Algorithmes 2/3

Soulever le capot Sortir de la phobie ou de la philie 2 parties : -les systèmes à apprentissage -les données	Avertissement : cette séance va traiter d'autres types d'algorithmes pour lesquels, comme pour APB, on va s'interroger sur les impacts qu'ils peuvent avoir sur nous. Mais, comme pour APB, et selon l'intention exprimée en introduction avec le texte de Dominique Cardon, il va falloir « ouvrir le capot ». Dans la première séance j'avais du parler de numérique, d'information, de calculateur, … cette fois on va parler de corrélation, de réseaux de neurones, d'apprentissage, de big data…. Ce sera évidemment sommaire mais ce me semble indispensable pour comprendre ces automatismes et pouvoir ainsi les analyser et, le cas échéant, les critiquer.
Différentes familles d'algo	- Les algorithmes d'appariement (matching) : très répandus : école, emploi, médecine, rencontres, ….(pas de prix= Les questions à se poser sont du même ordre que celles envisagées pour APB. - Les algorithmes de recommandation : de livres, de films, de formation, d'emploi, de vote, etc. Les questions importantes : les données de départ, la bulle ?, Loyauté difficile. - Les algorithmes d'aide à la décision : GPS, justice, médecine, police, Poids donné au passé (ex : liberté conditionnelle) et PB de responsabilité. - Les algorithmes de prédiction : décrochage scolaire, épidémie, anti-terrorisme, accueil des œuvres culturelles,… PB de choix des critères, des données, …. - Les algorithmes générateurs de connaissances : page rank de Google, logiciel de l'affaire Grégory, …
A croiser avec la technique Déterminisme Logique déductive Basé sur des études consultables et critiquables	L'algorithme utilisé par APB repose certes sur la trouvaille des deux mathématiciens (qui leur a valu le prix Nobel) mais l'idée une fois trouvée, il n'y a derrière qu'une programmation informatique tout à fait ordinaire L'algorithme utilisé ici est **déterministe** : il doit donner une solution. ou, au minimum, plusieurs solutions possibles entre lesquelles choisir. La plupart des algorithmes insérés dans le monde technique le sont aussi. Le problème à résoudre peut être compliqué (aller de Clermont à Barcelone en transport en commun…) mais les éléments du problème existent. Dans ce type d'algorithme, on est dans le domaine de la logique déductive : la réponse est certaine. Les résultats sont déduits des règles et/ou des résultats précédents : si l'élève a telles ou telles caractéristiques alors il est affecté à telle formation ; si le parcours x et plus court ou plus rapide que le parcours y alors choisir x, etc. Les caractéristiques qui vont déterminer le résultat du calcul peuvent être complexes, résultats accumulés de travaux, de connaissances, etc. Si l'expérience du magistrat qui doit décider ou non de la liberté conditionnelle d'un détenu observe que cette personne n'a ni emploi, ni qualification, ni famille, ni logement….son expérience ou les consignes qu'on lui donne peuvent lui faire refuser la mise en liberté. Si une étude des mises en liberté conditionnelles a été effectuée, des consignes ont pu être énoncées avec les explications adéquates. Ces consignes peuvent être incorporées dans un programme informatique ce qui donnera un algorithme d'aide à la décision du magistrat. On verra que cette automatisation provoquée par le passage de l'étude sociologique à l'algorithme pose déjà une série de problèmes (quelle est alors la liberté et la responsabilité du juge ?) mais l'étude

	sociologique existe, on peut la consulter, la critiquer, savoir pourquoi et dans quelle mesure on peut suivre l'algorithme….
Nouveauté radicale Logique inductive à partir des data	La nouveauté radicale des algorithmes par apprentissage automatique c'est qu'ils vont partir directement des faits pour induire les règles à appliquer. On passe de la logique déductive à une logique inductive. En soi où est le mal ? Nous pratiquons nous-mêmes ce type de logique : lorsqu'un même fait se répète on est tenté de penser qu'il va se reproduire. Par exemple, je sais qu'il arrive fréquemment que telle personne arrive en retard, j'agis en conséquence MAIS je sais que ce n'est pas certain : cette fois, elle peut être à l'heure. Où est le problème, sinon forcément le mal, c'est que l'algorithme va élaborer lui-même ses règles à partir des faits. En restant dans l'exemple de la remise de peine, il va d'abord remplacer l'étude des sociologues.

Traduction automatique

Pour illustrer cette évolution dans la méthode des algorithmes, prenons un exemple qui ne soit pas trop chargé d'humanité, celui de la traduction automatique d'une langue dans une autre.

1ère approche : Dès les années 50, des linguistes et informaticiens, pour élaborer un algorithme de traduction, prennent un dictionnaire, une grammaire dont ils extraient des règles pour construire un modèle (cf. les travaux de Jacobson, Noam Chomsky, …). Ça donnera SYSTRAN. Mais il reste beaucoup de problèmes (exemple, the « flesh is weak » est traduit « par la viande est molle »).

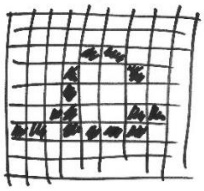
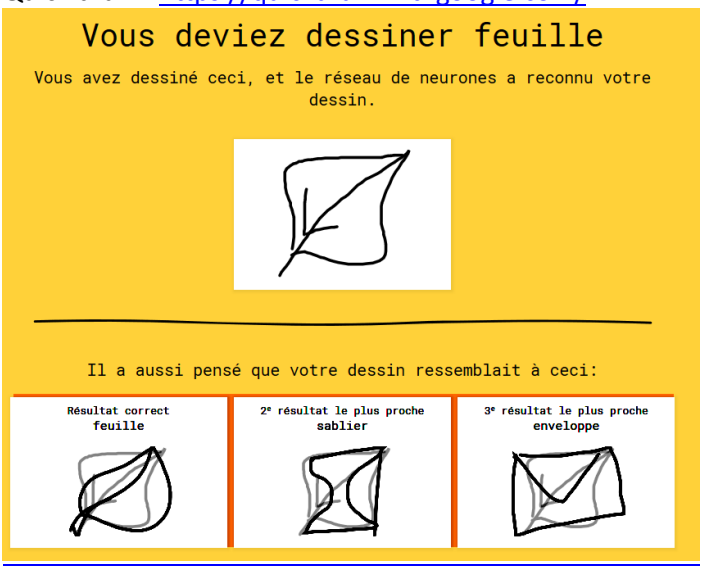
2ème approche : à partir de 80, au lieu de travailler comme un traducteur, la machine, recevant une phrase à traduire, va plonger dans des « corpus » bilingues (dans le cas de l'anglais, provenant du Canada ou de l'Union Européenne) pour trouver la phrase qui est la plus fréquemment trouvée. Aujourd'hui, dans GoogleTrad, « flesh is weak » est traduit par « la chair est faible ») : c'est l'exploitation de la « force brute » de l'ordinateur, rapidité d'exploration du fonds de référence + un peu de statistique. (voir différence avec le Gaffiot où le mauvais élève cherche une bonne traduction alors que le programme va choisir le traduction la plus fréquente).

3ème approche : en 2014 : application des réseaux de neurones artificiels : on transforme les textes en vecteurs sur lesquels on effectue les traitements (un pas de plus dans la numérisation !). On fait encore lire à la machine des millions de pages dans les deux langues mais le but n'est plus de traduire mais d'apprendre à traduire. Une fois l'apprentissage réalisé, on n'a plus besoin des bases d'apprentissage, la machine traduit selon le « modèle » (« pattern ») qu'elle s'est fait elle-même (ou presque[1]) par son apprentissage…. Modèle qui n'a rien à voir avec celui que recherchaient les linguistes-informaticiens dans les années 50

Et ça fonctionne ! Google, Systran, Facebook, DeepL…les grandes entreprises de traduction automatiques sont passées ou sont en train de passer à cette nouvelle méthode.

(Pour en savoir plus sur http://www.lemonde.fr/sciences/article/2017/11/27/la-traduction-dopee-par-l-intelligence-artificielle_5221041_1650684.html#lwRTavLiT7atL4dL.99)

[1] En fait, l'apprentissage est « supervisé » par des experts du domaine mis en apprentissage

L'apprentissage au XXème siècle DRAGON OMNIPAGE (OCR) ou plusieurs (gratuit) en ligne ex : https://convertio.co/fr/ocr/)	<u>Exemple de la reconnaissance vocale</u> : Au départ, dans les années 70, c'est le logiciel DRAGON qui permet de dicter un texte à un calculateur qui va l'écrire dans un fichier de traitement de texte, word par exemple. (L'apprentissage initial se fait par lecture d'un texte idoine, qui va s'enrichir au fur et à mesure des erreurs corrigées par le dictant. Utilisation très fréquente en médecine pour enregistrer les comptes rendus d'examens ou d'intervention). <u>Exemple de la reconnaissance d'images</u> : une image = un grand damier dont chaque case (pixel) est noire ou blanche. Il est facile de passer du damier à une matrice (de « 1 » et de « 0 ») et de la comparer à toutes les autres matrices de références (**des caractères reconnus**) pour trouver la plus proche. Cf OMNIPAGE 
L'apprentissage post 2010 deep-learning, machine-learning Quick draw Apprentissage supervisé Ne marche pas à tous les coups : « faux positifs » et « faux négatif »	Quick draw : https://quickdraw.withgoogle.com/  Supposons que l'on veuille apprendre à la machine à reconnaître un tigre parmi toutes les images d'animaux qu'on lui donne à voir. On lui donne toutes les images de tigre qu'on peut trouver, dans toutes les positions, de toutes les variétés. Après quoi, on lui présente une image de chat ou de lynx. S'il a bien appris, il « retoque » chat et lynx. Mais peut-être qu'il va retoquer aussi un tigre blanc, suffisamment rare pour n'avoir pas été présent dans le fichier d'apprentissage. Si, plutôt que d'utiliser ces systèmes de reconnaissance on voulait écrire un algorithme « à la main » comme dans DRAGON ou OMNIPAGE, il faudrait une telle quantité d'instructions que l'objectif serait inateignable.

Les réseaux de neurones	Cf NOTES p.3
-un bon niveau en maths ~~en biologie~~ -le procédé s'applique à une grande variété de domaines, -ça marche mais les créateurs ne savent pas toujours ni pourquoi ni comment ! (Il y a comme cela des domaines des mathèmatiques où subsistent quelqu mystère cf. par exemple, les automates cellulaires de J.Conway) https://www.youtube.com/watch?v=S-W0NX97DB0 http://jean-paul.davalan.pagesperso-orange.fr/divers/jeuvie/index.html	Une présentation par l'un de ses inventeurs, professeur au collège de France, Yann LeCun : www.youtube.com/watch?v=OzZoPVjv8iE Plus technique : https ://www.imagile.fr/creation-dun-petit-reseau-de-neurones-artificiels-netait-complexe/

« C'est fascinant de voir que cette technique, qui reste encore opaque et mal comprise, fonctionne aussi bien », constate François Yvon[2] ».

On peut avoir une bonne illustration des possibilités et de ces réseaux de neurones artificiels avec le film « Alpha-GO » visible sur NETFLIX qui retrace la préparation du programme qui, en 2017, a battu le champion du monde du jeu de GO, Se-Dol Lee. (Netflix offre un mois d'essai gratuit : ça peut servir à ça !)

En 1997, un ordinateur, Deep Blue, battait Kasparov, champion du monde des échecs. En 2017, le programme Alpha Go battait Se-Dol Lee, champion du monde de Go.

-IBM en 97, Google en 2017,

-Un ordinateur de 1,5T en 97, un programme dans un portable en 2017,

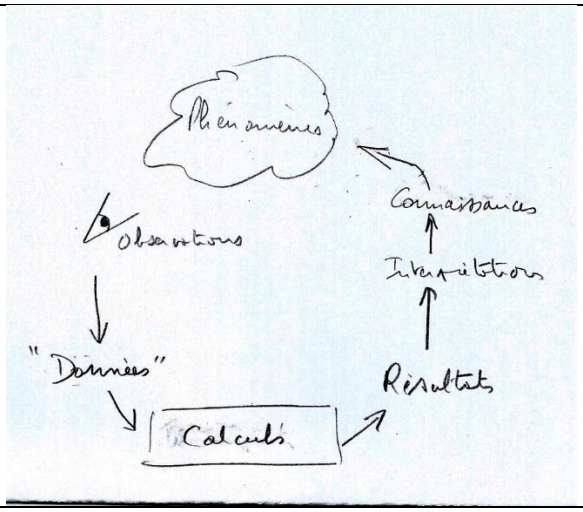
-Exploitation de toutes les parties jouées en 97, apprentissage du modèle du jeu de Go en 2017,

-étonnement du champion devant les coups inventés par le programme mais l'étonnement tout autant des ingénieurs sidérés des réactions de leur programme.

-Lorsque l'on dit qu'un algorithme est incapable de créer, il faut s'entendre sur ce qu'est une création : il ne peut créer à partir de rien mais, ayant appris un « modèle » il est capable de produire des résultats (ou, ici, des coups) auxquels nul n'avait encore pensé.

[2] François Yvon, directeur du Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Limsi-CNRS) à Orsay (Essonne)

Mais l'apprentissage se fait à partir des données : ce sont elles qui vont déterminer sa qualité.

Les Données Méfaits de l'école Les données de sont pas données -pas évidentes, -pas disponibles aisément -pas gratuites.	A l'école on nous a appris à recevoir des « données ». Elles étaient indiscutables, intangibles, il fallait les utiliser toutes, … si on était bon, elles permettaient d'atteindre un résultat. La vie nous a appris autre chose Les données sont des constructions, elles sont rarement évidentes, elles coûtent parfois très cher à produire. Elles donnent une vision du monde selon un certain point de vue. Elles ne sont pas le monde pas plus qu'une carte géographique n'est le territoire.
Cercle de la connaissance **observatio**n : point de vue, acuité**, …** **Données :** nombreux choix techniques **Calculs** : choix des méthodes et orga. **Résultats** : choix des représentations **Interprétation** Connaissances	

	À côté	Au-dessus	Dans	Au-dessous
Exemples	Médiamétrie, Google Analytics, affichage publicitaire	PageRank de Google, Digg, Wikipédia	Nombre d'amis Facebook, Retweet de Twitter, notes et avis	Recommandation Amazon, publicité comportementale
Données	Vues	Liens	Likes	Traces
Population	Échantillon représentatif	Vote censitaire, communautés	Réseau social, affinitaire, déclaratif	Comportements individuels implicites
Forme du calcul	Vote	Classements méritocratiques	*Benchmark*	*Machine learning*
Principe	***Popularité***	***Autorité***	***Réputation***	***Prédiction***

Cardon, p. 18

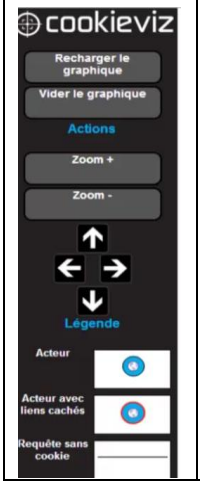
Quelles données ?	Tous les algorithmes dont nous parlons (par exemple, ceux du tableau donné par la CNIL) ont besoin de données mais, selon le type, ces données seront de natures très différentes. Pour APB, le système d'information sur les élèves, les vœux des bacheliers, les préférences des formations. Pour un GPS, une cartographie détaillée et si possible à jour, pour prédire les crimes et délits, les bases de données de la police et de la justice, etc. Pour les systèmes prédictifs ou prescriptif internes à une entreprise ou à une administration, ce sont les bases de données internes. Ce qui va importer, c'est la taille de ces bases : par exemple, Amazone a suffisamment de clients pour faire des suggestions d'achats, Les Volcans ne pourront pas….
« Quelle étude faire » devient « quelles data » donner à la machine.	Ex : satisfaction et souhait des adhérents des RPC ? Entretiens, enquêtes (face à face, par courrier, comptages stat…) Observation de l'accès au site web - Accès à Facebook, Amazone, Google, médiathèque, … EN….
Les données deviennent la ressource essentielle : d'où les slogans du type : l'or noir de demain Cf l'ouvrage : **L'empire des données** (Adrien Basdevant) à l'origine **d'un « coup data** » !)	

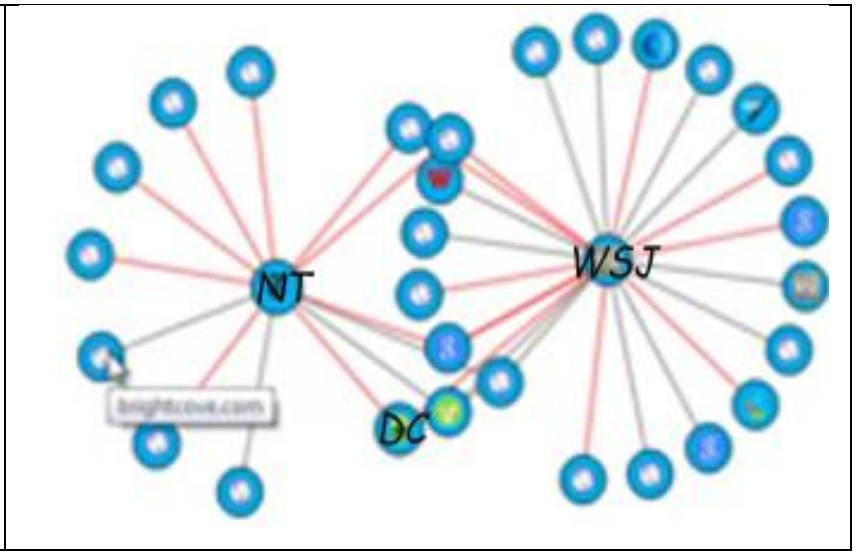
L'appariement de fichiers Ingénieurs sociaux, Société rationnelle	Le graal c'est de pouvoir rassembler sur une même personne des données venant de plusieurs sources différentes : santé (SNIIR), revenus, adresse, sexe et âge, …..C'était déjà l'idée, dès les années 30, d'un certain nombre « d'ingénieurs sociaux » dont l'exemple le plus notable est celui de René Carmille. René Carmille, après polytechnique, est contrôleur général de l'armement. Il s'intéresse aux fichiers de population utile à l'organisation de la conscription et de la mobilisation en cas de besoin. A ce titre il examine sur place l'utilisation des machines mécanographiques à cartes perforées qui sont en service dans les grandes entreprises américaines. Au retour, son projet est d'introduire cette technologie dans l'administration française : son grand projet c'est de rassembler derrière le même identifiant d'une personne toutes les informations dont disposent les différents services administratifs : démographiques, scolaire, justice, emploi, finance, … Il propose son projet au gouvernement de front populaire qui n'en veut pas, plus tard il le propose au gouvernement de Vichy qui n'en veut pas davantage mais, comme il a été nommé directeur du Service National de la

Liberté individuelle, Informatique et liberté	Statistique (SNS), il impose dans le fichier de l'état civil un **identifiant à 13 chiffres** qui sera adopté par la sécurité sociale en 1946. A cette date, le numéro n'est utilisé que par l'INSEE qui a succédé au SNS (pour le répertoire national des personnes physiques et le fichier électoral) et la sécurité sociale pour sa gestion interne (maladie et retraite)[3]. On est donc loin du projet initial de Carmille de rassembler toutes les informations administratives derrière cet identifiant unique Ce répertoire national (nom, prénom, date de naissance, identifiant) étant sur cartes perforées, la question s'est posées fin des années 60 de passer de la mécanographie à l'informatique …. une technologie qui offrait de nouvelles possibilités. De façon non complétement élucidée, l'idée que ce fichier pourrait servir à rassembler des informations provenant de sources différentes fit son chemin[4]. Tant et si bien qu'en 1974, un article du journal « Le Monde » titrait : « SAFARI ou la chasse aux français » dans lequel son auteur, Philippe Bouchet, dénonçait le secret qui entourait le projet conjoint de l'INSEE et du ministère de l'intérieur de « ficher » l'ensemble de la population. L'affaire fait grand bruit, les débats sont nombreux qui aboutiront en 1978 à la Loi Informatique et libertés qui encadre très strictement les rapprochements de fichiers de personnes. En particulier, elle interdit de rapprocher des fichiers à des fins commerciales ou politiques.
Les cookies **Cookies tiers**	**Il fallait donc que le marketing trouve autre chose….** Quand je visite le site web philo63, quand je lis la page sur Montaigne, le site ne sait pas si j'ai d'abord été lire la page sur Erasme. Pour des sites qui font du e-commerce, ce serait très gênant. De ce fait, a été inventé un fichier (dénommé cookie) qui va, en particulier, contenir un « panier » dans lequel vont venir s'inscrire les achats effectués au fil des pages parcourues. Pour livrer la commande, le site de commerce va demander des informations de livraison et, quand le visiteur va quitter le site, le cookies va être récupéré par le commerçant. Et, comme le client va peut-être revenir, le cookies va être laissé sur l'ordinateur où il a été créé. En allant un peu plus loin, le commerçant ou media va essayer de « fidéliser » le client, lui faire ouvrir un « compte » en essayant de lui faire livrer des informations complémentaires. Où ça se complique c'est quand des sociétés, des régies publicitaires[5], viennent installer leurs cookies (dits « cookies tiers ») pour en faire commerce. Par ce système, chaque site affilié sait ce qu'a fait chacun de ses clients sur tous les sites affiliés à la même régie. La régie publicitaire se constitue ainsi une banque de données rapidement géante (big data !). C'est par ce système qu'après avoir commandé une perceuse électrique sur un site de bricolage je vais retrouver une publicité pour perceuses électrique lorsque je vais consulter ma messagerie sur le portail d'orange. Et il y a des mécanismes encore plus « vicieux » du genre augmentation du prix lorsqu'un visiteurs revient sur le site qu'il avait

[3] Voir wikipedia pour en savoir plus sur René Carmille.
[4] Le projet a pris le nom de *Safari* (Système automatisé pour les fichiers administratifs et le répertoire des individus).
[5] Ces régies sont connues sous le terme de « ad-network » : weborama, double-click, critéo, …

	d'abord consulté avant de rechercher, par exemple, un meilleur tarif. **En fait, l'inventivité des gens de marketing paraît sans limites !** Ces dispositifs sont le moyen de rassembler sur un même utilisateur des données qui proviennent de différentes sources : faute de pouvoir utiliser un identifiant commun style n°insee/SS (ce qui est interdit) l'appariement se fera sur un nom, un pseudo, un numéro de carte bleu, une adresse IP-, …)[6]. Nos consommations en disent long sur nous-mêmes et, même si nous n'achetons rien sur Internet, la liste des sites et des pages que nous consultons en disent long sur nous-mêmes.





Exemple : soit une visite du site du Wall Street Journal (WSJ sur la carte). Il est en relation avec une vingtaine d'autres sites pour leur signaler que l'on vient d'arriver chez lui.. Si, dans la même session on consulte le New-York Time (NT sur la carte), celui-ci, à son tour va signaler à une dizaine de sites que l'on vient d'arriver (dont quelques-uns communs au WSJ). Il y a un site commun particulier, celui de la régie publicitaire Double-Clic (noté DC sur la carte[7]) qui va communiquer à ses deux clients NT et WSJ (et à tous les autres intéressés) tous les cookies du visiteur.
Voir le site de la CNIL :
https://linc.cnil.fr/fr/cookieviz-une-dataviz-en-temps-reel-du-tracking-de-votre-navigation

Les biais dans les données **Biais =/ erreur**	Une donnée peut être erronée : le bachelier a eu 18 en physique mais c'est 13 qui a été écrit dans la base de données. C'est une **erreu**r de lecture, de saisie, de traitement, …Elle est le plus souvent ponctuelle. Le **biais** désigne une erreur de méthode : penser que tout le monde se connecte sur Internet par exemple.

[6] L'imagination des gens du marketing dans l'utilisation du web est impressionnante … et il n'y a pas que du marketing commercial. Voir le dossier CNIL sur la publicité en ligne
https://www.cnil.fr/sites/default/files/typo/document/Publicite_Ciblee_rapport_VD.pdf
[7] Racheté par Google en 2007.

Les erreurs les plus graves imputées aux algorithmes proviennent de biais dans les données d'apprentissage	-Apprendre à traduire une langue à partir des textes de l'Union Européenne fait l'impasse sur l'existence de la première personne. - dans un système de recherche d'emploi, on s'aperçoit que le système propose des emplois moins bien payés aux femmes plutôt qu'aux hommes. -dans un système (US) de prescription de la liberté conditionnelle, les personnes de couleurs sont moins bien traitées que les blancs, -tag « gorille » attribué automatiquement à des photos d'identités de personnes noires, ……(rapport cnil p.32) Ce n'est pas le mécanisme d'apprentissage qui est défaillant mais bien les données que leurs concepteurs leur ont mis en apprentissage Cf.Cathy O'Neil : « Weapons of math destruction » (Armes de destruction mathématique)
La nécessité d'un prétraitement A supposer que les sources de data soient pertinentes, il faut des traitements préparatoires pour qu'elles soient utilisables.	Avant qu'un jeu de données soit utilisable par l'algorithme, il faut pratiquement toujours opérer des traitements de mise en forme, de tri, de codage, de correction, etc. Par exemple, des chercheurs travaillant sur le SNIIRAM[8] viennent de mettre en évidence une corrélation entre un médicament antidiabétique et l'apparition d'un cancer de la vessie. Après analyse par les médecins, il y a reconnaissance d'un lien de causalité. Ce type d'information qui n'avait pas été mise en évidence jusque-là peut avoir une grande importance et les résultats de l'étude incitent à exploiter ce SNIIRAM à grande échelle pour explorer des corrélations de ce type. Mais, comme cette base de données existe depuis près de 40 ans, on peut se demander pourquoi elle n'a pas été explorée plus tôt. La réponse donnée dans l'article relatant cette découverte (Le Monde du 30 janvier 2018) est que la base de données a été conçue pour gérer les remboursements de l'assurance maladie et pas pour faire de la recherche épidémiologique. **Pour cette recherche il a fallu un travail préparatoire de six ingénieurs pendant deux ans**. Ces traitements préparatoires sont décidés par les responsables de chaque projet. Modifiant plus ou moins profondément le fichier d'apprentissage ils modifient les résultats fournis par l'algorithme. On peut vouloir « redresser » le fichier pour qu'il soit plus conforme à ce que l'on sait du phénomène décrit mais on peut aussi le « déformer » au profit de telle ou telle politique.
Les big-data **(**mega données, données massives, …) Clics, likes, pages, liens, …	Au sens précis, il s'agit de méthodes (et donc de programmes) qui ont été conçues pour traiter les masses de données générées par l'utilisation d'Internet, des clics, des likes, des traces de navigation (pages et liens), ….des données souvent hétérogènes (textes, adresses, images, sons, vidéos, …) le tout à traiter rapidement. Ceci a obligé à trouver de nouvelles façons de stocker

[8] Le **Système national d'information inter-régimes de l'Assurance maladie** (**SNIIRAM**) est une base de données française maintenue par la Caisse nationale de l'assurance maladie des travailleurs salariés. Il contient les données correspondant à plus d'un milliard de feuilles de soins par an.

Data mining	l'information (bases noSQL) et à utiliser de nouvelles procédures d'exploration de données (data mining).
Nouvelle approche ? Avant : -hypothèses, -système d'observation, -collecte et contrôle, puis analyse pour vérifier ou non les hypothèses de départ. Avec le big data : -trouver des data, -mettre en forme, -lancer les programmes, -obtenir les résultats (?)	Ainsi comprises, les big data sont à la source d'une nouvelle approche de la méthode scientifique traditionnelle. Auparavant et jusqu'ici, lorsqu'il s'agissait d'explorer un domaine, on formulait des hypothèses, mettait en place un système d'observation, collectait et contrôlait les données obtenues. On pouvait alors entreprendre une analyse par laquelle on testait les hypothèses de départ. Prenons un exemple : dans les années 90, la confection en France est en crise, les entreprises ferment, le commerce se concentre, les marques étrangères prennent des part de marché de plus en plus importantes. Le CNIS demande à l'INSEE de faire une enquête auprès de la population pour mieux savoir qui achète quoi, où, à quel prix, avec quelle fréquence, etc. On réunit des représentants des différents acteurs du domaine pour dégager la liste des points à observer …. On fait des hypothèses dont on va déduire des questions à poser et à traiter. Cela aboutira à un questionnaire d'enquête. Après quoi on va définir la population que l'on va interroger, « construire » un échantillon à la fois représentatif de la population totale mais suffisamment vaste pour donner des résultats fiables sur des aspects peu fréquents. Passons les phases de collecte et de constitution du fichier des réponses pour aller à la phase de calcul des résultats. Supposons qu'on s'intéresse à l'achat de chaussettes pour homme. Dans le fichier des résultats on sait qu'une personne de sexe féminin, factrice, née en 1987, demeurant à Clermont-Ferrand a acheté une paire de socquettes pour homme de marque DIM à Monoprix à 15€…. Pour organiser ces données détaillées et produire des résultats compréhensibles, on va créer des catégories : une catégorie « chaussettes pour homme » dans une nomenclature de produits, une catégorie « moyenne surface généraliste» dans une nomenclature de points de vente, une catégorie « employé fonction publique » dans une nomenclature de catégorie socio professionnelle, etc. De la sorte, le calculateur va pouvoir classer les informations élémentaires dans ces différentes catégories. Par exemple, donner le nombre de personnes de sexe féminin qui ont acheté des chaussettes selon le point de vente, le prix, l'âge et la catégorie socio professionnelle de l'acheteuse. Avec le big data, on va d'abord rechercher les fichiers qui contiennent les traces des achats. Par exemple, les fichiers des tickets de caisse des grandes surfaces avec, si possible, le fichier des cartes de fidélité. On se fait un fichier rassemblant les tickets de caisse par client et on demande aux programmes de « data mining » de le faire parler : ils vont construire eux-mêmes avec des algorithmes de classification automatique des