que son service soit interrompu du jour au lendemain. Le détenteur de données a lui aussi intérêt à préciser le cadre juridique, technique et économique de réutilisation des données. La riposte juridique (les mises en demeure) fonctionne dans un premier temps, comme en témoigne la prudence affichée par les ré-utilisateurs concernés, mais *in fine* cela ne saurait constituer une politique en matière de diffusion et de valorisation des données. Bien plus, pour certains, le fait de détenir des données qui ne font pas l'objet d'une ré- utilisation sans accord doit faire réfléchir rapidement à la politique *open data* suivie.

Big Data : surveiller

Nous avons tous noté les récentes révélations concernant les pratiques de la National Security Agency (NSA), qui collecte directement toutes les données qui l'intéresse sur les serveurs américains de l'Internet. Tous les fichiers peuvent être extraits qu'elle qu'en soit la nature. Les modalités de ces extractions de données s'intégrent notamment dans le programme PRISM qui englobe Microsoft, Yahoo, google, skype, AOL, Apple. Récemment, Facebook a déployé un logiciel de recherche interne Graph Search, permettant de dresser rapidement une liste de personnes répondant à des critères complexes sociaux variés

Ainsi, le Monde a pu identifier les membres français de Facebook salariés de la RATP, aimant le livre Mein Kampf, les hommes mariés qui se disent intéressés par les hommes, une liste de soldats français partis en Afghanistan, une partie de ces informations étant supposées être cachées. Ces informations sont désormais à portée de clic. Certes, les résultats obtenus ne prétendent pas à une grande précision, mais ont un pouvoir stigmatisant très élevé et mettent grandement en danger la vie privée. Dans une interview du Monde (11/07/2013), Erin Egan, responsable de la vie privée chez Facebook, insiste sur la responsabilité individuelle des informations mises sur le réseau, en suggérant que le seul effet de Graph Search est de faire re-sortir des informations anciennes enfouies mais ouvertes. Le logiciel de reconnaissance faciale est évoqué. Concernant les demandes de requêtes par le gouvernement américain, Erin Egan, responsable vie privée chez Facebook indique 9000 requêtes impliquant 19000 utilisateurs pour 2012.

Nous nous souvenons de ce film de Sydney Pottier (1975), "Les trois jours du condor",

avec R Redford et F Dunaway, où le personnage principal était employé par la CIA dans un département discret à lire et écouter les media étrangers pour traquer l'anomalie qui doit éveiller le soupçon, étrangement prémonitoire, si ce n'est mission bien classique d'une centrale de renseignements étatique. Que dire de l'étonnement apparent de l'utilisation des relevés d'écoutes téléphoniques (Verazon) permettant de déterminer lieu, date et durées des appels, armes banales de la guerre économique que se livrent les états et les grandes multinationales ?

Ne soyons pas naïfs : nous acceptons assez facilement qu'une grande surface proche utilise la liste de nos achats pour nous adresser des publicités adaptées à nos goûts. Ainsi, notre collègue Suisse rappelait en début de ce colloque comment un père avait appris indirectement la grossesse de sa fille via les annonces publicitaire adressées à son domicile par la chaîne Target, qui avait noté l'achat récent de tests de grossesse à son adresse. Dans le champ biologique, où le mot "génétique" est prompte déclencher des réactions suspicieuse, une information, pourtant non génétique, comme l'abondance en ARN permet d'inférer l'ADN correspondant et donc identifier un sujet. Dans le domaine de l'éducation, l'intérêt suscité par les MOOCs (Cursera, Harvard, etc...) tient, pour les employeurs ultérieurs, autant à la transmission d'un savoir qu'à la possibilité d'avoir une trace du comportement de l'étudiant au cours de son apprentissage, enregistré au même titre que les éléments du diplôme obtenu.

Les Open Data et le mouvement "Donate your data" `we.consent.org`

La réutilisation non autorisée des données se développe et fait l'objet de nombreux conflits. L'ouverture des données (open data) peut-elle être une réponse à ces confrontations entre détenteurs et ré-utilisateurs de données ? C'est en tous les cas l'impression que donne la directive européenne récente (Directive, 2013) qui modifie la directive 2003/98/UE concernant la réutilisation des informations du secteur public. Cette directive souligne, d'emblée, que "les documents produits par les organismes du secteur public des Etats membres constituent une réserve de ressources vaste, diversifiée et précieuse, dont peut bénéficier **l'économie de la connaissance**.

Elle concerne plus directement les données présentes dans les bibliothèques publiques, dont les bibliothèques universitaires, les musées et les archives. Cette directive souligne l'augmentation exponentielle des données du secteur public depuis la directive de 2003, parallèlement à une constante évolution des technologies d'analyse, d'exploitation et de traitement. Le terme **métadonnées** apparait plusieurs fois dans la directive

Cloud computing et espionnage industriel

(Editor, 2013) "Les protestations de Mme Reding sont à mettre en écho avec les pressions qu'elle a du accepter des géants de la Silicon Valley pour diluer la législation européenne sur la protection des données privées. ...Intégré à nos vies, l'univers numérique expose en permanence à l'ingérence des gouvernants et des géants de l'internet. Visiblement les agences américains en profitent au maximum. On aimerait être sûrs que nos services secrets n'en font pas de même. La disparition de la notion de territoire d'application de règles nationales ! Les entreprises américaines ne veulent pas se voir imposer des règles supplémentaires sur le territoire européen."

Le Programme américain de surveillance PRISM et sa déclinaison britannique TEMPORA ont permis à la NSA la mise au point d'un programme de duplication des données du net à des fins d'utilisation rétroactive (site de Bluffdale, Utahet) avec arsenal légal associé du cloud : la loi *Foreign Intelligence Surveillance Amendements Act (FISAA)* (2008) permet de "siphonner" les données du net, avec une portée extra- territoriale valable jusqu'en 2012 prolongée de 5 ans. Elle permet d'agir sur les données de citoyens non américains

(les citoyens américains sont eux protégés parle 4ième amendement de la constitution). Les derniers amendements FISAA (1881-a) étendent la couverture de la surveillance à toutes les données sur le cloud. Ces données, obtenues par Amazon, Google, Microsoft, Apple, toutes compagnies américaines, sont fournies à la NSA à sa demande via des systèmes de scanner. Or, dans le même temps, un accord avec la CE avait fait de ces centres de cloud des "safe harbors" (zones sûres) bénéficiant de transferts de données sans l'autorisation de la CNIL ou équivalents européens.

Ainsi, chaque mois, la NSA "espionnerait" 500 millions de communications (téléphone, courriels, sms...)venant d'Allemagne, 8 millions de France, 4 millions d'Italie.... Le réseau américain Echelon existe depuis 1970. Le logiciel Eagle (Amesys, filiale de Bull) permet le Deep Packet Inspection (DPI) qui consiste à récupérer et trier tout ce qui passe par un point du réseau (c'est ce logiciel qui fit l'objet d'une vente au gouvernement Lybien de M Khadafi).

L'activité de renseignement et surveillance reste pourtante très officielle et active : le salon Milipol (Salon international de la sécurité intérieure des états) se tiendra à Paris en 2013, au Qatar en 2014, avec des sponsors tout à fait officiels (Renault Trucks Defense) Parmi les domaines d'activités couverts, nous retrouvons le traitement des fichiers et base de données

- Brouilleur de GSM
- Cryptographie
- Détecteurs et capteurs divers
- Détection d'écoute
- GSM
- Reconnaissance vocale
- Surveillance électro-acoustique
- Sécurité Internet
- Traitement des fichiers et bases de données
- Emetteur-récepteur-émetteur-récepteur

Pourtant, la réutilisation non-autorisée de données n'est pas une invention de l'ère Internet. Les premiers services d'annuaire inversés proposés sur Minitel étaient déjà basés sur une réutilisation non-autorisée des données publiées par France Telecom.

Aujourd'hui, ce ne sont pas seulement la liste des abonnés au téléphone que l'on peut retrouver sur Internet, mais la plupart des services et administrations publics : localisation et horaires des équipements, informations détaillées sur les transports et leur qualité ... Ce qui demandait, à l'époque du Minitel, une batterie de serveurs, est aujourd'hui accessible à n'importe quel individu équipé de logiciels disponibles dans le commerce. La "barrière à l'entrée" pour la collecte non-autorisée de données s'est donc très largement abaissée (Campbell, 2013). Dans un éditorial récent du quotidien Libération, N. Demorand parlait de Panopticon :

Panopticon, (Demorand, 2013) *"La sphère privée n'existe plus....Rien de ce qui est montré, dit ou écrit sur Internet n'échappe à ce Panopticon électronique qui scanne, copie et stocke tous les échanges. Que cette masse inimaginable de données puisse ou non être exploitée finit par sembler mineur au regard de l'infraction brutale, durable, sans doute même irréversible contre ce qui fonde toute démocratie : le respect de la vie privée. ...[Le vieux continent] dèjà incapable de collecter l'impôt des géants américains qui commercent sur leur sol et ouvrant grands leurs serveurs aux agences de renseignements de leur pays...comment même songer qu'is puissent créer, appliquer et défendre un Habeas Corpus numérique. ....La souveraineté numérique fonde désormais une part essentielle de la puissance d'un état et s'impose de fait comme le sujet éthique, politique, économique, international majeur de l'époque."*

8 Quelques éléments de conclusion

Nous voudrions apporter quelques pistes de réflexion autour de la notion de progrès, autour du renouveau du dialogue soignant/ soigné et sur les trois piliers d'une recherche fructueuse en santé qui restent : savoir utiliser, savoir imaginer, savoir réaliser, dans le respect de la volonté des participants volontaires, des règles et conventions sociales, en laissant des traces ouvertes aux autres (des cahiers ouverts)

Sur le progrès

L'accumulation de données (open ou big data) n'est-elle que l'illusion d'une source de connaissances ? Nous l'avons dit, les données ne sont jamais neutres. Les nouvelles technologies ne se substituent pas simplement aux anciennes : elles témoignent d'un monde nouveau et de nouveaux modes de communauté qui rejettent le monde ancien plus ou moins brutalement. Le monde futur n'est pas le monde ancien amélioré. Il est autre, avec de nouveaux modes de penser/ vivre, qui sont contenus dans les fonctionnalités de ses technologies . L'économie des nouvelles technologies est contenue dans leurs conception et leurs modalités de diffusion, sans distanciation possible. La notion de progrès reste relative.

L'apparition de nouveaux modes d'apprentissage d'un savoir interroge sur le sens d'un terme comme culture et le rôle dans le développement d'un individu des processus de mémorisation. Acquérir un savoir ne passe plus forcèment par une reconstruction selon sa propre logique d'une chronologie de savoirs. Les modalités d'échange entre les humains évoluent. La sophistication des modes d'extraction, l'augmentation des données disponibles renvoient à la naturalisation de notre vision du monde et de nos décisions à travers le prisme de ces instruments. Les pratiques professionnelles changent profondément : l'avalanche d'information remplace le temps de réflexion avec **une perte de l'épaisseur du temps**

Dans un entretien paru en mai 2012 dans la revue Sciences et Avenir (Wismann and Kahn, 2012), le philosophe H Wismann et le généticien A Kahn discutaient la notion de progrès et son lien avec l'humanisme. Ils notaient que la notion de progrès n'apparaissait qu'avec la Renaissance en Europe et la possibilité d'un futur pour l'homme et non une simple et permanente révolution (Copernic, découverte du nouveau monde,...). A Kahn cite Pascal "Toute la suite des hommes, pendant le cours de tant de siècles, doit être considérée comme un seul homme qui subsiste toujours et qui apprend continuellement". Il pointe, avec justesse, l'opposition, toujours présente, entre partisans de Socrate, pour qui le Vrai, qui caractérise le savoir, conduit au Bien et seul celui qui est dans l'ignorance peut s'engager dans la voie du Mal, et les partisans de Protagoras, figure du sophiste, où la quête du Bien est séparée de celle du Vrai. Une innovation scientifique belle et brillante peut ne pas être bonne. Peut- on pour autant empêcher l'explorateur des frontières de les franchir ? Ainsi A Kahn souhaitait re-définir "le progrès comme la mobilisation de l'intelligence et de la créativité pour produire des connaissances et des techniques, et générer des richesses dans un but humaniste".

Sur le renouveau du dialogue soignant- soigné

La consultation est devenue un lieu de confrontation des savoirs. Les p atients patients s'organisent et proposent, que ce soit sous forme d'associations de patients et leurs familles (mais quel est le rôle éventuel en coulisses des Big Pharma ?), ou sous forme de réseaux internet organisés et puissants (initiative Patients Like Me & Value for Openess), association Europa Bona Dona en cancérologie. Les registres des essais en cours sont accessibles à tous `http://www.ClinicalTrials.gov`. Les dernières années nous ont appris que la vigilance exercée par la seule communauté des professionnels de la santé au sens large ne permettait pas

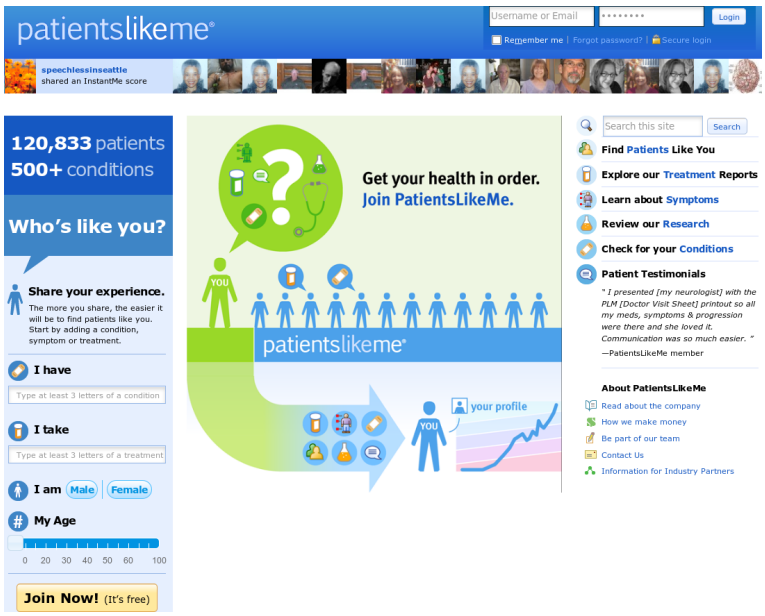


FIG. 14 – Google, internet et la santé : Patients Like Me

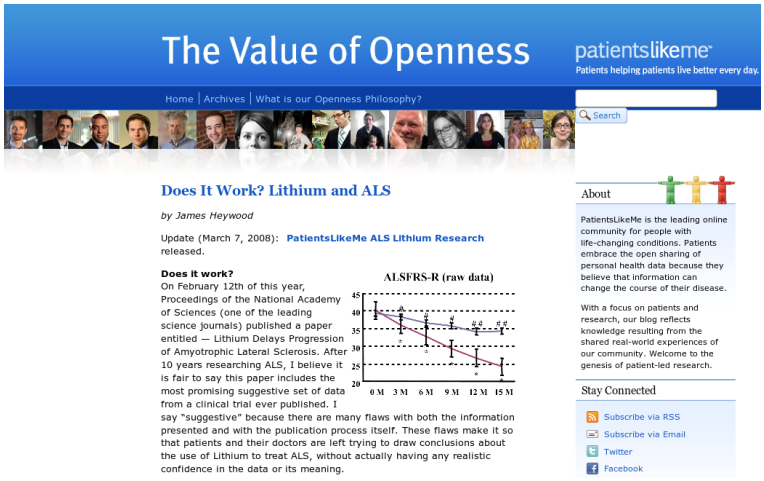


FIG. 15 – Google, internet et la santé : Value for Openess

d'éviter des phénomènes de *cécités collectives* (Médiator, contraceptifs oraux, GH, etc...).Le citoyen consommateur de soins se doit de participer à cette vigilance. Mais comment ?

La santé devient au coeur d'un marché prometteur

L'économie de ces technologies est contenue dans leur conception et leurs modalités de diffusion. Elle n'est pas la simple condition de leur apparition. La pensée économique n'offre guère d'espace critique pour qui aurait l'ambition d'observer de l'extérieur le rôle des technologies sur la transformation des manières d'être ensemble.

Il suffit de regarder des groupes d'adultes ou d'adolescents ensemble actuellement où le portable est devenu le *chapelet- komboloi*, les réunions sont devenues un mini- mur d'écrans, tablettes, branchés, où la pause n'est plus ce moment privilégié des échanges. Un staff médical se transforme souvent en un festival de connexions sur un iphone de tous les usages et toutes les distractions. Le sacerdose imposé par la lecture fastidieuse des *current contents* et la demande de tirés à part s'est relaché. La consultation bibliographique se fait à tout moment, bref regard qui accroche et stocke tout en même temps. Le temps du copiste soigneux en

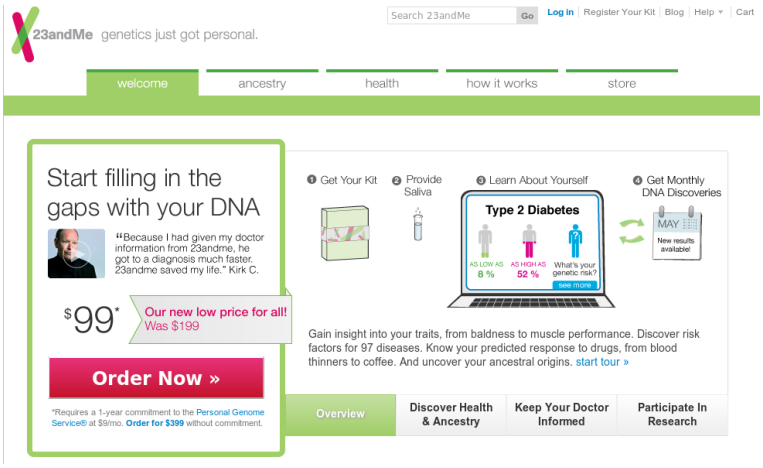


FIG. 16 – Google, internet et la santé : 23 and me

FIG. 17 – La santé devient une e-santé et enjeu de marchés prometteurs

salle de bibliothèque n'est plus remplacé par un autre. Acquérir un savoir ne passe plus forcèment par une reconstruction selon sa propre logique d'une chronologie de savoirs. A quoi bon savoir "Que j'aime à faire savoir un nombre utile aux sages. Immortel Archimède....", retenir une poésie ou une chanson. Le karaoke permet de passer des soirées agréables en chansons, moyennant une instrumentation. Comprendre une classe médicamenteuse n'a plus cette même nécessité.

Une illusion nous guette : la sophistication des modes d'extraction, l'augmentation des données disponibles et produites pourraient nous faire croire que les unes et les autres constituent des connaissances, naturalisant à la fois notre vision du monde et les décisions prises au nom des data.

Des pistes de réflexion s'ouvrent comme l'idée semi- utopiste d'impliquer des scientifiques citoyens dans des exploitations originales des Big Data. Individuellement, chacun ne pourrait pas forcément répondre à un appel d'offre, mais pourrait, par contre, répondre à des petits appels d'offre citoyen (environ 1000 euros), à travers des agences internationales, des fondations. Cette piste de micro- grants s'inspire du développement des micro- crédits (Özdemir et al., 2013) (DELSA Global, Seattle).

Penser/ créer pour redonner de l'épaisseur au temps

Nous voudrions terminer notre propos en ramenant une part de cette épaisseur qui caractérise l'acte de penser, en nous situant résolument dans le domaine de la recherche, permettant, au passage, de souligner combien les avancées des savoirs dans un domaine donné ont toujours imprégnés des concepts ambiants, aujourd'hui plus que jamais, témoignant de leur porosité. Ainsi la physiologie s'est -elle longtemps inspirée des développements de la physique électronique (circuits RLC et modélisation neuronale, modèles compartimentaux), le concept hormone- récepteur est né au moment des développements de l'information distribuée (réseau à jeton d'IBM), l'horloge hypothalamique, qui commande l'axe reproductif dans toutes les espèces, mise en évidence dans les années 1970, s'apparente aux tops d'horloge de la logique digitale, pour ne citer que quelques exemples.

Plus récemment la programmation nucléaire (le terme programmation, utilisé à dessein est pathognomonique du monde contemporain), a été totalement repensée avec les possibilités de clonage illustrées par la brebis Dolly, en hommage à la chanteuse de country Dolly Parton. Ces travaux s'inscrivaient dans la continuité de travaux de Gurdon sur les cellules souches embryonnaires pluripotentes (ESC) et ont ouvert les espoirs d'une médecine régénératrice. Ces espoirs se heurtaient à la notion bien établie d'une flêche du temps figeant les temps possibles de reprogrammation au stade embryonnaire, avec tous les débats de société autour de l'utilisation des embryons. La course au décryptage du génome humain et de son expression a amené des concepts et outils nouveaux (biologie des systèmes)

Yamanaka, prix Lasker 2011, a rcemment renversé ce point de vue sur la programmation nucléaire en s'appuyant sur ces outils et concepts. Il a pu mettre en évidence l'existence de cellules souches pluripotentes (iPS) obtenues à partir de cellules de peau de souris adultes en montrant que la pluripotence était possible en transférant 4 gènes (Oct4, Sox2, Klf4, c-Myc) parmi un pool de 24 gènes candidats, sélectionnés dans les banques existantes, renouvellant les rôles respectifs du conservateur et du lecteur dans une bibliothèque. Accueillis initialement avec scepticisme, ces résultats sont en voie d'applications avec, par exemple, la correction des anomalies de l'hèmoglobine dans la drépanocytose (iPS provenant d'une biopsie de peau) J Goldstein, en présentant les travaux de Yamanaka, soulignait les liens étroits entre innovation scientifique et art, la technique jouant en permanence un rôle d'aiguillon, incapable par elle-même de prévoir les retombées de ses développements futurs

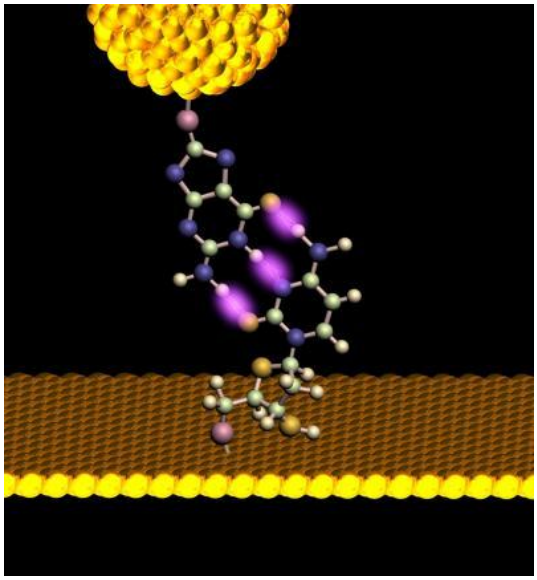


FIG. 18 – Science Daily, 2009. Microscope à effet tunnel et lecture des paires A-T ou C-G. IBM et le financement de l'innovation

FIG. 19 – L'innovation en peinture à travers la représentation papale ou l'art du renversement. De gauche à droite : Raphaël 1511, Titian 1543, Velazquez 1650, Bacon 1953, Brown 2008. (Goldstein, 2009)

Références

Foldit : solve puzzles for science, 2008.

Laurent Alexandre. Le docteur et l'astrophysicien. *Le Monde*, 20977, Juin 2012.

Emmanuelle Amoros, Jean-Louis Martin, Mireille Chiron, and Bernard Laumon. Road crash casualties : characteristics of police injury severity misclassification. *J Trauma*, 62(2) :482–490, 2007a.

Emmanuelle Amoros, Jean-Louis Martin, and Bernard Laumon. Estimating non-fatal road casualties in a large french county, using the capture-recapture method. *Accid Anal Prev*, 39(3) :483–490, 2007b.

Emmanuelle Amoros, Jean-Louis Martin, Sylviane Lafont, and Bernard Laumon. Actual incidences of road casualties, and their injury severity, modelled from police and hospital data, france. *Eur J Public Health*, 18(4) :360–365, 2008.

Gabriel Antoniu, Alexandru Costan, Benoit da Mota, Bertrand Thirions, and Radu Tudoran. A-brain : using the cloud to understant the impact of genetic variability on the brain. *ERCIM News*, 89, 2012.

Avner Bar-Hen and Antoine Flahault. Donnons aux citoyens accès aux données de santé. *Le Mon*, (9) :8, Mars 2013.

D. J. Barker and C. Osmond. Low birth weight and hypertension. *BMJ*, 297(6641) :134–135, Jul 1988.

D. J. Barker, P. D. Winter, C. Osmond, B. Margetts, and S. J. Simmonds. Weight in infancy and death from ischaemic heart disease. *Lancet*, 2(8663) :577–580, Sep 1989.

D. J. Barker, C. N. Martyn, C. Osmond, C. N. Hales, and C. H. Fall. Growth in utero and serum cholesterol concentrations in adult life. *BMJ*, 307(6918) :1524–1527, Dec 1993.

Pascal Bovet and Jude Gedeon. Life expectancy in seychelles. *Lancet*, 382 :23, 2013.

Berthold Brecht. *La vie de Galilée*. 1990.

Duncan Campbell. Le royaume- uni, maître- espion. *Le Monde*, page 18, 02 Jul 2013.

Jiajia Chen, Fuliang Qian, Wenying Yan, and Bairong Shen. Translational biomedical informatics in the cloud : present and future. *Biomed Res Int*, 2013 :658925, 2013. doi : 10.1155/2013/658925. URL `http://dx.doi.org/10.1155/2013/658925`.

Jean de Kervasdoué and Didier Sicard. Plus grave que le débat sur la pilule : l'affaire des données de santé publique. *Le Monde*, page 18, 2013.

Patrice Degoulet. Réutilisation des données d'un système d'information clinique : principes et applications, 2013. URL `http://lertim.timone.univ-mrs.fr/Ecoles/infoSante/2013/supports_ppt/Mar%di_AM/Degoulet.Corte.130716.V2.pdf`.

K. Demissie, N. White, L. Joseph, and P. Ernst. Bayesian estimation of asthma prevalence, and comparison of exercise and questionnaire diagnostics in the absence of a gold standard. *Ann Epidemiol*, 8(3) :201–208, Apr 1998.

Nicolas Demorand. Panopticon. *Libération*, 9994, 02 Jul 2013.

Alain Desrosières. *La politique des grands nombres : histoire de la raison statistique*. La Découverte, 1993.

Directive. Directive 2013/98/ue du parlement européen et du conseil du 26/06/2013 modifiant la directive 2003/98/ue concernant la réutilisation des iinformation du secteur public. *Journal Officiel de l'Union européenne*, 2013.

Cory Doctorow. Big data : Welcome to the petacentre. *Nature*, 455(7209) :16–21, Sep 2008. doi : 10.1038/455016a. URL `http://dx.doi.org/10.1038/455016a`.

Aisling O' Driscoll and Roy D Sleator. Synthetic dna : The next generation of big data storage. *Bioengineered*, 4(3) :123–125, 2013.

Editor. L'oncle sam se comporte très, très mal. *Le Monde*, 21290 :1, 02 Jul 2013.

Khatib Firas, Frank Dimaio, Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popovic, Mariusz Jaskolski, and David Baker. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat Struct Mol Biol*, 18 :1175–1177, 2011.

Antoine Geissbuhler. Réutilisation des données cliniques pour la recherche et la santé publique : Opportunités et enjeux transnationaux, 2013. URL `http://lertim.timone.univ-mrs.fr/Ecoles/infoSante/2013/supports_ppt/Lun%di_matin/Antoine_reutilisationDonnees_enjeuxTransnationaux.pdf`.

Joseph L Goldstein. Lasker awards and papal portraiture : turning fields upside down. *Nat Med*, 15(10) :1137–1140, Oct 2009.

Simon I. Hay, Dylan B. George, Catherine L. Moyes, and John S. Brownstein. Big data opportunities for global infectious disease surveillance. *PLoS Med*, 10(4) :e1001413, 2013.

Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, and Seung Yon Rhee. Big data : The future of biocuration. *Nature*, 455(7209) :47–50, Sep 2008. doi : 10.1038/455047a. URL `http://dx.doi.org/10.1038/455047a`.

John P A Ioannidis. Informed consent, big data, and the oxymoron of research that is not research. *Am J Bioeth*, 13(4) :40–42, 2013. doi : 10.1080/15265161.2013.768864. URL `http://dx.doi.org/10.1080/15265161.2013.768864`.

L. Joseph, T. W. Gyorkos, and L. Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*, 141 (3) :263–272, Feb 1995.

David Larousserie. Des microfiches en saphir pour affronter l'éternité : une solution française promet un stockage de l'information sans perte pendant des millénaires. *Le Monde*, 2012.

B. Laumon. Epidemiologic research and road traffic accidentology in europe. *Rev Epidemiol Sante Publique*, 46(6) :509–521, Dec 1998.

Christina H Liu, Da-Yu Wu, and Jonathan D Pollock. Bioinformatic challenges of big data in non-coding rna research. *Front Genet*, 3 :178, 2012.

J-L. Martin, S. Lafont, M. Chiron, B. Gadegbeku, and B. Laumon. Differences between males and females in traffic accident risk in france. *Rev Epidemiol Sante Publique*, 52(4) : 357–367, Sep 2004.

Vural Özdemir, Kamal F Badr, Edward S Dove, Laszlo Endrenyi, Christy Jo Geraci, Peter J Hotez, Djims Milius, Maria Neves-Pereira, Tikki Pang, Charles N Rotimi, Ramzi Sabra, Christineh N Sarkissian, Sanjeeva Srivastava, Hesther Tims, Nathalie K Zgheib, and Ilona Kickbusch. Crowd-funded micro-grants for genomics and "big data" : an actionable idea connecting small (artisan) science, infrastructure science, and citizen philanthropy. *OMICS*, 17(4) :161–172, Apr 2013. doi : 10.1089/omi.2013.0034. URL `http://dx.doi.org/10.1089/omi.2013.0034`.

Georges Perec. *Penser/ Classer*. Hachette, 1985.

Hilary Rose and Steven Rose. The legacies of francis galton. *Lancet*, 377(9775) :1397, Apr 2011.

JA Salomon, H Wang, and MK Freeman. Health life expectancy for 187 countries : a systematic analysis for the global burden of diseases 2010. *Lancet*, 380 :2144–2162, 2012.

Eric E Schadt. The changing privacy landscape in the era of big data. *Mol Syst Biol*, 8 : 612, 2012. doi : 10.1038/msb.2012.47. URL `http://dx.doi.org/10.1038/msb.2012.47`.

Maida J Sewitch, Mengzhu Jiang, Lawrence Joseph, Robert J Hilsden, and Alain Bitton. Developing model-based algorithms to identify screening colonoscopies using administrative health databases. *BMC Med Inform Decis Mak*, 13 :45, 2013. doi : 10.1186/1472-6947-13-45. URL `http://dx.doi.org/10.1186/1472-6947-13-45`.

Lucia Sillig. Les conseils du docteur watson : l'ordinateur qui a appris le langage des humains, au point de les battre lorsdu jeu télévisé "jeopardy", se lance dans l'oncologie. *Le Monde/ Le Temps*, 2013.

Nate Silver. *The Signal and the Noise*. 2012.

Nassim Nicholas Taleb. *The Black Swan, The Impact of the Hyghly Improbable*. Les Belles Lettres, 2007.

Emmanuel Todd. *La Chute finale : Essai sur la décomposition de la sphère soviétique*. Robert Laffont, 1976.

Alain Viari. Big data en biologie. *Médecine/ Sciences*, 28 :1027–1028, 2012.

S. D. Walter, P. Macaskill, Sarah J Lord, and L. Irwig. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Stat Med*, 31(11-12) :1129–1138, May 2012.

Stephen D Walter, Corinne A Riddell, Tatiana Rabachini, Luisa L Villa, and Eduardo L Franco. Accuracy of p53 codon 72 polymorphism status determined by multiple laboratory methods : a latent class model analysis. *PLoS One*, 8(2) :e56430, 2013.

Haidong Wang, Joshua A Salomon, and Christopher J L Murray. Life expectancy in seychelles - authors' reply. *Lancet*, 382(9886) :23–24, Jul 2013.

Heinz Wismann and Axel Kahn. Repenser le progrès dans une perspective humaniste. *Sciences et Avenir*, pages 46–50, 2012.