Collège Louise Wegmann

Première scientifique 1

Année scolaire 2017-2018

LE BIG DATA ET LA MEDECINE

AL-ASSAAD Bassam

DADA Karim

HASSAN Rayan

SAAD Jad

Sommaire :

Introduction

Pour le cadre des TPE nous avons choisi de traiter le sujet du Big Data en médecine qui rentre dans le thème transports et transferts. Le Big Data désigne des ensembles de données si grandes qu'ils dépassent l'intuition el les capacités humaines d'analyse et même des outils informatiques classiques de gestions de base de données ou de l'information. L'analyse approfondie du Big Data peut nous permettre d'avancer dans de multiples domaines tels que l'astronomie et la médecine et facilitera le transfert de l'information commet les imageries médicales et les données des patients. Comme le dit un chercheur du Knight Cancer Institute Jeff Tyner : « C'est là que l'idée d'un biologiste travaillant avec des informaticiens est si importante ». En effet plusieurs géants de l'informatique comme Google, qui à partir de la rétine peut détecter les maladies cardiovasculaires et peut prévoir un cancer à partir d'un bracelet, se sont lancés dans le domaine de la médecine. On peut aussi prendre l'exemple de Microsoft et son projet Microsoft Hanover qu'in détaillera plus tard.

Le Big Data : un véritable allié pour l'avenir de la médecine ?

Dans un premier temps nous nous attarderons sur ce qu'est le Big Data avant de parler de son utilisation en médecine. Ensuite nous évoquerons le machine learning, l'apprentissage automatique, et son implication dans le domaine de la santé. Le développement du Big Data et du Machine Learning possède aussi plusieurs limites qui disparaitront avec le développement de nouvelles techniques.

LA PRESENTATION DU BIG DATA

A- Définition du Big Data

Le « Big Data » représente à la fois la production à croissance rapide de données et leurs traitements pour pouvoir extraire des informations cruciales, utilisées dans plusieurs domaines pour faciliter leurs analyses et interprétations.

Le « Big Data » est caractérisé par les « 5 V » :

1- Volume

Le volume, mesuré en bytes, est la propriété la plus importante et la plus caractéristique du Big Data. Chaque journée, à peu près 2.5 quintillion (10^{18}) de bytes de données sont produites. En effet, de nos jours, tous les domaines qu'ils soient à aspect scientifique ou commercial produisent des nombres immenses de données. Ce phénomène évolue d'autant plus facilement à cause du coût faible de stockage des données.

D'une autre part, plus la taille est élevée, plus les résultats de leurs analyses sont précises, des résultats très importants pour extraire l'information utile et la transformer en connaissance pour la branche scientifique tout comme pour les branches commerciales. Cette multitude de données représente le défi majeur à relevé pour les systèmes supportant le Big Data, alors le traitement du grand volume de données exige un traitement divisé sur plusieurs sites distants.

2- Vélocité

Cette propriété tient pour la grande allure à laquelle les données sont générées, stockées et analysées. On réfère à ce phénomène par « Flux de données » qui représente la façon dont le système d'information d'une entreprise est alimenté automatiquement et en permanence par un gros volume de données pouvant donner lieu à des utilisations dans différents domaines.

Il est même attendu que la majeure partie des données composant le « Big Data » soit collectée en temps réel. Autrement dit, la vitesse à laquelle elles seront collectées surpassera la vitesse avec laquelle on peut les produire artificiellement.

3- Variété

Les données composant le Big Data proviennent de différentes sources et se présentent ainsi sous différents formats. En effet, elles peuvent être structurées sous forme de PDF, JavaScript, C++ ou de tables Excel ou sous forme de textes, images ou vidéos. Les différentes données sont analysées grâce au Big Data pour extraire les informations utiles provenant de données hétérogènes.

Ainsi, construire des systèmes pouvant cohabiter les différents types de données est primordial pour tirer pleine puissance du Big Data. D'un autre côté, la variété peut aussi référer aux différents genres de ces systèmes, que ce soit en matière d'infrastructures matérielles, plateformes de traitement ou langages de programmation.

4- Véracité

Le Big Data prenant des formes différentes est victime a l'instabilité de ses manifestations et de son évolution continue, la correction, précision et qualité des données deviennent douteuses ce qui altère leur valeur. Les sources de ces données étant différentes ont des risques à cause des capteurs qui peuvent être défectueux et résultent en de fausses données se mêlant de manière indiscernable aux données correctes. Les systèmes de traitement doivent offrir la possibilité de paramétrer l'analyse de sorte à assurer une certaine qualité des résultats au détriment du temps de traitement.

5- Valeur

La valeur des données représente ce qu'elles peuvent apporter comme gain, à la fois à la communauté l'extraction de connaissances scientifiques sur les phénomènes du monde mais aussi aux secteurs de l'industrie et de l'entreprise de manière générale à travers l'étude effective des marchés.

D'un autre côté, les systèmes de traitement Big Data doivent offrir différents niveaux de performances puisque les données possèdent des valeurs différentes provenant d'une multitude de secteurs.

B- L'historique du Big Data

En 1961 Derek Price publie annonce que le nombre de nouvelles revues a augmenté de façon exponentielle et non linéaire, doublant tous les quinze ans.

Tandis qu'en 1981 le Centre des statistiques en Hongrie lance un projet de recherche pour tenir compte des industries de l'information du pays, y compris la mesure du volume d'informations. Les recherches se poursuivent à ce jour.

Dans l'année 1996, IBM annonce que le stockage numérique est devenu plus rentable que le stockage des données en papiers. Deux ans plus tard, le trafic de données dépasse le trafic téléphonique

En 1999 Steve Bryson, David Kenwright, Michael Cox, David Ellsworth et Robert Haimes publient «Visually Exploring Gigabyte Datasets in Real Time». C'est le premier article qui emploie le terme « Big Data ».

Enfin, en 2000, des études de l'université d'UC Berkeley quantifient, en termes de stockage informatique, le montant total des informations nouvelles et originales (sans compter les copies) créé dans le monde chaque année et stocké dans quatre milieux physiques: papier, film, optique et magnétiques. Ces études concluent qu'une grande quantité d'informations uniques est créée et stockée par des individus qualifiant cette conclusion par une domination du numérique.

C- Comment enregistrer les données

1- L'origine des données

Les données utilisées en analyse dans la majorité des secteurs sont contenus générés par machine ainsi qu'à travers de contenus générés par les utilisateurs.

i- Sensorique

D'importantes quantités de données sont créées grâces aux capteurs et détecteurs situes dans tous les appareils technologiques comme par exemple les services de localisation avec GPS (Géo-positionnement par satellite) qui génèrent des informations sur la géographie de la terre provenant de millions d'utilisateurs. Dans le domaine des transports, les nouveaux véhicules sont connectés à Internet et peuvent transférer des données d'état au constructeur, ce qui permet de mettre au point de nouveaux modèles commerciaux pour satisfaire les besoins des consommateurs. De plus, les capteurs situés dans la turbine d'un Airbus A380, par exemple, transmet à elle seule 20 Téraoctets de données par heure, qui peuvent être utilisées pour une analyse et une surveillance système techniques.

ii- Données de connexion et de médias sociaux

Les données touchant au comportement de l'utilisateur sont dorénavant saisies de manière très détaillée par les systèmes de Web-Tracking dans les Sites Web et surtout dans les boutiques de commerce électronique. D'une autre part, le Web social est considérée être comme source de Big Data, Il est donc évident que de telles entreprises Internet comme Facebook et Twitter, utilisent et fournissent des données énormes intéressantes et utiles dans le Big Data.

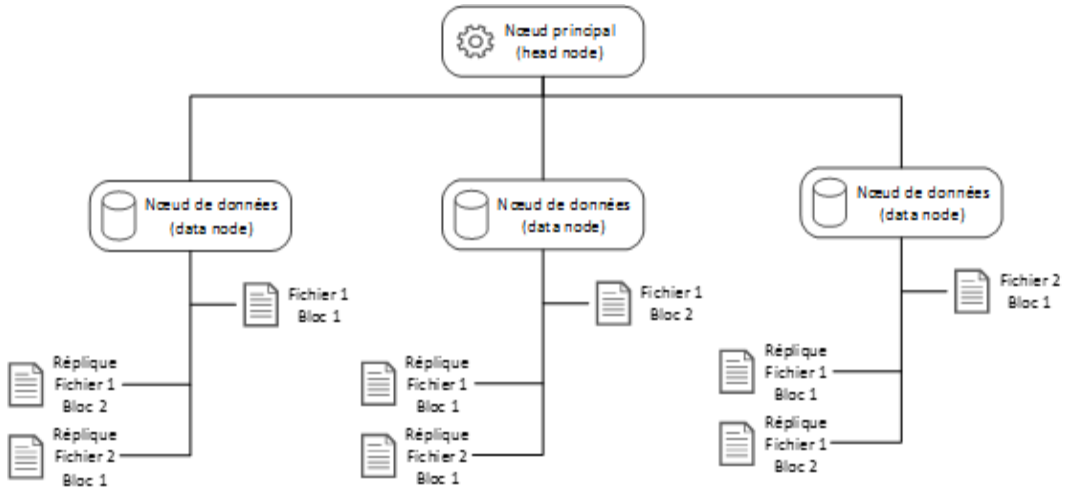
2- Les technologies qui vont aboutir au grand stockage d'informations

De nombreuses techniques sont actuellement en concurrence sur le marché pour l'enregistrement du Big Data, comme par exemple le Hadoop File System (HDFS) ou les bases de données NoSQL. Mais dans la pratique, les architectures d'entrepôt de données classiques dans lesquelles les entreprises peuvent utiliser des données structurées

dominent jusqu'à présent. Ces dernières se voient cependant confrontées au défi de devoir traiter des interrogations de bases de données de plus en plus nombreuses et complexes, d'assurer l'actualité des données et de les rendre plus rapidement disponibles.

i- Hadoop :

Son architecture générale consiste en un nœud central de contrôle appelé NameNode et plusieurs autres nœuds qui stockent les données appelés DataNodes. Le rôle du NameNode est de détenir les données et de les partager et les fragmenter sur les nœuds de stockage qui les sauvegardent.



ii- Bases de données NoSQL

C'est une approche de la conception des bases et de leur administration particulièrement utile pour de très grands ensembles de données distribuées. NoSQL englobe des technologies et d'architectures, afin de résoudre les problèmes de performances en matière d'évolutivité et de Big Data NoSQL est particulièrement utile lorsqu'une entreprise doit accéder, à des fins d'analyse, à de grandes quantités de données non structurées ou de données stockées à distance sur plusieurs serveurs virtuels du Cloud.

Or les 3 propriétés fondamentales pour les systèmes distribués sont : la Consistance lorsque tous les nœuds du système voient exactement les mêmes données au même moment, la Disponibilité ou la perte de nœuds n'empêche pas les survivants de continuer à fonctionner correctement alors les données restent accessibles et enfin la Résistance au partitionnement puisque le système étant partitionné, aucune panne

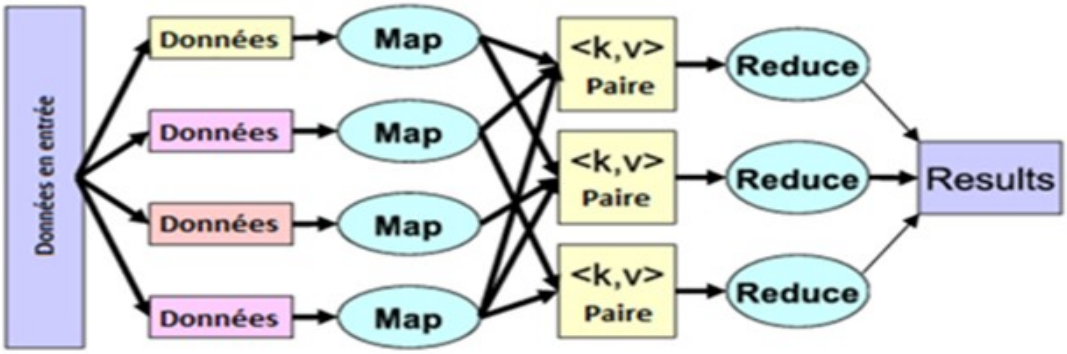
moins importante qu'une coupure totale du réseau ne doit l'empêcher de répondre correctement.

En effet, d'après le Théorème de CAP, dans un système distribué, il est impossible d'obtenir ces 3 propriétés en même temps, il faut en choisir 2 parmi les 3. NoSQL emploie les propriétés de la Disponibilité et de la Résistance et cela pour pouvoir enregistrer des nombres de données immenses sans prendre compte des pertes et des pannes et cela assure le fonctionnement du Big Data.

D- Comment traiter les informations

1- L'algorithme Map-Reduce

Map-reduce est un modèle de programmation massivement parallèle adapté au traitement de très grandes quantités de données. Les programmes adoptant ce modèle sont automatiquement parallélisés et exécutés sur des clusters d'ordinateurs.



Les deux étapes principales sont :

- Map: Emission de paires <clé,valeur> pour chaque donnée entrée
- Reduce: Regroupement des valeurs de clé identique et application d'un traitement sur ces valeurs de clé commune

Un système très utilisées en Big Data grâce à l'algorithme Map-Reduce, le PageRank, qui est un système de distribution de probabilité sur les pages web qui représente la chance qu'un utilisateur naviguant au hasard ouvre une page $P_{initiale}$ située entre P_1 et P_n appartenant à M, l'ensemble des pages ayant un lien vers $P_{initiale}$. Elle permet de voir aussi la probabilité que cet utilisateur naviguant l'Internet arrive à accéder une autre Page Web appartenant à M en cliquant sur un des liens L appartenant à $L(P_{initiale})$ qui est le nombre de liens sortant de la Page $P_{initiale}$.

Cette formule aidera donc à voir les sites Web les plus utilisés et pouvoir les classes à l'aide de la méthode « Mapping » employé par l'algorithme Map-Reduce.

$$PR(p_i) = \sum_{p_i \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

- $p_1, p_2, \ldots, p_N$ sont les pages web (les nœuds du graphe)
- $M(p_i)$ est l'ensemble des pages ayant un lien vers p_i
- $L(p_j)$ est le nombre de liens sortant de la page p_j
- N est le nombre total de pages web

Ces probabilités peuvent se retrouver dans un algorithme de Map-Reduce sous forme de :

Mapper:
Entrée:
clef: URL_{page}
valeur: «PR; [$URL_{lien,\ldots}$]»
Sortie:
Pour chaque URL_{lien}, émettre:
clef: URL_{lien}
valeur: «URL_{page}; PR, nb_url_{lien}»

Où: nb_url_{lien} est le compte de URL_{lien}

Reducer:

Entrée:

clef: URL_{page}

valeurs: [«$URL_{inverse}$;PR, $nb\_url_{page\_inverse}$», …]

Traitement: calculer le PR

Sortie:

clef: URL_{page}

valeurs: « PR;[URL_{lien}]»

Le concept du PageRank peut être appliqué dans diffèrent domaines en classant les résultats à l'aide de « Mapping », pour trouver des chances d'apparitions et de récurrence de quelques résultats à l'aide de la méthode « Reduce » dans cet algorithme.

2- Les supercalculateurs

Un supercalculateur est un très grand ordinateur, réunissant plusieurs dizaines de milliers de processeurs, et capable de réaliser un très grand nombre d'opérations de calcul ou de traitement de données simultanées. Les superordinateurs sont utilisés par les scientifiques et les industriels pour concevoir de nouveaux systèmes, des matériaux ou des médicaments. Il est capable de simuler des phénomènes physiques complexes comme les séismes et les formations des étoiles et des galaxies. Il peut aussi réaliser des prévisions en météorologie ou accomplir virtuellement des expériences difficilement réalisables en laboratoire.

Le Tianhe-2 est un supercalculateur situé, à l'université nationale de technologie de la défense à Guangzhou en Chine. C'est le supercalculateur le plus performant au monde à présent.

3- Les algorithmes

Le travail des algorithmes est de guidé le travail des supercalculateurs afin d'analyser les données et en tirer les informations nécessaires dans les secteurs différents. L'avancée des recherches sur les algorithmes a ainsi permis de faire émerger ce qu'on appelle l'apprentissage automatique qui désigne la capacité d'un algorithme à repérer des tendances ou des corrélations dans un très grand volume de données, en adaptant ses analyses et ses comportements. C'est notamment utilisé pour détecter des régularités dans le comportement des consommateurs et des indices économiques, c'est l'analyse prédictive

E- Le Machine Learning

1- Définition et utilisation

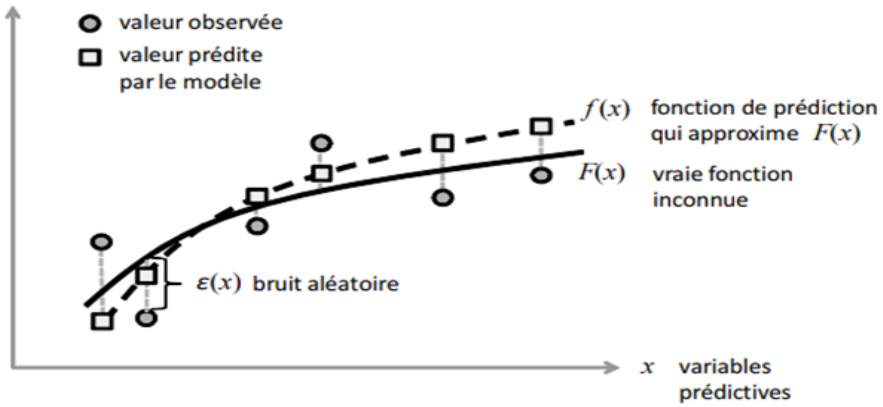
Le Machine Learning est une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Le Big Data est la base du Machine Learning qui est une technologie permet d'effectuer des prédictions à partir de données en se basant sur des statistiques, sur du forage de données et sur les analyses prédictives. Pour l'analyse de telles données, le Machine Learning se révèle nettement plus efficace que les méthodes traditionnelles en termes de précision et de vitesse.

Pour pouvoir fonctionner, l'algorithme du Machine Learning repose sur :

1. Une fonction F(x) des variables prédictives. C'est donc une contribution entièrement déterminée par les variables prédictives x de l'observation. C'est le signal que l'on souhaite mettre en évidence.

2. Un bruit ε(x) aléatoire. C'est une fonction qui englobe les effets conjugués d'un grand nombre de paramètres dont il est impossible de tenir compte.

Aussi bien F que ε resteront à jamais inconnus mais l'objectif d'un modèle de machine Learning est d'obtenir une « bonne approximation » du signal F à partir d'un ensemble d'observations. Cette approximation sera notée f, on l'appelle la fonction de prédiction.



La valeur d'une variable cible est la somme d'une fonction déterministe F et d'un bruit aléatoire ε. L'objectif du ML est de trouver une bonne approximation de F à partir des observations. Cette approximation f est la fonction de prédiction qui permet d'obtenir des estimations f(x) de F(x).

En conclusion, un modèle de Machine Learning est un procédé algorithmique spécifique qui permet de construire une fonction de prédiction f à partir d'un jeu de données d'apprentissage. La construction de f constitue l'apprentissage ou l'entraînement du modèle. Une prédiction correspond à l'évaluation f(x) de la fonction de prédiction f sur les variables prédictives d'une observation x.

2- Lien avec le Big Data

Le Machine Learning est idéal pour exploiter les opportunités cachées du Big Data. Cette technologie permet d'extraire de la valeur en provenance de sources de données massives et variées sans avoir besoin de compter sur un humain. Elle est dirigée par les données, et convient à la complexité des immenses sources de données du Big Data.

Plus les données injectées à un système Machine Learning sont nombreuses, plus ce système peut apprendre et générer des résultats de meilleures qualités. Le Machine Learning permet ainsi de découvrir les phénomènes périodiques enfouies dans les données avec plus d'efficacité que l'intelligence humaine.

Les analyses prédictives consistent à utiliser les données, les algorithmes statistiques et les techniques de Machine Learning pour prédire les probabilités de tendances et de résultats financiers des entreprises, en se basant sur le passé. Selon une étude, 75% des entreprises qui augmentent leurs investissements dans les technologies analytiques en tirent profit.

Toutes les entreprises accumulent au fil du temps de grandes quantités de données qui demeurent inutilisées. Il s'agit des dark data. Grâce au Machine Learning et aux différents algorithmes, il est possible de faire le tri parmi ces différents types de données stockées sur les serveurs.

Enfin, l'apprentissage des données permet d'organiser le stockage de données pour un meilleur accès. Au cours des cinq dernières années, les vendeurs de solutions de stockage de données ont mis leurs efforts dans l'automatisation de la gestion de stockage.

LE BIG DATA UN ALLIÉ POUR LA MEDECINE

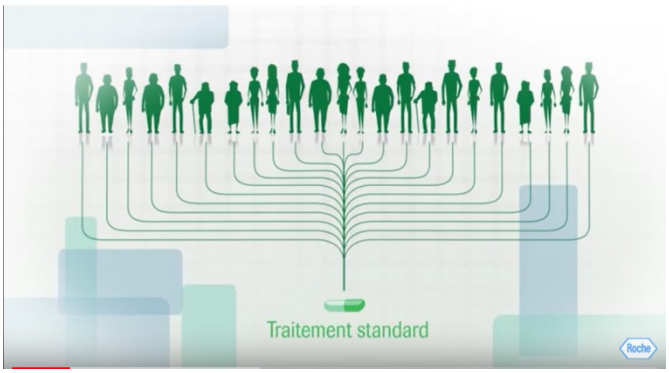
A-Médecine personnalisée

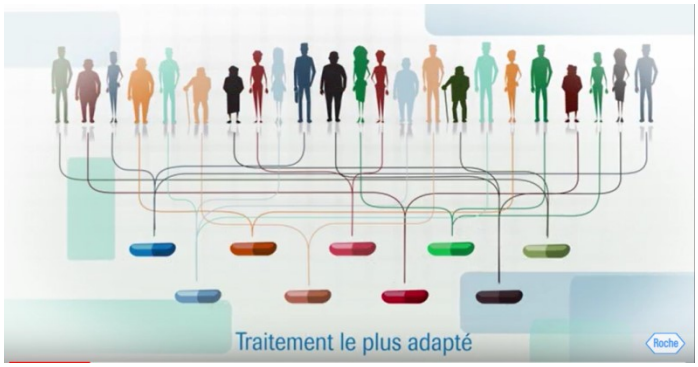
La médecine personnalisée est aussi une médecine prédictive, préventive et participative, d'où vient son nom : « la médecine P4 ».

1-Domaine de consultation médicale

a-La médecine stratifiée

De nos jours, les médecins prescrivent pour une même maladie un traitement standard. Mais les patients possèdent des patrimoines génétiques très différents et on retrouve pour chaque maladie plusieurs sous-types caractérisés chacune par des propriétés moléculaires différentes. Pour un même traitement seulement la moitié en tire bénéfice d'après des recherches menées par Roche, l'une des premières entreprises pharmaceutiques du monde, elle est basée en Suisse et comporte le secteur pharmaceutique ainsi que le secteur diagnostic. La médecine personnalisée consiste à choisir le traitement le plus adapté en fonction du patrimoine génétique du patient et en fonction de la maladie et ceci présente plusieurs avantages. En effet, la médecine personnalisée permet l'augmentation de l'efficacité et la tolérance d'un traitement. En observant les effets macroscopiques du médicament sur le patient on peut aussi varier la dose du médicament pour éviter les effets secondaires comme les allergies, phénomène qui apparait lorsqu'on fait face à un corps étranger à notre organisme. La médecine stratifiée fait partie de la médecine personnalisée puisqu'elle consiste à traiter le patient ayant recours à





ses données biologiques. Par exemple, en regroupant les données biologiques du patient ainsi que ses données familiales, environnementales avec les données cliniques et les recherches pour pouvoir détecter les maladies chez un patient.

Mais les données biologiques et familiales d'un patient sont gigantesques et leur analyse n'est pas possible avec les technologies utilisées alors on doit avoir recours au Big Data qui va permettre l'analyse de ces données pour pouvoir détecter la maladie. Pour mettre en exemple l'usage du Big Data en médecine on peut prendre l'exemple suivant : un patient rentre chez son médecin et lui fait part de ce qu'il ressent. Le médecin entrera alors ces données sur son ordinateur qui les analysera avec les données biologiques, familiales et environnementales du patient. Si le patient possède un allèle morbide qui est aussi présent chez plusieurs membres de sa famille et ressent les symptômes de cette maladie ou bien s'il vit dans des conditions favorables à l'apparition de cette maladie alors le médecin pourra tout de suite trouver de quoi souffre le patient.

Un chercheur de l'institut Knight, Jeff Tyner dit : « C'est là que l'idée d'un biologiste travaillant avec des informaticiens est si importante ». En effet on peut prendre le cas du projet Microsoft Hanover qui utilise les technologies Machine Learning pour le traitement de précision du cancer en collaboration avec le Knight Cancer Institute. Cette collaboration a commencé lorsqu'un cancérologue (médecin qui traite les personnes atteints de cancer) du Knight Cancer Institute a dit que chaque tumeur (cellule cancéreuse) est différente de l'autre et qu'il faudrait avoir recours à une médecine personnalisée. Un chercheur de Microsoft, Holfang Poon a mis en place un algorithme capable de trouver un traitement à l'aide du Machine Learning. Cet algorithme renferme dans une « librairie » (ensemble de documents qui peut être introduit dans l'algorithme pour qu'il soit analysé) un ensemble de recherches faites sur le cancer et les tumeurs. On introduit donc dans cet algorithme cette librairie et des données et des informations concernant la tumeur du patient qu'on veut traiter. L'algorithme rentre alors dans la phase « Map » où il analyse les données pour pouvoir ensuite passer dans la phase « Reduce » où l'algorithme sélectionnera les documents en relation avec le cas étudié. A l'aide de ces documents, les médecins pourraient éventuellement trouver un traitement efficace.

En analysant aussi ces données, le médecin va aussi pouvoir prescrire au patient un médicament qui sera toléré par le malade et qui aura pour but de guérir son cas, alors le médecin lui donnera un traitement « sur mesure ».

b-La surveillance à distance :

La médecine personnalisée consiste aussi à avoir recours à des moyens de modélisation individuelle. Avec le développement de l'analyse du Big Data dans

l'apprentissage automatique l'utilisation de micro biocapteurs et de dispositifs capables de surveiller la santé du patient est possible et en cas de problème ces dispositifs enverront une alerte au médecin traitant. En effet, ces dispositifs qui représentent des moyens de modélisation individuelle vont analyser les paramètres biologiques du patient et les transmettre au médecin. Ce type de surveillance présente des avantages sur le plan médical puisque le médecin va pouvoir consulter chaque patient et ses données à tout moment. En effet, le tensiomètre (pour mesure la tension) connecté envoie les données d'un patient au médecin à travers une application (IHealth) sur le téléphone.

Alors on peut en déduire que le Big Data permet le développement de la médecine personnalisée en permettant la médecine stratifiée avec l'analyse des séquences génomiques mais aide aussi dans le domaine de la surveillance à distance avec l'usage de l'analyse du Big Data en apprentissage automatique dans les micros biocapteurs.

2-Domaine pharmaceutique

Comme le dit les laboratoires Roche la médecine personnalisée doit avoir recours à deux équipes : l'équipe diagnostic et l'équipe pharmaceutique. Cette compagnie nous détaille dans une vidéo les étapes de développement d'un médicament. On commence par sélectionner les molécules qui représentent le principe actif avant de les tester pour trouver le rapport efficacité/tolérance de ce médicament.

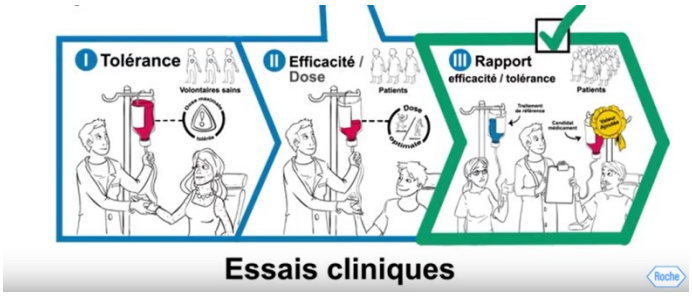
a-Trouver la molécule guérisseuse

De nos jours beaucoup de maladies restent mortelles puisqu'aucun remède n'a été trouvé. Avec le Big Data et l'apprentissage automatique de plus en plus de médicaments sont trouvés. Puisque leur utilisation dans la découverte de médicaments préliminaire (au stade précoce) peut être utilisée à diverses fins, depuis le criblage initiale de composés médicamenteux jusqu'au taux de succès prévu basé sur des facteurs biologiques. D'après un rapport qui date de 2015 par Pharmaceutical Research Manufacturers d'Amérique (la plus grande compagnie de commerce dans le domaine pharmaceutique aux Etats-Unis) plus de 800 médicaments contre le cancer sont en cours d'essais. On peut aussi prendre en exemple le MIT Clinical Machine Learning Group (MITCMLG) qui essaye à l'aide de l'apprentissage automatique de mieux comprendre les processus pathologiques et de concevoir des

traitements efficaces contre le diabète de type 2, qui touche généralement les personnes en surpoids. Pour guérir les patients les médecins prescrivent tout d'abord des médicaments accélérant le métabolisme pour que la personne perde du poids plus rapidement. Les personnes atteintes de cette maladie n'ont généralement pas le même poids. C'est pour cela que le MITCMLG a décidé à l'aide d'un algorithme comportant l'apprentissage automatique de déterminer la quantité nécessaire de la molécule constituant le principe actif, comme l'insuline, chargé de régler le métabolisme du patient selon ses besoins.

b-Trouver le médicament idéal pour le patient

La médecine personnalisée consiste à utiliser pour chaque patient un médicament dont le rapport efficacité/tolérance est le maximum, donc chaque patient possède son propre médicament. Le médicament est formé du principe actif (la molécule guérisseuse) et des excipients (molécules qui déterminent le gout et la couleur du médicament). Le médicament peut avoir des effets secondaires sur le patient si la quantité du principe actif est très élevée ou bien si le patient est allergique à certains des excipients. Donc pour avoir le meilleur rapport efficacité/tolérance il faut choisir une concentration de principe actif spécifique à chaque cas et lui ajouter des excipients qui n'auront pas d'effets secondaires sur le patient. Avec le Big Data même sans tests cliniques le médecin peut lui trouver son médicament à l'aide de ses données sauvegardées dans les hôpitaux et industries pharmaceutiques.



Alors on peut conclure que le Big Data présente plusieurs avantages dans la médecine P4 dans le domaine pharmaceutique :

- Avec un accès aux données des patients, une meilleure compréhension des effets des médicaments serait possible.
- Meilleur identification des cibles biologiques des molécules.
- Détermination des candidats des essais cliniques.

B-Médecine prédictive

La médecine prédictive est une branche de la médecine personnalisée qui désigne les capacités nouvelles de la médecine, notamment de prévoir, parfois très à l'avance les affections qui frapperont le patient. Son but est d'empêcher voire retarder l'émergence de certaines pathologies telles que l'hypercholestérolémie (excès de cholestérol, molécule dont l'excès peut causer des maladies). Pour arriver à ceci il faudra analyser le génome humain et le stocker pour pouvoir l'analyser avec les données familiales du patient ainsi que l'environnement dans lequel on vit.

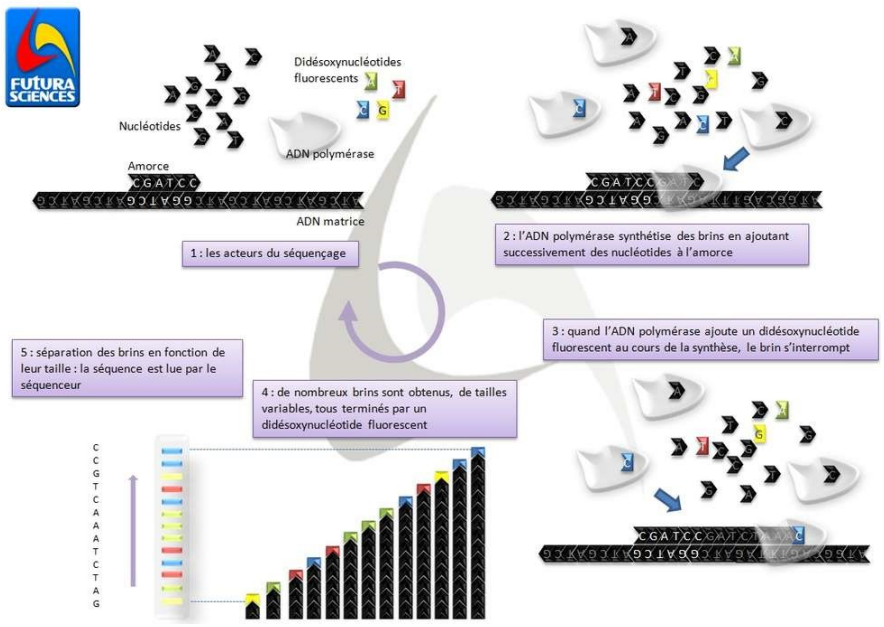
1-Analyse du génome humain

Notre génome est constitué de 23 paires de chromosomes, supports de notre information génétique. Deux grandes approches différentes permettent son analyse : la génétique moléculaire avec le séquençage qui permet une analyse ciblée des gènes et la cytogénétique avec le caryotype qui permet une analyse globale du nombre et de la structure des chromosomes. La cytogénétique est mise en place pour améliorer le niveau de résolution (détection d'anomalies sur une plus grande partie de l'ADN) du caryotype. En effet, le niveau de résolution des résultats de la cytogénétique est cinq fois plus élevé que celui du caryotype possédant le meilleur niveau de résolution. L'utilisation de la cytogénétique a permis le développement de nouvelles techniques permettant une analyse globale du génome avec un très grand niveau de résolution. Actuellement, la plus performante d'entre elles est l'hybridation génomique sur micro réseau ou *CGH array* dont les applications en diagnostic ont déjà débuté. Cependant, ces nouvelles méthodes nécessitent d'être encadrées, car elles peuvent poser des problèmes d'interprétation et soulever des questions d'ordre éthique.

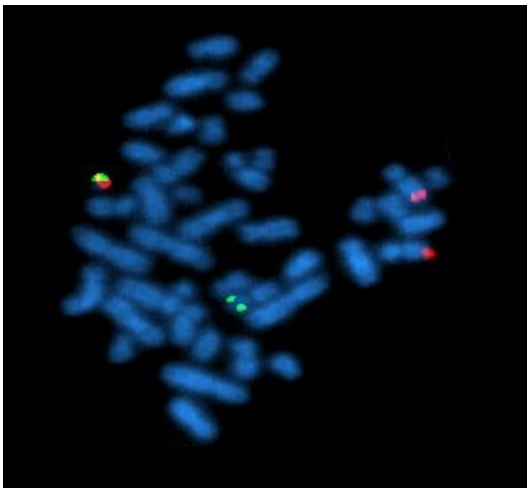
a-Méthodes d'analyse du génome humain :

- **La génétique moléculaire avec le séquençage** : La séquence de l'ADN constitue en quelque sorte l'anatomie d'un génome : elle indique les formules des protéines dont celui-ci dirige la synthèse. Sa connaissance est donc cruciale pour la biologie, mais les techniques nécessaires pour la déchiffrer ne sont apparues qu'assez récemment. D'abord très lentes et laborieuses, elles ont ensuite progressé au point qu'il a été possible de lire intégralement l'ADN. Alors que la méthode de Maxam et Gilbert (la première réalisée) consistait à utiliser les propriétés chimiques des nucléotides, la

méthode de Sanger, qui repose sur la biologie moléculaire, est désormais la plus utilisée. Bien que le séquençage ait beaucoup évolué et soit désormais automatisé, il repose généralement sur l'utilisation de composants biologiques qui existent naturellement dans les cellules. La méthode de Sanger repose sur plusieurs étapes. On commence par insérer en grande quantité dans un tube à essai de l'ADN, des nucléotides, de l'amorce et de l'ADN polymérase. La réaction de séquençage fait donc intervenir de multiples réactions. On utilise aléatoirement les nucléotides pour former une séquence d'ADN complémentaire à l'amorce. Si un didésoxynucléotide est choisi à la place d'un nucléotide la chaine s'interrompt et prend une couleur dépendant du dernier nucléotide. Toutes les chaines qui se terminent par le même nucléotide sont de même longueur puisque toutes sont faites à partir de l'amorce. On place ces fragments dans un gel d'acrylamide alors les chaines de même taille migrent à une même distance alors la suite des couleurs de fragments détermine la séquence nucléotidique de l'ADN.



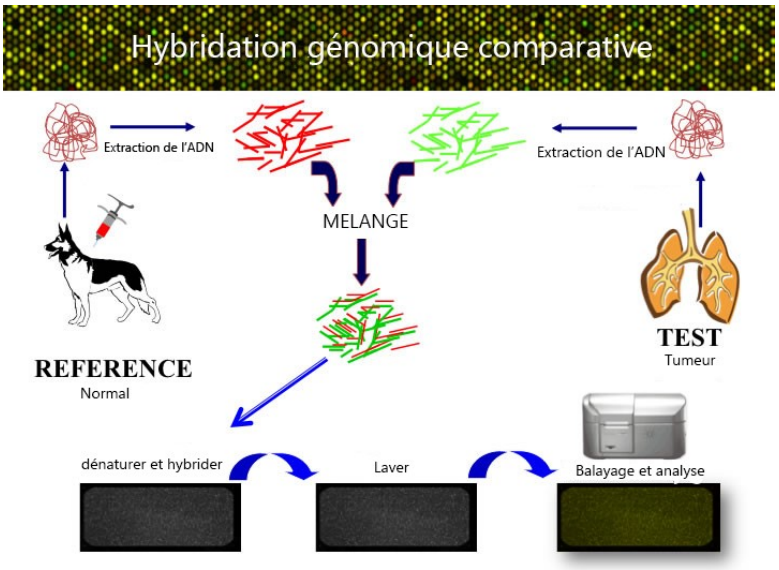
- **La cytogénétique avec le caryotype :** La cytogénétique est l'étude des phénomènes génétiques au niveau de la cellule. C'est une technique de marquage appelé aussi méthode FISH qui permet la détection des arrangements pathogènes (bactéries) sur les chromosomes à l'aide de sondes qui peuvent être- envoyer sur l'ADN, l'ARN ou les protéines.

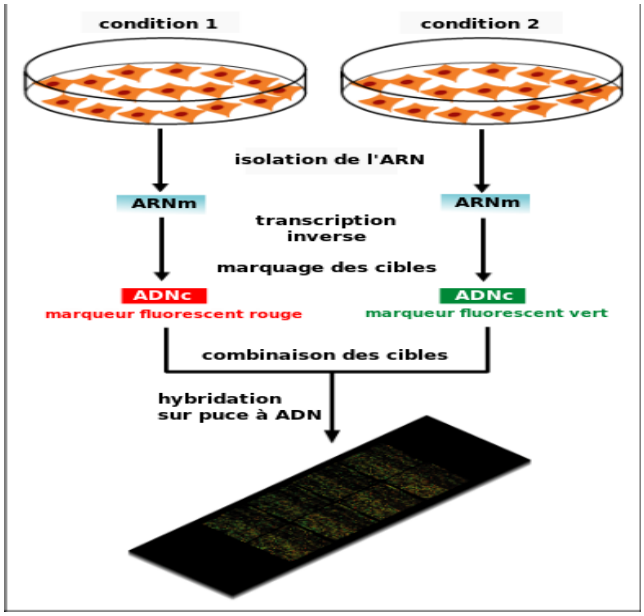


Méthode FISH

- **La cytogénétique moléulaire:** La technique CGH (hybridation génomique comparative), est une technique de cytogénétique moléculaire permettant d'analyser les variations du nombre de copies de l'ADN. Cette méthode permet d'observer s'il y a ajout (syndrome de Down: maladie causé par l'ajout d'un chromosome à la paire 21) ou délétion (syndrome de Turner: anomalie présente chez les femmes qui possèdent un chromosome X manquant) de chromosomes. Cette méthode consiste à marquer des ADN et des ADN tumoraux et les ajouter à de l'ADN de contrôle en métaphase. Le mélange d'ADN marqué est mélangé avec de l'ADN de contrôle en métaphase. L'ADN marqué va alors s'hybrider de façon compétitive avec l'ADN métaphasique mais avant ceci on lave le mélange avec de l'ADN COT-1 (échantillon capable de supprimer les séquences répétitive d'ADN) pour supprimer les excès. A l'aide d'un microscope épi fluorescence on cherche s'il y a une anomalie en ce qui concerne le nombre de chromosomes. Si le nombre de copies a augmenté, la sonde tumorale s'hybridera et le segment de chromosome apparaitra en vert, si le nombre de copies a diminué, la sonde de tissu sain s'hybridera préférentiellement et le segment de chromosome apparaitra en rouge, si le nombre de copies reste le même, les deux sondes s'hybrideront en quantité équivalente et le segment de chromosome apparaitra en orangé. La méthode CGH array est identique à la technique CGH mais in retrouve l'utilisation de la puce à ADN qui permet une analyse à plus haut débit ainsi que des résultats à une meilleure résolution, et peut en plus de trouver des anomalies au niveau du nombre de chromosomes, mais aussi en e qui concerne les mutations. Ces

mutations sont trouvées après une hybridation sur une puce à ADN qui analyse la séquence d'ARNm non traduite pour détecter la présence d'une mutation.





Puce à ADN

b-Où et comment on va stocker le génome humain :

L'analyse du génome étant accompli, on doit maintenant chercher où et comment on va stocker ce génome. Ceci crée un problème à cause de la taille de ce génome dont la valeur se donne Mb (méga bases), et où 1Mb représente un million de nucléotides.

- **Dans l'ADN :** Le génome humain est gigantesque. Pour le stocker il faut alors un « disque dur » capable de supporter une telle grandeur. L'université Harvard a réussi a stocké dans moins d'un gramme d'ADN 5 millions de bits. Pour tester la fiabilité de l'encodage sur la densité de stockage, c'est-à-dire la capacité de transcrire des informations dans l'ADN, l'European Bioinformatics Institute de l'université de Cambridge en Angleterre, a stocké dans 115 000 brins d'ADN, 6 millions de bits. La fiabilité d'encodage était d'à peu près 99.9%. Alors, la possibilité de stocker de l'ADN dans l'ADN est une possibilité de stockage du génome humain. Mais, il y aura un obstacle à surmonter qui est l'obligation de synthétiser de l'ADN en quantité et une base azotée coute 1000$. Alors, Sylvain Gariel, Thomas Ybert et Xavier Godron, les trois fondateurs de « DNA script », une compagnie biotechnique, ont eu l'idée de

créer l'ADN artificiel à l'aide de catalyseurs biologiques tel que les enzymes. Mais, la manipulation des enzymes est très complexe et il faudra trouver une enzyme facile à « dompter » pour qu'elles puissent contrôler le processus d'addition, c'est-à-dire l'ajout d'un nucléotide bien déterminé (et pas un autre). Pour cette compagnie, cette technique de fabrication de l'ADN permettra au temps de production de vingt jours à seulement une seule journée. Le processus de stockage dans l'ADN est très complexe. Après la production des ADN de « stockage », il suffirait d'hybrider l'ADN sur puce à ADN sur un verre pour ensuite les transférer vers l'ADN synthétisé.

- **Dans des ordinateurs à grande mémoire :** Il est possible de stocker le génome humain sous forme numérique dans des supers ordinateurs à grande mémoire de stockage (mémoire capable de supporter les données gigantesques du génome humain). Ces ordinateurs ne sont pas encore ouverts au public et sont encore en train d'être testé ou développé pour permettre ce genre de stockage.

- **Différence :** L'ADN a une forme tridimensionnelle ce qui permet à l'information stocké d'être plus compacte ce qui facilite son exploitation alors que le stockage dans des ordinateurs permet à l'aide du Big Data une analyse plus facile à manipuler.

Donc on peut déduire que pour avoir recours à la médecine prédictive on doit commencer par analyser le génome humain (soit par le séquençage soit par la cytogénétique) et le stocker dans des supers ordinateurs ou dans l'ADN pour qu'ensuite les analyser, avec l'aide du Big Data, avec d'autres données pour pouvoir prédire les maladies et de les prévenir.

2- Prédire les maladies et les épidémies

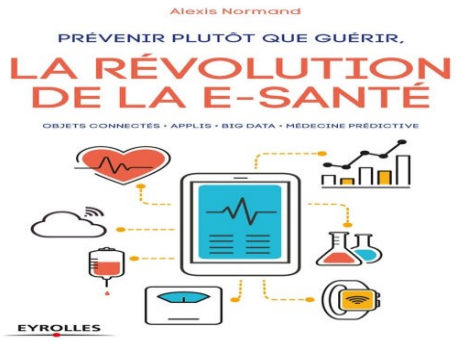
a-Prédire les maladies

Il y a de façons de prédire les maladies : soit plusieurs années ou mois à l'avance, la prédiction, soit quelques jours à l'avance, la prévention.

- **La prédiction :** Ce genre de médecine prédictive est dû à l'analyse du génome du patient, son phénotype macroscopique, son pedigree et l'environnement dans lequel il vit. Le génome seul représente quatre milliards de bits ce qui déjà impossible à

analyser par les ordinateurs de nos jours (ceux qu'on utilise au quotidien). Alors pour pouvoir avoir recours à cette branche de la médecine, il faudra avoir recours au Big Data et son analyse pour analyser toutes ces données. Ce travail est possible dès la naissance de l'individu. En effet, si la majorité de sa famille est atteint d'une maladie quelconque, il y a de grandes chances qu'il soit lui aussi atteint de cette maladie à un moment dans sa vie.

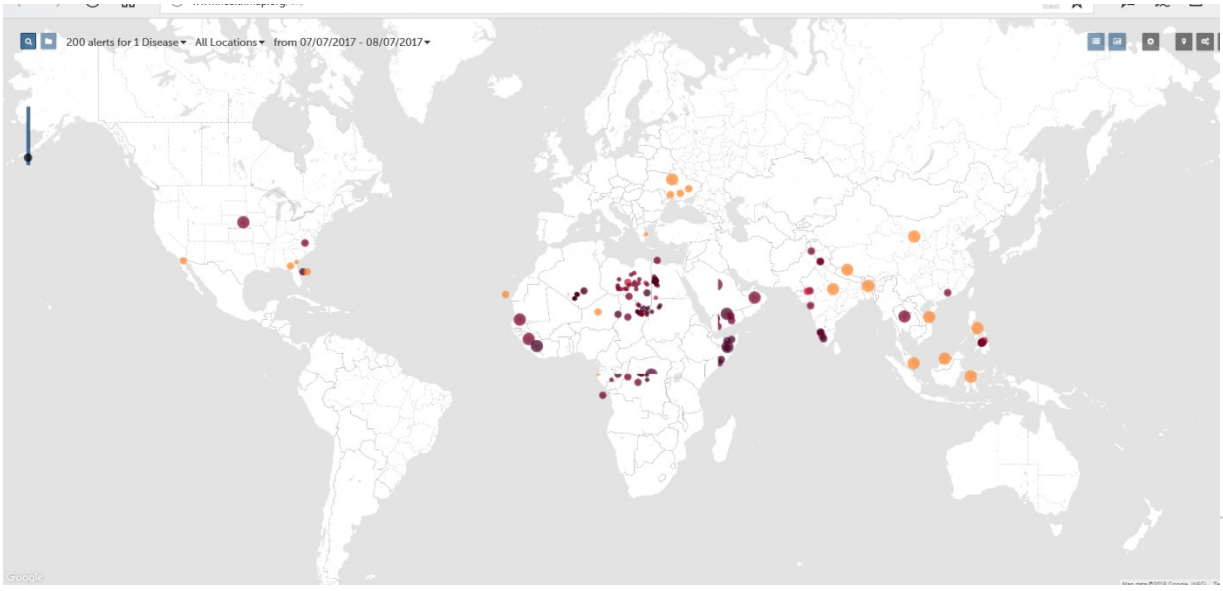
- **La prévention :** La prévention, est aidée par la présence de montre intelligente qui trouve les paramètres hémodynamiques d'un individu. On peut trouver la montre Fit Bit ainsi que Diabeo qui suivait l'évolution du glucose et partager les données par smartphone avec le médecin. Mais la prévention est aussi possible sans objets connectés: s'il ressent quelque chose d'anormal, il peut passer chez son médecin qui analysera son génome (il déterminera s'il est porteur, malade ou sain), son pedigree et son environnement. Deux cas se présentent : soit il est atteint d'une maladie génétique, soit il est atteint d'une maladie causé par l'environnement dans lequel il vit (pollution…).



b-Prédire les épidémies (augmentation rapide du nombre de personnes atteintes d'une même maladie) :

- **A l'aide de satellites :** Grâce à l'apprentissage automatique dans les satellites, ces derniers sont capables de trouver les températures, les précipitations, l'humidité du sol, le type de végétation et l'utilisation des terres sur les différentes régions de la Terre. Grâce au Big Data, ces satellites pourront analyser ces informations et prédire l'apparition d'une épidémie en les combinant à des informations de santé dans un modèle informatique. Cet analyse se fait suivant le modèle de l'algorithme « MapReduce ». L'algorithme commence par regrouper toutes les informations citées au-dessus et les numériser. Ensuite il rentre dans la phase Reduce où il regroupe toutes les informations nécessaires à la prédiction de la maladie et ceci se fait à l'aide du Machine Learning.

- **A l'aide de Health Map:** Health Map, est une compagnie qui à l'aide d'analyse de données peut détecter l'apparition des épidémies ainsi que leur évolution. Elle transmet ses alertes à travers leur site web ainsi qu'à travers des organisations mondiales comme l'organisation mondiale de santé, les centres de prévention et de contrôle des épidémies en Europe et aux Etats-Unis. A l'aide du Big Data, Health Map, a prédit la propagation d'Ebola en Afrique de l'Ouest neuf jours avant l'annonce officiel de l'OMS en décembre 2013. Ce logiciel a réussi de prédire à l'aide de l'analyse des réseaux sociaux, des bulletins d'informations locaux et d'autres bases de données. Ce logiciel peut ensuite suivre la propagation des épidémies en analysant les flux humains. Les technologies analytiques permettent en effet de suivre l'évolution d'Ebola, H1N1 ou Malaria. HealthMap se base aussi sur l'algorithme MapReduce, mais fait entrer d'autres informations en plus comme les réseaux

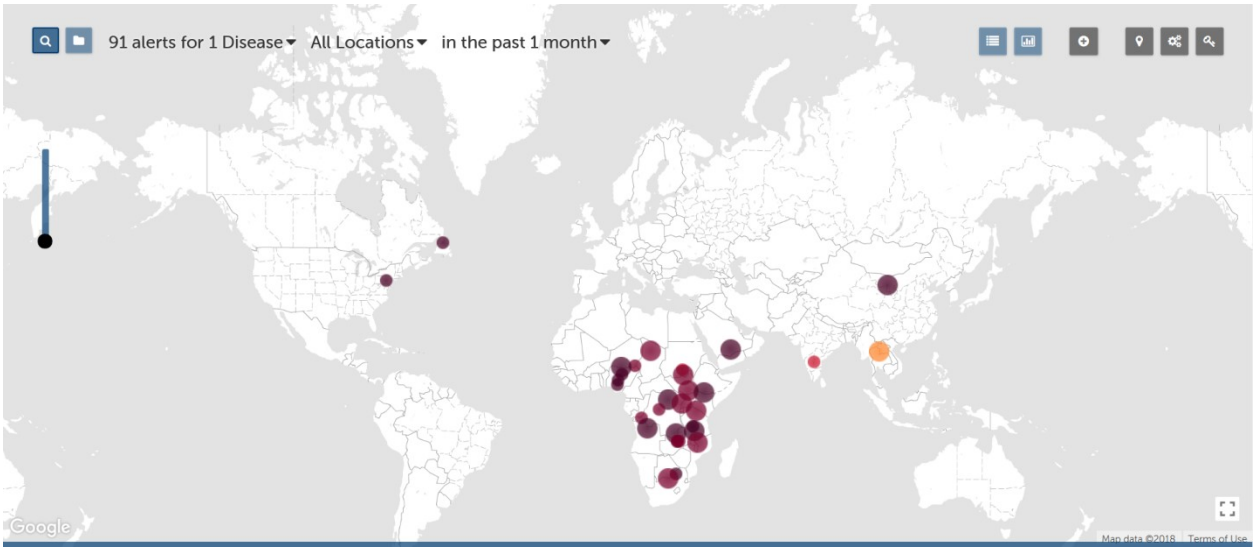


sociaux. Pour suivre leur évolution on fait entrer des données qui permettent de montrer l'évolution comme les déplacements des personnes.

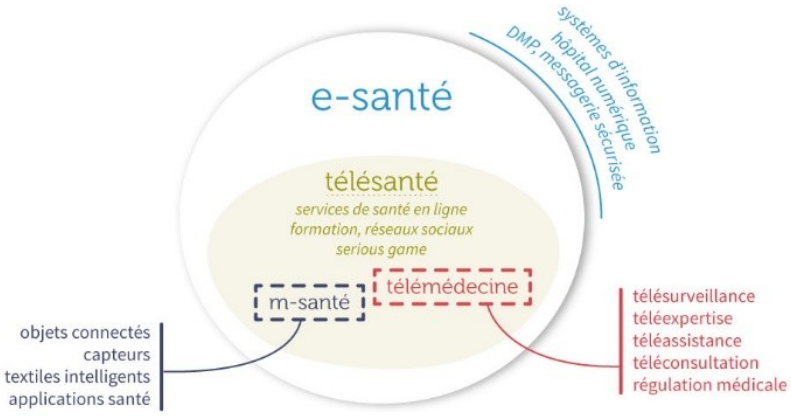
Alors on peut conclure que la médecine prédictive permet la prédiction d'épidémies et de maladies ainsi que leur prévention, et ceci grâce à l'évolution du Big Data et du Machine Learning qui permettent de détecter certaines informations et d'analyser de grandes bases de données.

C-Télémédecine

Aujourd'hui, le domaine de la santé est constamment en développement avec l'application des technologies de l'information et de la communication (TIC) à laquelle on attribue le terme e-santé. La télésanté, quant à elle, fait référence au développement du numérique permettant par exemple la transmission de dossier ou d'informations médicales ou



la coordination de professionnels de santé autour de la prise en charge ou du suivi du patient… A l'intérieur de ce concept global, on différencie deux actes ; le m-santé et la télémédecine à laquelle on s'intéressera dans cette partie.



Source : Qu'est-ce que la télémédecine ? *Calendovia Blog*
http://blog.calendovia.com/telemedecine-teleconsultation-esante

La télémédecine est une forme de pratique médicale qui s'effectue entre un médecin et un patient situés dans des lieux séparés. Elle utilise les technologies de l'information et de la communication (TIC) pour assurer leur interaction, qui se fait à travers les ordinateurs. Ainsi, grâce à l'analyse des données recueillis par ces machines, le médecin pourra établir un diagnostic, assurer un suivi médi cal, préparer une décision thérapeutique ou même prescrire des médicaments ou des actes. La télémédecine assure d'autre part l'interaction entre des médecins à distance, échangeant leurs avis professionnels pour des diagnostics plus précis et décisions encore plus exacte.

La télémédecine comprend généralement cinq actes :

- Téléconsultation (ou télédiagnostic) : Consultation médicale à distance entre un médecin et un patient qui permet au premier de poser un diagnostic, d'après les informations qui lui seront fourni par l'ordinateur.
- Télésurveillance médicale : Interprétation à distance des données nécessaires au suivi médical du patient. La transmission de ces données peut être automatisée ou réalisée par le patient lui-même ou par un professionnel de santé.
- Télé-expertise : Echanges des avis, à distance, entre le médecin et un ou plusieurs collègues sur la base d'informations recueillis sur le patient.
- Téléassistance médicale : Aide à distance d'expert(s) lors de l'accomplissement d'un acte médical.

- Régulation médicale : Appel du centre 15 du Service d'Aide Médicale Urgente (SAMU) pour déclencher à distance une réponse adaptée à l'état du patient.

La télémédecine offre pluridisciplinarité. En effet, plusieurs spécialités médicales ont recourent à la télémédecine telles que la chirurgie, la gynécologie ou la dermatologie.

1. Big Data et télémédecine

a) La transmission à distance

La télémédecine est basée sur l'échange de données informatiques entre plusieurs postes éloignés. Pour permettre la transmission d'informations, plusieurs supports physiques ont été testés tels que les faisceaux Hertziens qui permettent de relier des points qui ne peuvent être reliés par des câbles en envoyant des signaux radioélectriques entre deux stations fixes équipés d'antennes directives.

La fibre optique est également l'une des méthodes utilisées pour la transmission d'informations numériques en conduisant de la lumière entre deux lieux distants. Le signal lumineux envoyé est capable de transmettre une grande quantité d'information et sera ensuite transformer en signal électrique qui pourra être compris par l'ordinateur ou la machine capteur.

Les liaisons par satellite est une technique plus récente. Le satellite est un relais qui reçoit des ondes depuis une station terrestre émettrice et les retransmet simultanément à une ou plusieurs stations réceptrices. Les signaux électro-électriques reçus sont amplifiés et retransmis sur une bande de référence différente de la bande de fréquence émettrice.

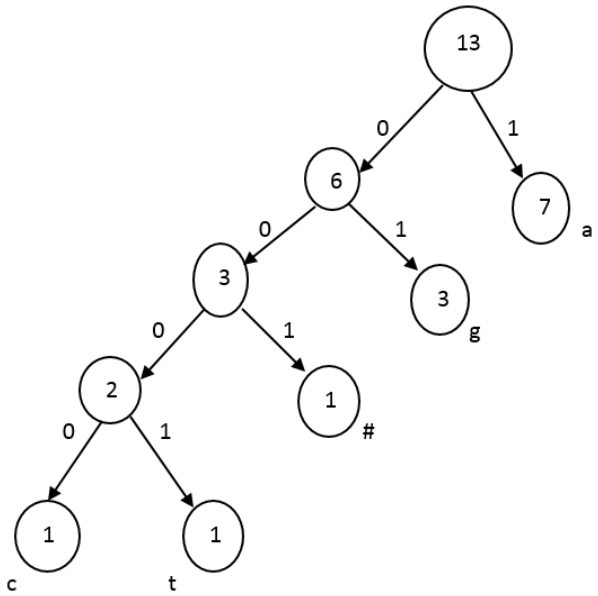
b) Le transfert de données en imagerie médicale

Les images médicales donnent des informations sur la forme et la fonction des organes du corps humain, ceci étant l'un des moyens les plus importants pour le diagnostic. En effet, les images provenant des appareils radiographiques sont directement affichées sur l'écran du radiologue et sauvegardées sous forme numérique. Le problème avec les images et les vidéos est qu'elles nécessitent une grande quantité de bande passante, pour la transmission et la réception. Par conséquent, il est nécessaire de réduire la taille de l'image qui doit être envoyée, et ceci en utilisant le codage de Huffman, un algorithme de compression qui maintien la qualité de l'image originale. Celui-ci consiste à remplacer les caractères les plus fréquents de l'information par des codes courts et les caractères les moins fréquents par des

codes longs. Prenons l'exemple de l'information cagataagagaa# à compresser. Chaque caractère étant associer à 7 bits (d'après l'ASCII, code américain normalisé pour l'échange d'information), et ayant 13 caractères dans cagataagagaa#, on peut calculer le nombre de bits que cette information occupe : 13 x 7 = 91 bits. Pour la compresser, on relève premièrement les fréquences d'apparition des caractères de l'information.

a	c	g	t	#
7	1	3	1	1

On peut ainsi dresser un arbre binaire créé suivant un principe simple : on associe à chaque fois les deux nœuds de plus faibles poids, pour donner un nouveau nœud dont le poids équivaut à la somme des poids de ses fils. On répète ce processus jusqu'à n'en avoir plus qu'un seul nœud : la racine. On associe ensuite par convention le code 0 à chaque embranchement partant vers la gauche et le code 1 vers la droite. Dans ce cas, on rejoint le « c » et le « t » pour former un nœud d'un poids de 1+1=2. De même, on rejoint le caractère « # », peu fréquent aussi, avec le nœud précédemment crée, pour former un autre nœud avec un poids de 2+1=3. Et ainsi de suite, jusqu'à arriver à la racine de l'arbre dont le poids vaut dans ce cas 13.

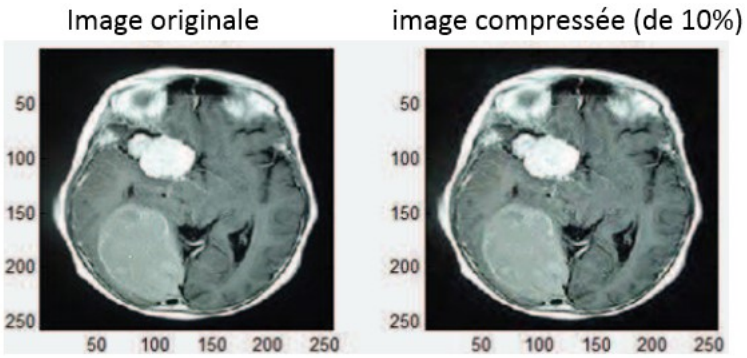


Pour obtenir le code binaire de chaque caractère, on remonte l'arbre à partir de la racine jusqu'aux feuilles en rajoutant à chaque fois au code un 0 ou un 1 selon la branche suivie.

Ainsi, nous pouvons lire les codes correspondant à chaque caractère :

a	c	g	t	#
1	0000	01	0001	001

Le code binaire de l'information initiale : « cagataagagaa# » devient alors :
0000 1 01 1 0001 1 1 01 1 01 11, soit 21 bits au lieu de 91 bits.



Radiographie montrant différence entre l'image normale et l'image compressée

c) L'analyse de données en télémédecine

Les appareils utilisés lors de la téléconsultation et qui sont reliés aux ordinateurs capteurs vont constamment générer des données qui seront recueillies et analysées par le médecin à distance. Cette analyse sera assurée par des systèmes tels que le programme Watson. Au lieu d'échantillonner les données pour l'analyse, ce système détectera modèles de maladie et alertera le patient et le médecin au premier signe d'anomalie, ce qui est particulièrement important pour les maladies chroniques, c'est-à-dire qui dure longtemps. La variabilité des données sera si grande que nous serons capables de détecter les différences sur la façon dont la même maladie affecte sous-population différente et administrer le bon médicament selon le phénotype approprié lors de la régulation médicale.

Le logiciel d'intelligence artificielle Watson est capable d'emmagasiner, de comprendre et de croiser d'importants volumes d'informations. Il est donc capable de lire des articles de presse ou de recherche, des tweets, des romans ou encore des « posts » de blogs écrits en anglais et dans sept autres langues, grâce à un entrainement solide réalisé par les chercheurs d'IBM (International Buisness Machines) avec des technique de traitement du langage naturel, et peut ainsi « apprendre » de plus en plus en enregistrant toutes les informations qu'il croise. Dans notre cas, ce sont des informations médicales se rapportant au corps humain et à ses organes.

2. Machine Learning et télémédecine

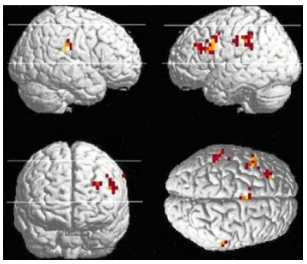
a) L'apprentissage automatique en imageries médicales

En télémédecine, l'apprentissage automatique (Machine Learning) revient à apprendre à l'ordinateur à détecter une maladie en observant des images provenant d'imageries médicales, tels que l'IRM qui est une imagerie par raisonnante magnétique permettant d'obtenir des vues en deux ou trois dimensions de l'intérieur du corps de façon non invasive. Pour « apprendre » à l'ordinateur à analyser ces images en 2D, des milliers d'images de patients analysées par des spécialistes sont chargées sur le programme. L'ordinateur utilise un algorithme qui permet d'extraire les caractéristiques de certains traits du visage par un processus semblable à l'appareil visuel humain. En effet, notre cerveau enregistre inconsciemment des caractéristiques du visage humain et les associe à des particularités comme le sexe ou l'âge pour nous permettre de mieux reconnaître les personnes autour de nous. Afin de permettre à l'algorithme « d'apprendre » ses caractéristiques, on doit préalablement l'alimenter de nombreuses images, au moins des milliers, qui seront utilisées comme références. L'algorithme associe les caractéristiques extraites de la nouvelle image aux caractéristiques semblables des images de références et estime par la suite les traits d'une personne de la photo en faisant une pondération des résultats obtenus.

Echographie (exemple d'imagerie médicale 2D)

Quant aux images médicales 3D, les IRM cérébrales par exemple, des techniques sont misent en place pour apprendre à l'ordinateur de les analyser. Des caractéristiques de cerveaux de gens atteints de certaines maladies sont comparées à celles de cerveaux sains afin

de les classifier. Ce genre de classification pourrait être utilisé dans plusieurs applications, notamment le diagnostic de maladies avant l'apparition des premiers symptômes, l'évaluation de la progression de maladies et la prédiction de la gravité d'une maladie.



IRM cérébrale (exemple d'imagerie médicale 3D)

b) La robotique en chirurgie

b.1) Télé-chirurgie

La télé-chirurgie est une intervention chirurgicale effectuée à distance, ayant recours aux nouvelles technologies de l'information et de la communication. Elle comporte deux aspects : l'assistance chirurgicale à distance d'un médecin expert et la chirurgie à distance assistée par ordinateur ou par robot. Les deux types de télé chirurgie nécessitent la transmission des images d'un patient. La télé-chirurgie offre aux chirurgiens un nouvel outil de précision dans leur pratique. Les informations envoyées par les outils virtuels, manipulés par le chirurgien, sont traitées par un ordinateur qui transmet simultanément les ordres aux appareils situés sur le champ opératoire. Le chirurgien suit l'intervention à partir d'un écran de contrôle. L'ordinateur peut être programmé pour qu'il modifie l'ampleur des gestes, par exemple d'un facteur dix. Ainsi, lorsque le chirurgien bouge sa pince sur un centimètre, le bras robotisé peut effectuer le même geste en réduisant la distance à un millimètre. Plus précisément, l'image vidéo émise par la camera endoscopique du bloc opératoire est une image analogique. Ce signal analogique est transformé en signal numérique par un appareil appelé transcodeur qui va coder puis décoder l'image. L'image est donc décomposée à la source, transmise, recodée puis recomposée sur le site distant en un laps de temps à peine détectable à l'œil avec un débit de 25 images par seconde. A ce transfert de l'image se superpose un transfert de son synchrone de l'image.

b.2) Chirurgie robotisée

La chirurgie telle qu'on la connait maintenant va plus tard disparaitre, due au développement de la technologie qui permet de crée un robot qui se chargera d'opérer le patient. En effet, il

sera désormais possible, avant d'opérer un malade, de créer son « clone digital » à partir de ses images médicales. On pourra pratiquer l'opération de façon simulée sur ce clone pour ensuite opérer le patient grâce à la chirurgie assistée par ordinateur, c'est à dire la robotique.

L'application de la chirurgie robotisée couvre tout un processus opératoire depuis l'acquisition et le traitement des données jusqu'à l'intervention chirurgicale et au contrôle post-opératoire.

En phase pré-opératoire, il s'agit de modéliser, pour chaque patient, les organes rigides (comme les os) ou déformables (comme le coeur), qu'il faut opérer. Pour faire ceci, il faut exploiter les spécificités des modalités d'imagerie médicale et les informations pertinentes qu'elles proposent. Les structures anatomiques mises en évidence sont ensuite utilisées lors de la préparation du planning opératoire et de sa simulation. Ces deux phases intègrent les modèles mécaniques du robot pour décrire et simuler les mouvements, et éventuellement les forces réalisables. Ce planning est ensuite mis en corrélation avec le patient au bloc, en phase post-opératoire. Le système robotique peut alors fournir une aide active en appliquant les mouvements du chirurgien pour la réalisation précise de la procédure planifiée. Dans certains cas, le robot peut agir de façon autonome pour réaliser tout ou partie de la procédure opératoire envisagée.

UN ALLIÉ ENCORE INCERTAIN

Le Big Data peut être révolutionnaire dans la médecine. Il peut toucher beaucoup de domaines, et potentiellement sauver des vies qui, autrement, ne pourront pas être sauvées. Mais l'utilisation du Big Data en médecine n'est pas simple, et a plusieurs difficultés et limites. On s'intéressera aux problèmes du Big Data en général et les solutions, puis les problèmes du Big Data spécifiques à la médecine et les solutions.

A- Problèmes dans l'utilisation générale du Big Data

Le Big Data promet beaucoup dans un grand nombre de domaines. Mais toutes ces promesses sont faites sans considérées un grand nombre de limites. C'est un "buzz World", avec lequel il faut être prudent.

Lors d'un projet de Big Data, des difficultés peuvent apparaitre à plusieurs niveaux. Au niveau méthodologique, un projet de Big Data doit créer un nouveau besoin aux clients.

Un projet de Big Data doit aussi faire face à des problèmes d'infrastructure. Mais ce ne sont pas de simples problèmes de cout. Les infrastructures du Big Data doivent être très sécurisées, pour pouvoir lutter aux bugs d'une part, et avoir des solutions de secours en parallèles dans ce cas, et surtout lutter aux hackers, qui pourront exploiter des données secrètes. Il faut donc que des experts s'occupent de ces infrastructures, sans que le cout soit une limite. Des programmes et des algorithmes très sophistiqués devront être mis en place. Les infrastructures doivent aussi supporter une grande base de données, facilement et rapidement accessibles, pour qu'elles soient efficaces.

On ne sait pas vraiment combien de données on a bel et bien et besoin pour opérer la Big Data. Des sites internet comme Amazon, Facebook, et Google ont un Traffic de données journalier de qui peut arriver aux pétotes, soit un ordre de grandeur de 10*15. Mais ce ne sont pas des exemples, et bels et bien des exceptions, leur Traffic étant les plus grands mondialement.

Sur le long terme, des problèmes d'organisation peuvent survenir. Si un projet de Big Data devient assez grand, il faudra une équipe de plus en plus grande pour s'en occuper, et l'équipe devra être organisée en plusieurs catégories, chacune s'occupant d'un aspect du data Lake.

Il faut donc assurer un grand budget dédies exclusivement au Big Data, pas seulement pour l'infrastructure, mais pour assurer des compétences technologiques nécessaires pour faire évoluer l'architecture technologique du data Lake. Ce budget peut être très variable selon les secteurs et le progrès de la technologie, les compagnies n'ont donc pas une estimation ne précise de l'argent nécessaire.

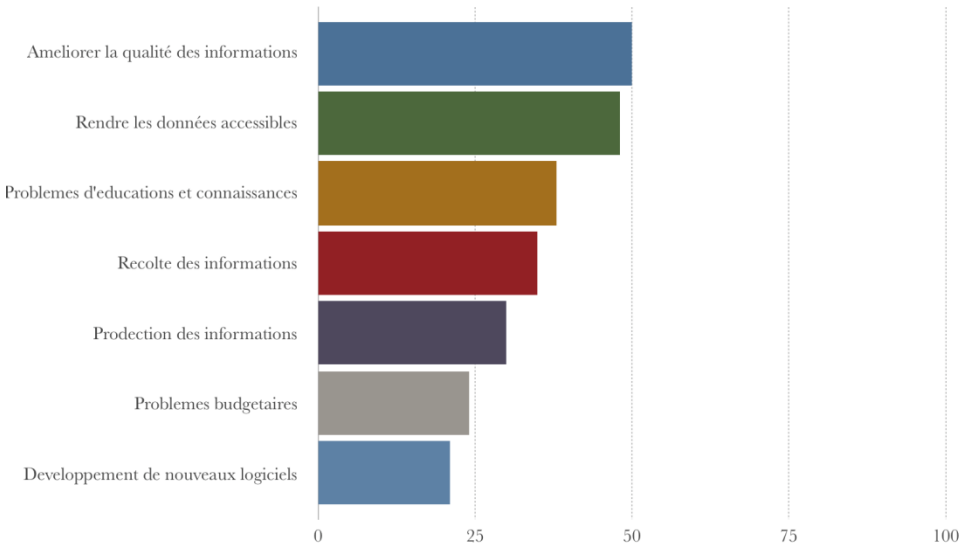
L'utilisation du Big Data n'est donc pas aussi simple qu'on peut le croire. Il faut donc modérer son enthousiasme.

Le Big Data pose aussi des problèmes face à son exploitation, qui, en plus, ne prennent pas compte des problèmes face à son initiation.

Le Big Data nécessite beaucoup de données. Mais pas n'importe quelles données, il faut que les données soit de qualité, récentes, cohérentes au sujet et assez stable. On a aussi besoin d'anciennes données, historiées, pour quelques utilisations du Big Data.

Solution : Une entreprise qui veut utiliser la Big Data a besoin de beaucoup de données variées. Il ne faut donc pas que les entreprises se limitent à leur données internes, car celles-ci, mêmes si elles sont très intéressantes par rapport à l'entreprise, ne lui sont pas suffisantes. Il faut donc ouvrir des banques de données publics, l'open data, et aussi généralement sur le web. Cela peut être contre les principes de certaines entreprises, car leurs informations pourront aider des concurrents, mais l'ouverture des données à l'extérieur est nécessaire à l'utilisation efficace de la Big Data. C'est ce que recommande GARTNER, très optimiste par rapport à l'utilisation du Big Data.

Il faudra former des experts dans le domaine de la Big Data. Il faut former des Data-scientists, qui seront capables, d'exploiter des algorithmes de machine learning, et bâtir des programmes pour pouvoir répondre aux besoins spécifique de l'utilisateur



Les plus grands problèmes face à la mise en place du Big Data : (données d'après bigdatasurvey.nl)

B- Les problèmes du Machine Learning et les algorithmes

Les algorithmes sont considérés comme une révolution technologique, qui permet beaucoup d'avancements dans un grand nombre de domaines, notamment le machine learning. Mais les algorithmes et le machine learning ne sont pas parfaits.

Les algorithmes permettent de résoudre des problèmes décidables. Mais ils ne sont pas capables de résoudre des problèmes ouverts, “indécidables”. Par exemple, un algorithme peut facilement répondre à la question : les enfants trouveront ils la baguette magique au bout de la 5eme étape de leur parcours ? Mais si la question devient : À quelle étape du parcours les enfants trouveront ils leur baguette magique ? La résolution peut s’avérer impossible. Si les enfants trouvent leur baguette magique après un milliard d’étapes, l’algorithme aura besoin d’un temps indéfini pour trouver la réponse, et n’est donc plus efficace. Si le parcours n’a en fait pas de fin, l’algorithme essayera les étapes infiniment sans trouver de réponse. Heureusement, la plupart des problèmes sont “décidables”. Mais les problèmes indécidables, ne pouvant pas être résolus en pratique, posent problème. Certaines questions indécidables ont un prix d’un million de dollars pour être résolues à l’aide d’un algorithme, elles sont donc qu’elles sont pratiquement impossibles à résoudre.

Même si le temps de calcul est le principal problème des algorithmes, il existe d’autres problèmes, par exemple l’espace mémoire pour stocker l’information, et l’énergie consomme par les ordinateurs faisant les calculs.

Les algorithmes ne sont pas parfaits. Ils peuvent présenter des erreurs dans leurs codes, et provoquer des bugs. Par exemple, un SMS envoyé à notre grand-mère peut être reçu par notre ami à cause d’un bug. Mais cela n’est qu’un petit problème. Il faut imagine que ces bugs peuvent toucher par exemple un robot chirurgien, ou les donnes de machine learning concernant la médecine. Ce genre d’erreurs peut mettre des vies en ligne, et voilà une des grandes limites des algorithmes. Un algorithme complexe est souvent constitue de millions de lignes de codes, et une erreur dans ce code ne serait pas surprenante. On peut corriger les erreurs, mais le risque d’en faire et la difficulté de correction est un grand point d’interrogation sur le futur des algorithmes en santé et médecine.

La sécurité des algorithmes pose aussi problème Les algorithme peuvent d’une part envahir des données personnelles, ce qui peut poser problème a un grand nombre de personnes. D’autre part les algorithmes peuvent être compromis par des sources externes. Les “hackers” ont pour seul but d’accéder aux algorithmes d’autres personnes et sociétés. Même avec beaucoup de précautions, les algorithmes peuvent toujours être envahis par les hackers.

Le machine learning étant dépendent des algorithmes, il pose donc un grand nombre de limites et problèmes, mais qui peuvent être résolve avec du travail.

Exemple d’algorithme, d’après lemonde.fr. On peut voir que l’algorithme est très complique, d’où les erreurs et la difficulté de correction.

C- Les problèmes dans l'utilisation du Big Data en médecine

Actuellement, les médecins et les concernés de la médecine ne sont pas formés pour manipuler la donnée fournie par la Big Data. Ils ont fait leurs études dans un système qui s'oppose à l'informatique.

La Big Data ne servira en médecine sans certaines procédures :

Il faut commencer à introduire la formation dans les écoles de médecine, pour que les prochaines générations de médecins soit capables d'opérer la Big Data. Des études en maths et plus précisément en algorithmes et manipulation d'ordinateurs seront requises dans les programmes de médecine. Mais cela fera des médecins des ingénieurs aussi, et leurs études seront beaucoup plus difficiles. La meilleure solution sera donc de former des professionnels dont la spécialisation est le Big Data en médecine.

Une autre solution pourrait être la création de logiciels qui pourront afficher les données clairement a des personnes qui n'ont pas de formation en Big Data. Il faudra donc que les scientifiques qui n'ont aucun rapport avec la data puissent explorer les données facilement.

Un autre problème est le grand écart qui existe entre le milieu de recherche et la pratique actuelle concernant le Big Data dans la ma médecine. Cet écart mène à une mauvaise productivité ainsi qu'a des erreurs d'exécution. Ce phénomène est du à plusieurs raisons, les restrictions financières et réglementaires du programme de santé, la résistance des médecins aux nouvelles technologies et leur manque de formation dans ces domaines.

Pour résoudre ce problème, des applications sont développés pour pouvoir faciliter l'interprétation des données concernant la santé. Par exemple, l'application Asthmapolice qui fait un suivie GPS des asthmatiques. En trouvant les locations ou les patients souffre le plus des crises d'asthme, elle permet de reconnaitre les facteurs qui catalysent l'asthme, la pollution par exemple.

La Big Data est basée sur des algorithmes. On sait que les algorithmes ne sont pas toujours parfaits, et certains bugs ou erreurs dans ces algorithmes qui affectent le Big Data en relation avec la santé sont de grands points d'interrogations sur le futur du Big Data en santé. Des vies peuvent être perdues en raison de certains bugs. Il faut résoudre ces bugs et ses erreurs à l'aide de plus de programmeurs pour assurer le futur de la Big Data en médecine.

Mais le vrai problème de la Big Data par rapport à la médecine n'est pas technique ou économique, mais bel est bien politique et juridique.

Les problèmes économiques et techniques par rapport à la Big Data sont de mieux en mieux maitrisés, même si la sécurité des systèmes d'information peut encore être améliorée. Mais les questions d'accès aux données et les conditions de leur exploitation sont encore un problème au niveau politique. Il faut que tous les partis politiques dans tous les grands réglementent l'utilisation de ces données. La mise en commun de toutes les données personnelles par rapport à la santé des individus nécessite des changements et des reformes juridiques, qui nécessitent des décisions politiques. Mais cela n'est pas irréaliste, car l'amélioration de la prise en charge des patients et la prévention des maladies est une grande ambition politique. Les politiciens qui proposeront des procédures pour améliorer la santé du peuple seront favorisés par ce peuple, et il existe donc une vraie motivation politique pour implémenter le Big Data dans la médecine au niveau national.

Conclusion

Alors on peut conclure qu'avec les grandes capacités de l'analyse du Big Data et malgré ses limites, la médecine du futur sera bien développée, les maladies et les épidémies seront prédictibles. Le développement de nouvelles techniques en diagnostic médicale et télémédecine de nouveaux traitements seront accessible ce qui limitera les erreurs en médecine .De plus le développement du Machine Learning va accentuer l'évolution de la médecine. Avec la création de super computer, computer quantique et l'accroissement des algorithmes les limites du développement du Big Data seront presque inexistantes.

Bibliographie

"Santé : l'intelligence artificielle et le machine learning au service de la lutte contre le cancer." *L'intelligence artificielle au service de la lutte contre le cancer*, rslnmag.fr/innovation/sante-intelligence-artificielle-lutte-contre-cancer-microsoft-research/

Peltier, Futura Par Claire. "Le séquençage de l'ADN pour les nuls !" *Futura*, www.futura-sciences.com/sante/actualites/genetique-sequencage-adn-nuls-26754/.

"Magazine business & Internet des Objets." *Objetconnecte.com*, www.objetconnecte.com/-de-demain-besoin-big-data/.

Raynal, Juliette. "L'ADN, le disque dur de demain ?" *Industrie et Technologies : Veille des technologies émergentes et des solutions innovantes. Ingénieur de l'année, CNISF et derniers brevets*, 21 Aug. 2015, www.industrie-techno.com/l-adn-le-disque-dur-de-demain.39324.

"Médecine personnalisée." *Médecine personnalisée, la médecine de demain | Roche*, www.roche.fr/pharma/medecine-personnalisee.html.

"Médecine personnalisée en cancérologie: le rôle du Big Data." *YouTube*, YouTube, 31 Aug. 2016, www.youtube.com/watch?v=QNkaSY4brho&t=50s.

Céline Poirier, Univadis Marketing Operations Manager Follow. "Le Big data en santé et l'éthique, sont- ils compatibles ?" *LinkedIn SlideShare*, 6 Mar. 2016, www.slideshare.net/CelinePOIRIER/le-big-data-en-sante-et-lthique-sont-ils-compatibles?qid=6d365eca-8a30-42c4-a9b4-1e4a883ea03d&v=&b=&from_search=5.

"Que signifie NoSQL (Base de données « Not Only SQL »)? - Définition par WhatIs.Com." *LeMagIT*, www.lemagit.fr/definition/NoSQL-base-de-donnees-Not-Only-SQL.

"Big data, c'est quoi ? Définition simple du big data ? Avoir peur du big data." *Culture Informatique*, 12 Sept. 2017, www.culture-informatique.net/cest-quoi-le-big-data/.

Cea. "Le Big Data." *CEA/Découvrir & Comprendre*, CEA, 16 Juin 2017, www.cea.fr/comprendre/Pages/nouvelles-technologies/l-essentiel-sur-le-big-data.aspx.

"Définition : Qu'est-Ce que le Big Data ?" *LeBigData.fr*, www.lebigdata.fr/definition-big-data.

L, Bastien. "Machine Learning et Big Data – Les données sont l'essence de l'apprentissage automatique." *LeBigData.fr*, 5 Jan. 2018, www.lebigdata.fr/machine-learning-et-big-data.

Written by Bernard Marr. "A brief history of big data everyone should read." *World Economic Forum*, www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/.

Lemberger, Pirmin, et al. *Big Data et Machine Learning: Manuel du data scientist*.

Béranger, Jérôme. « E-santé, m-health, big data médicaux vers une hiérarchisation des données médicales. » *Revue hospitalière de France*, n°562, janvier-février 2015, p. 70-74.

Disponible sur : http://www.adel-label.com/wp-content/uploads/2016/05/39eme-article-beranger.pdf

Chirurgie robotique. *Institut Mutualiste Montsouris*, 2014. Disponible sur : https://imm.fr/loffre-soins/medecine-experte/chirurgie-assistee-robot/

"Codage de Huffman." *CommentCaMarche*, Disponible sur : www.commentcamarche.net/contents/1209-codage-de-huffman

Simon, Pierre et Acker, Dominique, novembre 2008. « La place de la télémédecine dans l'organisation des soins ». *Ministère de la Santé et des Sports. Direction de l'Hospitalisation et de l'Organisation des Soins*. Rapport Mission thématique n° 7. Disponible sur : http://solidarites-sante.gouv.fr/IMG/pdf/Rapport_final_Telemedecine.pdf

Le médecin, la télémédecine et les technologies de l'information et de la communication - Guide d'exercices, février 2015. *Collège des médecins du Québec*. Disponible sur : http://www.cmq.org/publications-pdf/p-1-2015-02-01-fr-medecin-telemedecine-et-tic.pdf

"Watson : comment marche l'IA d'IBM dans la santé, la banque..." *Journaldunet.com*, JDN, Disponible sur : www.journaldunet.com/solutions/reseau-social-d-entreprise/1196452-ibm-watson/.