ÉCOLE DE GUERRE

PROMOTION *GÉNÉRAL GALLOIS*

2016 -2017

Les enjeux du *data mining*

CBA Daniel ZENI

Sous la direction de :

M. Axel Le Poupon

Docteur en communication et électronique

« Les idées et opinions émises dans ce mémoire n'engagent que la responsabilité de son auteur et ne reflètent en aucun cas celles de l'Ecole de Guerre. Par ailleurs, ce mémoire ne peut être diffusé en partie ou dans sa totalité sans l'accord préalable de son auteur»

Sommaire

La multiplication exponentielle des données est la conséquence logique de la convergence de deux facteurs. Le premier, technologique, a permis l'essor du numérique en l'espace de trois décennies, lorsque le second, socio-économique, a visé la mutation de notre modèle en exploitant le premier, selon le principe de la transformation numérique. Aujourd'hui cette digitalisation de la société, d'abord occidentale, se diffuse à l'ensemble de la planète, aidée par la volonté des grandes multinationales du numérique soucieuses de leurs intérêts et conscientes de leur puissance et leur primauté dans un monde davantage interconnecté. Cette expansion ne pourrait se faire sans un public séduit par la plus-value quotidienne apportée par les nouvelles technologies de l'information et de la communication (NTIC) : accès à l'information, échanges et services facilités, interconnexions démultipliées …

Face à cette multiplication des acteurs et utilisateurs, l'homme doit être capable de gérer de la façon la plus efficace ce flot de données pour en extraire l'information utile au citoyen ou à une organisation : c'est bien l'enjeu du *data mining* d'exploiter ces *data*, de trouver les mécanismes techniques et statistiques pour extraire l'information pertinente, l'enrichir, lui donner du sens, du relief en l'insérant dans des schémas et concepts de causalité voir de prédiction. Maîtriser le *data mining* permet donc de donner du sens, de la clarté à notre quotidien particulièrement lorsque l'infobésité à tendance à l'obscurcir.

Dans ce cadre, les Armées comme les autres organisations doivent anticiper, adapter et moderniser leurs outils pour en tirer pleinement parti mais en conservant cette capacité de résilience supplémentaire attendue d'un ministère régalien qui doit demeurer l'*ultima ratio regum*. A ce titre, l'utilisation du *data mining* permettrait, par simple transfert de technologie, l'amélioration de nombreuses fonctions duales présentes aussi bien dans la société civile qu'au sein de la Défense. De manière complémentaire, de par leurs spécificités, certaines fonctions plus complexes voire exclusives comme le renseignement pourraient engrangées les dividendes de cette infobésité croissante dans le cyberespace.

L'objectif de ce mémoire réside bien dans l'explication de ces phénomènes : de la génération de « toujours davantage de données » jusqu'à une exploitation poussée pour garantir la sécurité des citoyens, aider les décideurs de demain voire anticiper de nouvelles tendances.

Enfin ce tour d'horizon s'attachera également à dresser les problématiques inhérentes à la donnée en elle-même et à son exploitation pour davantage de cohérence.

Table des matières

Figures

Acronymes utilisés

Ces acronymes seront développés à nouveau lors de leur première apparition dans le corps de ce mémoire. Ils sont également regroupés ci-dessous par ordre alphabétique pour permettre au lecteur un autre accès, plus rapide. Les acronymes et les mots en italique correspondent à des termes anglo-saxons communément utilisés dans leurs domaines.

ANSSI : Agence nationale de sécurité des systèmes d'information
BSS : bande sahélo-saharienne
CMI : Cellule de management de l'information
CNIL : Commission nationale de l'informatique et des libertés
DAS : délégation aux affaires stratégiques
DCP : Données à caractère personnel
DHS : Department of Homeland Security
DRHAT : Direction des ressources humaines de l'Armée de Terre
DRSD : Direction du renseignement et de la sécurité de la Défense
EPS : Etude prospective et stratégique
FBI : Federal Bureau of Investigation
GAFAM : Acronyme de Google, Amazon, Facebook, Apple, Microsoft
HUMINT : Human Intelligence
IDS: Information detection system
IPS: Information protection system
IEC: International Electrotechnical Commission
ICANN: Internet Corporation for Assigned Names and Numbers
IKM: Information and Knowledge Management
IMINT: Imagery Intelligence
IoT : Internet of Things
MCO : Maintien en condition opérationnelle
NBIC : nano et biotechnologies, intelligence artificielle et sciences cognitives
NTIC : les nouvelles technologies de l'information et de la communication
OIV : opérateur d'importance vitale
PAM : plan annuel de mutation
R&D : Recherche et développement
RFID : Radio frequency identification

RH : Ressources humaines

RMA : Revolution in Military Affairs

ROC : Renseignement d'origine cyber

ROHUM : Renseignement d'origine humain

ROEM : Renseignement d'origine électromagnétique

ROIM : Renseignement d'origine imagerie (ROIM)

SCORPION : Synergie du Contact Renforcée par la Polyvalence de l'infovalorisatiON

SECOPS : Sécurité des opérations

SA : Situation awareness

SI : Système(s) d'information

SIEM : Security information and event management

SIGINT : Signal Intelligence

SNR : Stratégie Nationale de Recherche

UE : Union Européenne

Remerciements

En préambule de ce mémoire, je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide, leur soutien et qui ont contribué de près ou de loin à la réalisation de ce mémoire.

Je tiens à remercier tout particulièrement, Monsieur Axel Le Poupon (directeur du mémoire), pour son implication et ses conseils prodigués dans le cadre de cette étude. Son investissement sur cette thématique qui lui est chère ainsi que ses connaissances sur le cyberespace ont été particulièrement utiles et enrichissants. Sa disponibilité est à souligner et ce malgré de lourdes charges managériales et professionnelles.

Mes remerciements s'adressent également à l'équipe du centre de documentation de l'Ecole de Guerre pour leurs conseils dans l'élaboration du mémoire, la méthodologie à adopter et la documentation choisie sur cette thématique à cheval entre technique et sociologie.

Je tiens également à souligner ma reconnaissance à certains camarades de cette promotion pour les conseils et leurs échanges de points de vue sur des thématiques plus ou moins proches de mon sujet ainsi que sur leurs solides compétences acquises dans ce domaine dans leurs fonctions précédentes.

Merci à toutes et tous.

Introduction :

Dans son article « Données le vertige », Gabriel Siméon rappelait en 2012 que *« L'humanité produit autant d'informations en deux jours qu'elle ne l'a fait en deux millions d'années. L'avenir appartient à ceux qui sauront utiliser cette profusion »*[1]. Lorsque ces volumes de données réussissent à être traités en quasi temps réel, ils apportent de nouvelles avancées dans de nombreux domaines, à l'image du logiciel *HealthMap,* développé et utilisé pour l'anticipation et le suivi des crises sanitaires. Fonctionnant sur l'analyse des données de l'Organisation Mondiale de la Santé, des requêtes sur le moteur de recherche Google, des informations issues de Google News et de Twitter, ce logiciel *« a permis de suivre l'évolution d'une épidémie de choléra en Haïti avec près de deux semaines d'avance sur les observations des autorités de santé »*[2]. Cette capacité d'appréciation de situation déportée, délocalisée et en temps réel peut permettre dans ce cadre de contenir la propagation de l'épidémie voire d'agir avec d'avantage d'efficacité pour l'éradiquer tout en diminuant d'autant le nombre de victimes. Ce simple exemple prometteur laisse entrevoir les possibilités futures offertes par le *data mining*, compris comme *« l'extraction d'un savoir ou d'une connaissance à partir d'une grande quantité de données »*[3]. Il questionne cependant sur le traitement et la fiabilité des sources utilisées.

Par ailleurs ces collectes et exploitations de données nécessitent d'être encadrées quelques soient les finalités souhaitées. Dans son roman d'anticipation 1984, Georges Orwell décrivait une intrusion possible de *Big Brother* dans la vie des citoyens en portant atteinte aux libertés fondamentales et à la vie privée de chacun. En ce début de XXI^{e}, cette vision orwelienne que l'on pensait réservée à la science-fiction est en passe d'être réalisée dans le cyberespace avec l'avènement d'organisations assimilables à des *Big Brothers* numériques : les GAFAM[4]. Ceux-ci répondent en quelque sorte à cette définition par leur capacité à collecter des données produites par les utilisateurs ou clients pour en optimiser l'exploitation à des fins mercantiles. Ce travail de *data mining* à l'échelle planétaire s'avère plus que concluant économiquement puisque Apple et Google constituent les deux plus grosses capitalisations mondiales, mais

[1] http://www.liberation.fr/futurs/2012/12/03/donnees-le-vertige_864585
[2] Idib
[3] Wikipedia - recherche sur le terme *data mining*
[4] GAFAM : Acronyme utilisé pour évoquer les grandes entreprises américaines du numérique que sont : Google, Amazom, Facebook, Apple, Microsoft.

interroge sur la conservation et l'utilisation faite des données et donc sur la notion même d'espace privé.

Traiter du *data mining*, loin d'être un effet de mode, constitue un véritable sujet d'actualité et d'avenir, placé au rang de priorité immédiate dans la stratégie nationale de recherche (SNR) afin *«d'assurer notre place parmi les premières puissances de recherche mondiale et de mobiliser les énergies sur les défis scientifiques, technologiques, environnementaux et sociétaux du XXI^e siècle »*[5]. Ainsi le *data mining,* et plus généralement le *big data,* apparaît comme un des cinq *« enjeux à fort impact potentiel, […] devant être traités avec une urgence particulière, compte tenu de la diversité de leurs impacts économiques et sociaux, des dynamiques internationales en cours, et de la maturité des actions envisagées :*

- *L'explosion du volume de données numériques dans l'ensemble de la société et des domaines de la science, qui représentent un gisement exceptionnel de connaissances nouvelles et de croissance économique »*[6];

D'ailleurs, les précédents exemples liés à l'utilisation des données, certes aux finalités et acceptabilités différentes, témoignent de l'importance et de la sensibilité croissantes accordées aux *data* dans des sociétés de plus en plus numérisées et dépendantes d'Internet. Ils montrent également l'intérêt porté par tous les acteurs de la société pour le *data mining* et le *big data*, terme plus général pour traiter ce qui touche à l'explosion de données. Ainsi les Etats, les organes régaliens, les multinationales, les organisations internationales, jusqu'aux citoyens, ont tous un intérêt dans le développement du *data mining*. En revanche, cette appétence se décline de plusieurs manières, avec des priorités et des approches différentes, parfois divergentes voire opposées suivant l'acteur concerné. Il paraît donc légitime de comprendre les enjeux liés à cet usage du *data mining* et de s'interroger sur ses modalités d'utilisation dans la société et plus particulièrement pour la Défense.

Au final, maîtriser cette croissance exponentielle des données passe par l'utilisation et l'adoption du *data mining*. C'est pourquoi il est indispensable que la Défense poursuive

[5] Stratégie nationale de recherche – France Europe 2020 - http://cache.media.enseignementsup-recherche.gouv.fr/file/Strategie_Recherche/26/9/strategie_nationale_recherche_397269.pdf

[6] Idib

sa transition digitale et intensifie son recours au *data mining* , notamment pour assurer la résilience de l'Etat. En effet son utilisation permettrait :

- **d'accroître la sécurité et la sureté en général dans ce cinquième milieu grâce notamment au renseignement numérique,**
- **de renforcer la présence et le contrôle de la couche cognitive, véritable champ d'affrontement idéologique,**
- **d'améliorer l'efficacité des pratiques numériques et donc l'efficience des services en général.**

Démontrer ces apports possibles du *data mining* oblige tout d'abord à comprendre plus en profondeur ce phénomène d'explosion des données, à travers les causes, les problématiques associées et les conséquences de manière globale. Ensuite, ce phénomène en accélération avec l'arrivée des objets connectés, nouveaux pourvoyeurs de *data*, oblige concomitamment, à trouver des solutions efficientes de traitement, d'exploration et d'analyse afin de pouvoir donner du sens à ces *data* sous peine de rapidement se perdre dans « l'infobésité ». C'est bien là l'essence même du *data mining*. Cette compréhension globale des *data* demeure le préalable pour acquérir de la connaissance et ainsi pouvoir entrevoir les nombreuses perspectives offertes par le *data mining* en matière de description ou de prédiction. Enfin il conviendra de déterminer, pour la Défense, le cadre et les champs d'application possibles du *data mining* dans un exercice prospectif.

1. Une croissance exponentielle des données, non sans conséquence…

Comprendre les enjeux liés au *data mining*, nécessite au préalable de connaître les raisons essentiellement techniques de cette explosion des données en l'espace de deux décennies. Par ailleurs, cette évolution exponentielle des *data* amène en parallèle à s'interroger sur un certain nombre de problématiques en rapport avec le milieu d'évolution, qu'est le cyberespace. Enfin ce phénomène résumé par le terme englobant de *Big-data* amène déjà, souvent à notre insu, un certain nombre d'avancées et de changements concrets dans notre quotidien, qu'il conviendra d'exposer.

1.1 Les raisons de cette croissance exponentielle des données numériques

Cette explosion des data ne pourrait avoir lieu sans la conjonction de plusieurs facteurs nécessaires comme le développement des principales technologies numériques, l'amélioration continue d'Internet favorisant l'interconnexion d'utilisateurs de plus en plus nombreux et la multiplication de nouveaux acteurs connectés.

➢ <u>Les lois de Moore, Butter et Kryder pour expliquer techniquement cette évolution</u>

Ces données ne peuvent être générées et traitées en nombre que par les nombreux progrès scientifiques effectués dans la deuxième partie du XX^{e} siècle particulièrement dans les domaines de l'électronique et l'informatique. La miniaturisation des composants (*downsizing*) avec notamment l'apparition du circuit intégré, des processeurs puis des microprocesseurs, les progrès effectués dans la programmation, l'industrialisation de la fabrication et la baisse des matières premières permirent à partir des années 80-90 de démocratiser l'usage des ordinateurs jusqu'alors réservé en partie aux administrations. Dès lors la performance des ordinateurs n'a cessé de croître d'année en année, suivant la loi de Moore – le nombre de transistors sur un processeur double tous les deux ans – permettant ainsi un nombre d'opérations toujours plus élevé. Mais ce dernier

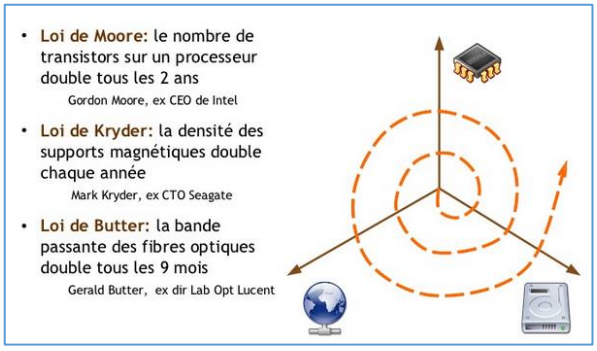


Figure 1: Différentes lois ayant un impact direct sur les data

n'aurait pas pu continuer à croître sans une augmentation en parallèle des capacités de stockage pour permettre ce traitement des données. La loi de Kryder – la densité des supports magnétiques double chaque année – traduit cette progression dans le domaine des supports magnétiques que sont les mémoires et autres disques durs mais également serveurs. Enfin, sans échange de ces données entre utilisateurs, le *data mining* resterait cantonné à un modeste moteur de recherche tout au mieux. C'est pourquoi l'avènement de nouvelles technologies dans les réseaux numériques (fibres, 4G…) avec des bandes passantes de plus en plus grandes ont permis d'exploser le débit de données transférées allant jusqu'à permettre actuellement le visionnage de vidéos haute définition en streaming par réseau cellulaire sans craindre de ralentissement dans la lecture. Cet essor des télécommunications – suivant la loi de Butter qui voit la bande passante doubler tous les 9 mois - a été rendu possible par les investissements conséquents des opérateurs de télécommunication, la normalisation des standards d'échanges numériques via l'*IEC*[7] ou encore *ICANN*[8] favorisant l'interopérabilité, les invitations et participations des Etats mais aussi des acteurs privés comme les géants d'Internet.

Ce triptyque : essors des télécommunications via internet, de la puissance de calcul et des capacités de stockage sont les principales raisons techniques ayant permis cette explosion des *data* dans notre quotidien. Elles expliquent en grande partie l'origine de cette 4^{e} révolution industrielle toujours en cours.

- L'humanité connectée d'ici 2030 : Les initiatives projet *Loon* et *Internet.org*

A la capacité technique de production de data s'ajoute une volonté politique et sociale d'ouverture d'Internet d'abord au plus grand nombre avec pour objectif final de relier l'humanité. Cette multiplication des acteurs physiques interconnectés - au travers d'Internet - contribue ainsi directement à l'accroissement exponentiel de données.

L'essor des télécommunications décrit supra, a permis de relier les continents (avec notamment les câbles sous-marins – cf. Figure 2) et démocratiser l'accès à Internet. Si cet essor n'a pu être réalisé par les pays eux même, il l'a été par les acteurs privés voyant dans ce besoin d'interconnexion de l'Homme, de futures perspectives de gains directs (clients) et

[7] *IEC: International Electrotechnical Commission*
[8] *ICANN: Internet Corporation for Assigned Names and Numbers*

indirects (usages des données d'un plus grand nombre d'internautes). Les projets *Loon* et *Internet.org*[9] témoignent de cette volonté ou de ce besoin de connecter l'humanité par les leaders de l'économie numérique que sont Google et Facebook. Ces initiatives privées viseraient à faciliter l'accès à Internet non seulement des pays en cours de développement mais aussi les pays les plus pauvres d'ici 2030. Elles réduiraient ainsi la fracture numérique et donc les inégalités à l'échelle de l'humanité en réunissant sur la toile non plus 3,2 milliards d'internautes[10] mais bientôt de 8 milliards. Cette « globalisation » des échanges de data s'accompagne également d'une modification de la consommation des données. Désormais les internautes disposant d'une connectivité suffisante consomment en direct - au détriment du téléchargement - permettant de fait d'améliorer la connaissance et le comportement en temps réel des utilisateurs. Cette évolution rend donc le *data mining* encore plus pertinent pour appréhender le comportement de chaque internaute.

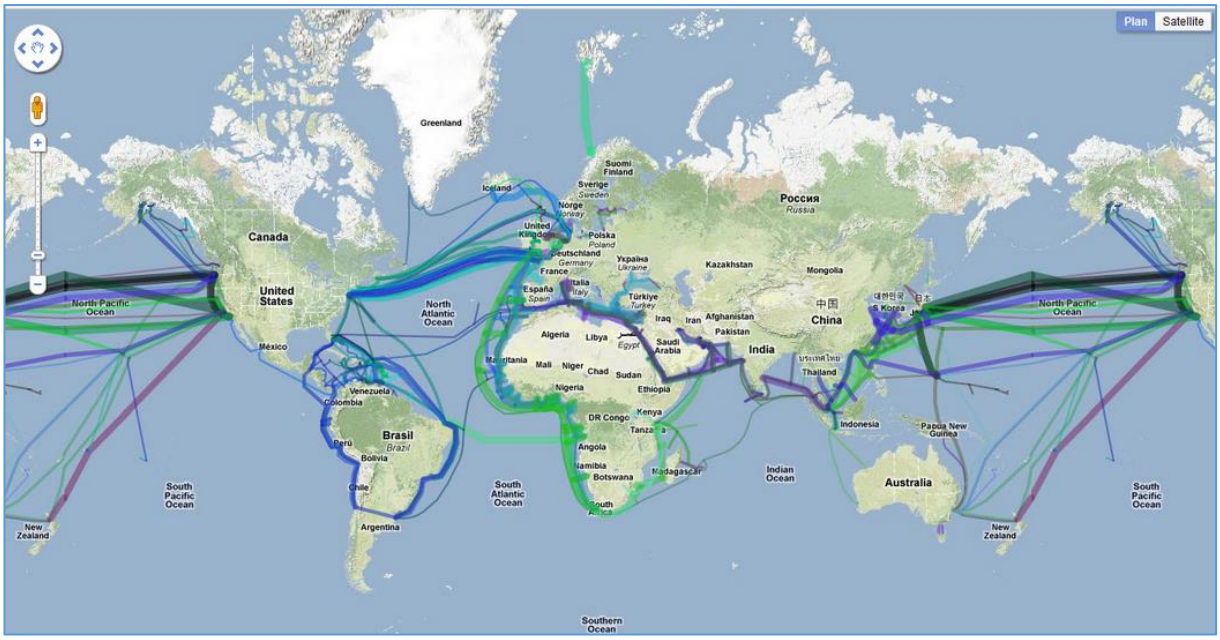


Figure 2: Carte mondiale des fibres optiques sous-marines (Source : Romain Decker)

- Le développement et la multiplication d'autres acteurs connectés : L'internet des objets (*IoT*[11])

[9] Projets *Loon* et *Internet.org* : Projets initiés respectivement par Google et Facebook pour connecter la population mondiale à Internet.
[10] Référence : http://www.itu.int/net/pressoffice/press_releases/2015/17-fr.aspx
[11] *IoT : Internet of Things*

Enfin à un nombre d'internautes humains toujours plus élevé vient s'ajouter de nouveaux acteurs connectés regroupés sous l'expression d' « internet des objets » ou Web 3.0. *« L'Internet des Objets est un réseau de réseaux qui permet, via des systèmes d'identification électronique normalisés et unifiés, et des dispositifs mobiles sans fil, d'identifier directement et sans ambiguïté des entités numériques et des objets physiques et ainsi de pouvoir récupérer, stocker, transférer et traiter, sans discontinuité entre les mondes physiques et virtuels, les données s'y rattachant. »* (Pierre-Jean Benghozi). Cette expression traduit donc la connexion croissante d'objets à Internet, connexion recherchée dans le but d'optimiser leur utilisation ou d'en maîtriser leur usage à distance via le téléphone portable par exemple. Grâce à une démocratisation de la technologie et une couverture Internet toujours plus grande, ces objets génèrent en permanence des flux de données conséquents au travers de protocoles de communication plus ou moins sécurisés. Cette technologie permet en temps réel aussi bien de gérer les caméras de son domicile pour une meilleure surveillance comme celles d'une ville comme Paris cherchant à davantage sécuriser ces arrondissements en complémentarité du travail de la préfecture de Police de Paris. On voit donc le panel des usages possibles allant du simple usage personnel à celui de l'optimisation de la gestion de l'énergie au sein des futures villes connectées via les *smart grids*[12]. Même si ces nouvelles technologies apportent de réels avantages elles peuvent encore trop facilement se retourner contre leurs utilisateurs par défaut de sureté, en atteste la dernière campagne d'attaques *DDOS*[13] en septembre dernier contre *« le numéro trois mondial de l'hébergement de sites internet et numéro un français, OVH »*[14]: A l' origine des hackers ont pris le contrôle d'objets connectés devenus des objets « zombies » par défaut de sécurisation (dans ce cas plus de 150 000 caméras), puis ont détourné leurs flux de donnés (supérieurs à 1 000 Gbit/s) pour les concentrer et saturer les serveurs d'hébergement d'OVH, paralysant ses services pendant la durée de l'attaque. Cet exemple démontre que les objets connectés au même titre que les internautes génèrent des *data* en conséquence. Leurs multiplications conduisent inéluctablement à une explosion de *data*.

---

[12] *Smart grids* : réseaux intelligents de distribution d'électricité utilisant des technologies informatiques d'optimisation de la production, de la distribution et de la consommation, et éventuellement du stockage de l'énergie, pour rendre plus efficient l'ensemble des mailles du réseau électrique, du producteur au consommateur final afin d'améliorer l'efficacité énergétique de l'ensemble en minimisant les pertes en lignes, en optimisant les moyens de production par rapport à la consommation, en temps réel (Source : Wikipedia)

[13] *DDOS : Distributed Denial of Service*

[14] Référence : http://lexpansion.lexpress.fr/high-tech/les-objets-connectes-nouveaux-relais-des-attaques-informatiques_1835475.html

1.2 Une explosion des *data* qui posent quelques risques techniques, juridiques et sécuritaires

Cette explosion de *data* circulant sur les réseaux pose un certain nombre de problèmes d'ordre technique liés au stockage et à la circulation de l'information notamment. En parallèle d'autres incertitudes juridiques et sécuritaires apparaissent avec la difficulté à matérialiser et appréhender les frontières intrinsèques du cyberespace mais aussi celles plus floues entre les données relevant de la sphère publique et celles du privé (contenus et contenants).

➢ Des risques techniques liés à l'explosion des données

Les trois lois citées supra (cf. 1.1) devraient nous laisser confiant et optimiste face à cette évolution exponentielle des données. Il convient malgré tout de prendre en compte certaines problématiques ou enjeux techniques qui commencent à apparaître avec cet essor des *data*, particulièrement vis-à-vis des capacités de stockage et de transfert.

En effet, la multiplication des données et leurs redondances (pour davantage de réactivité et éviter notamment toute perte technique et accidentelle de données) obligent à accroître le nombre de *data centers*. Or ces structures pour fonctionner correctement en continu, sans coupure, ont besoin d'énormément d'énergie :

- A la fois pour gérer le flux (sur les réseaux) et le stockage des *data*,
- mais aussi pour maintenir les équipements à bonne température, sans risquer de surchauffe voire d'incendie,
- enfin pour sécuriser 24h/24h et 7j/7 des locaux souvent extrêmement étendus.

A titre d'exemple, un des derniers *data centers* créés à Amsterdam par un des leaders du secteur s'étale sur plus de 17 000m^2 regroupe plus de 80 000 serveurs et consomme en électricité l'équivalent d'une ville *« comprenant entre 20000 et 50000 habitants »*[15]. Les *data centers* ont donc une énorme empreinte énergétique et calorifique, nécessitant de fait une proximité avec des sources de production d'énergie voire une capacité conséquente de refroidissement. C'est d'ailleurs pour ces raisons (économiques et écologiques) que l'on observe de plus en plus de sociétés placer leur dépôt de données près du cercle Arctique, dans

[15] Source : http://www.lemonde.fr/planete/article/2013/07/01/les-centres-de-donnees-informatiques-gros-consommateurs-d-energie_3439768_3244.html

les pays scandinaves, où l'énergie hydraulique demeure majoritaire et où les températures froides permettent un refroidissement quasi-naturel voire avec un faible apport énergétique. Ce choix génère un gain énergétique substantiel tout en diminuant les émissions de CO_2, en recourant à une énergie dite verte.

Malgré la loi de Butter explicitée précédemment (cf. 1.1), les flux de données croissent à un rythme supérieur à celui de la bande passante (de la fibre optique) laissant présager une saturation prochaine des réseaux (cf. Figure 3) : *« Without radical innovation in our physical network infrastructure—that is, improvements in the key physical properties of transmission fibers and the optical amplifiers that we rely on to transmit data over long distances—we face what has been widely referred to as a "capacity crunch" that could severely constrain future Internet growth, as well as having social and political ramifications»* (Richardson, 2010).

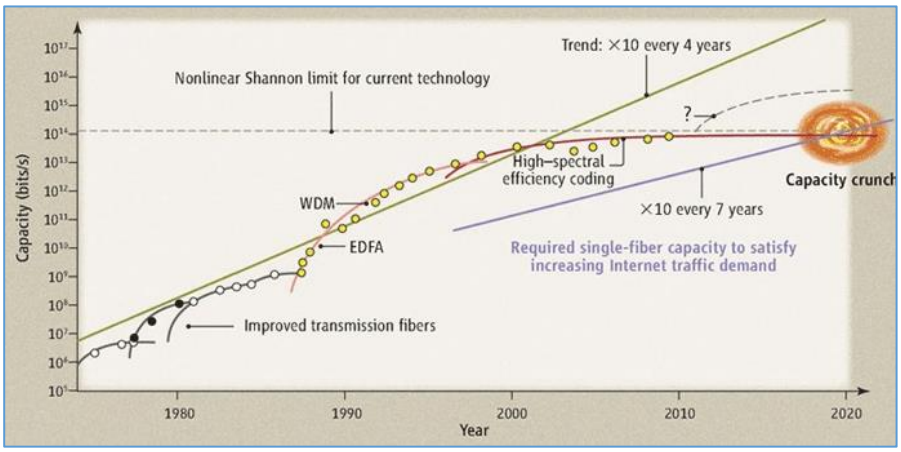


Figure 3: Evolution du débit en transmission en comparaison avec celui demandé (Source : Science 2010 - David J Richardson)

Dans les solutions envisagées, les opérateurs devraient mener des investissements très onéreux visant à doubler les réseaux actuels, ce qui n'est pas sans poser d'autres problèmes liés notamment cette fois ci à la consommation électrique. Pour l'instant les solutions retenues visent à optimiser la circulation sur le réseau actuel, en améliorant les protocoles, en répartissant au mieux la charge sur l'ensemble du réseau tout en mettant en place des bonnes pratiques, à l'image des régulations de vitesse imposées sur les autoroutes en fonction du trafic pour fluidifier la circulation et retarder l'apparition des bouchons.

- Territorialité des *data* et difficultés de définir les frontières du cyberespace

Avant d'exploiter une donnée il convient de connaître l'environnement légal dans lequel notre recherche s'effectue : *« L'idée d'un cyberespace libre, ouvert, où l'information circule sans restriction et où tout est accessible à tous à partir de n'importe où relève largement du*

mythe » (Amaël Cataruzza, 2014). Pour cela le cadre juridique est important puisqu'il restreint plus ou moins les possibilités d'exploitation en matière de *data mining* et donc la qualité de l'information obtenue . Comme souvent ce cadre juridique pourrait s'apparenter au droit du pays origine. Cependant dans le cyberespace et à l'heure de la mondialisation, cela devient nettement plus complexe puisqu'un simple mail peut transiter, entre l'émetteur et le récepteur, par de nombreux serveurs et donc de nombreux pays aux visions et intérêts juridiques différents voire parfois contradictoires dans l'usage des données.

Dans ce cas, quelle juridiction suivre pour leurs exploitations ? La plus favorable à notre étude, donc la moins liberticide ? Ou au contraire la plus protectrice des données des internautes ? Cette divergence d'appréhension juridique sur le cadre d'étude constitue un véritable enjeu pour des raisons à la fois sécuritaire, scientifique et économique. C'est pourquoi, les Etats-Unis ont très tôt compris l'intérêt de cette captation des données en étant le « grand aspirateur de données » :

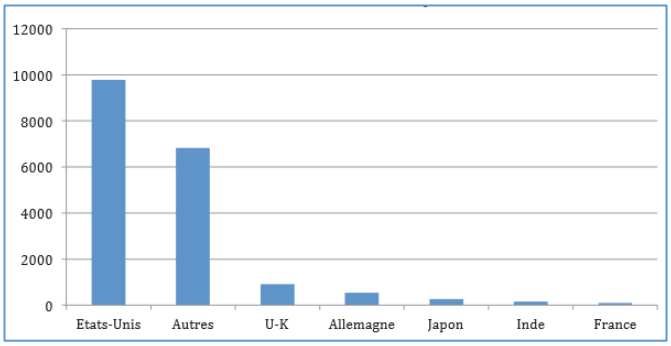


Figure 4: Répartition par pays du CA des entreprises du Top70 dans le domaine du Big Data (en M€)

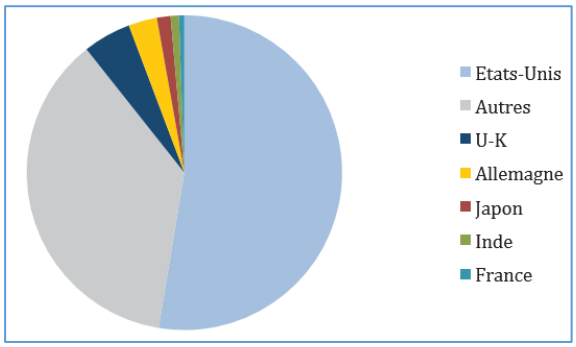


Figure 5: Répartition du top 70 des entreprises du Big Data par pays (Amaël Cataruzza, 2014))

« Comparant les données à un nouvel or noir, Stéphane Grumbach et Stéphane Frénot expliquent que « si des raisons géologiques expliquent la concentration des matières premières dans des régions particulières, des raisons économiques et politiques gouvernent leurs flux sur la planète. (…) La comparaison souvent faite avec le pétrole illustre parfaitement une caractéristique essentielle de l'économie des données personnelles : la concentration ». Dans ce cas, la localisation des données n'est pas fixée par des facteurs naturels mais par les concentrations économiques et les environnements règlementaires » (Amaël Cataruzza, 2014).

La carte ci-dessous réalisée par la chaire Castex de Cyberstratégie démontre bien la capacité des Etats-Unis à concentrer cette captation renforcée par les compétences extraterritoriales de la législation américaine.

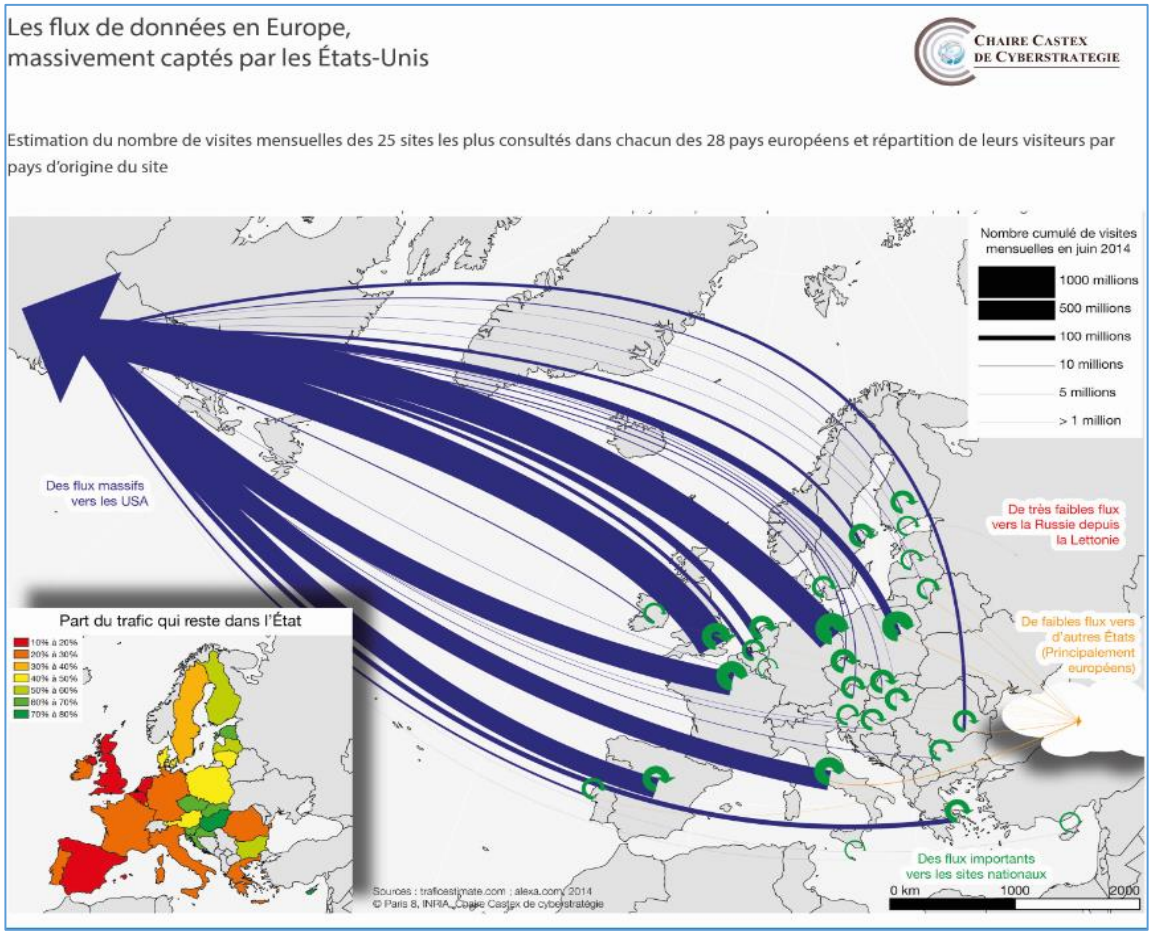


Figure 6 : Flux de données en Europe massivement captés par les Etats-Unis (Source : Castex)

« Plusieurs exemples ont montré combien la législation américaine permettait aux Etats-Unis de capter et de mobiliser un grand nombre de données collectées et stockées par les grands opérateurs américains. Les compétences extraterritoriales de la législation américaine constituent un véritable enjeu pour l'Union européenne en matière de protection de ses citoyens et bien sûr de souveraineté. Cette dynamique recouvre également des enjeux économiques, notamment l'émergence d'une offre européenne en matière informatique et de cybersécurité. Cependant dans un monde post-Snowden, les dynamiques ont tendance à évoluer comme le montre le cas récent de Microsoft qui a, pour l'instant, refusé de fournir à la justice américaine l'accès à des données hébergées sur le sol européen (en Irlande) malgré la demande d'un juge fédéral mais la procédure est toujours en cours. L'extra-territorialité de la législation américaine représente un coût sécuritaire pour l'UE mais aussi un coût

économique. La localisation des données est ainsi devenue un enjeu important au sein des Etats-membres qui y voient une opportunité de protéger les données de leurs citoyens et de leurs entreprises de la surveillance des Etats-Unis » (Amaël Cataruzza, 2014). Cette importance de la localisation physique (stockage des données) constitue donc un enjeu majeur de souveraineté numérique. A ce titre l'ANSSI[16], garante de la résilience des systèmes d'information et de communication de l'Etat et de ses opérateurs d'importance vitale (OIV) incite fortement à privilégier des offres nationales en matière de stockage de données, voire à rapatrier de l'étranger les données d'importance. Cette frilosité des Etats est légitime et compréhensible eu égard à la sensibilité de ces *data* et à la difficulté à sécuriser leur stockage. De même, l'Europe se sentant menacée dans sa souveraineté numérique, a également suivi cette tendance en développant un projet de *Data Protection Regulation* adopté le 12 mars 2014, visant à imposer aux entreprises d'héberger désormais les données des citoyens européens sur le sol européen.

- <u>Propriété des données, données à caractères privée (sphère public / privée)</u>

Lorsque l'on traite des *data* il convient de distinguer le contenant et le contenu. Le contenant regroupe l'ensemble des données extérieures, paramètres, environnements liés à l'objet numérique en question. Par exemple lorsque l'on traite d'un *tweet* ne dépassant pas les 140 symboles (le contenu), le contenant paradoxalement, peut s'avérer nettement plus prolixe en dévoilant des informations comme les différents rebonds (du *tweet*), l'auteur, les *followers*, les coordonnées géographiques, les liens, etc…autant d'informations qui, croisées avec des milliards d'autres *tweets*, peuvent permettre de définir de véritables arborescences numériques très utiles pour comprendre par exemple la propagation de l'information ou encore la structure d'un réseau de relations sociales.

Le contenu est lui encadré par les lois sur les données à caractère privé (DCP). Ces lois peuvent diverger d'un pays à un autre voire même s'opposer entre partenaires occidentaux. La France, et depuis peu l'Europe, ont privilégié la préservation des données personnelles au travers de lois extrêmement protectrices et d'organismes nationaux ou autorités indépendants chargés de veiller au respect de celles-ci. Pour la France, la Commission nationale de l'informatique et des libertés (CNIL) se charge de recenser puis de contrôler les organismes

[16] ANSSI : Agence Nationale de Sécurité des Systèmes d'Information.

utilisant les DCP. Ces lois européennes[17] sont davantage orientées vers la protection des droits du citoyen ce qui peut limiter les possibilités du *data mining*, pire, amener à délocaliser ces données vers des pays moins scrupuleux des droits individuels ou plus à la recherche d'une omniscience numérique (Google par exemple). Le traitement des *data* aux Etats-Unis, en revanche, peut d'une certaine manière se rapprocher de cette dernière tendance plus liberticide depuis les attentats du 11 septembre 2001. Le contrôle des flux sur le Web comme sur Internet constitue une priorité des services de renseignement américains à des fins sécuritaires, priorité sous-tendue dans le *Protect America Act of 2007* et dont la mise en œuvre au travers d'écoutes a été rendue public par le lanceur d'alerte Edward Snowden en 2013 (programme *PRISM*[18]).

En conclusion, l'aspect juridique des *data* constitue un enjeu majeur pour une optimisation du *data mining*. Envisagé dans un cadre sécuritaire, isoler une partie de ces données (comme les DCP) peut restreindre les performances de tels traitements . Dès lors, s'opposent la tolérabilité d'un effritement de la sphère privée du citoyen par rapport à son aversion au risque, terrosiste notamment. C'est pourquoi de plus en plus de pays adaptent leurs législations, en définissant des cadres particuliers, pour pouvoir bénéficier si nécessaire d'un outil supplémentaire capable d'offrir davantage de sécurité à leurs concitoyens. La loi n° 2015-912 du 24 juillet 2015 relative au renseignement permet à ce titre en France d'encadrer cette pratique au travers de services de renseignements compétents pour œuvrer dans le respect des lois de la République.

[17] Lois françaises et européennes sur la protection des DCP :
- Loi n°78-17 du 6/01/1978,
- loi n°2004-801 du 6/08/2004,
- RE n°611/2013 du 25/08/2013,
- RE sur la DCP du 14/04/2016.

[18] *PRISM , également appelé US-984XN, est un programme américain de surveillance électronique par la collecte de renseignements à partir d'Internet et d'autres fournisseurs de services électroniques. Ce programme classé, relevant de la National Security Agency (NSA), prévoit le ciblage de personnes vivant hors des États-Unis. (Source Wikipedia)*

1.3 Les conséquences générales de cette explosion de données

Après avoir vu les causes et les risques associés à cette explosion des données numériques dans les différents espaces numériques, il convient d'en présenter les conséquences majeures dans notre société en partant des domaines techniques et scientifiques propres aux *data,* pour ensuite migrer vers les mutations induites dans notre économie puis dans notre société au quotidien. Cette approche au travers de trois axes majeurs réellement impactés par l'apport des *data* parait indispensable pour comprendre conscience de l'importance des mutations générées par l'explosion des données en l'espace de deux décennies.

- Un changement de paradigme en cours pour la Science

La multiplication des données scientifiques résulte en partie du recours et de l'utilisation toujours plus grande de l'informatique comme outils non seulement de calcul mais aussi et surtout de simulation. Les performances technologiques sans cesse repoussées dans l'informatique (cf 0) permettent ainsi de multiplier les expérimentations afin d'infirmer ou confirmer les modèles présupposés. Cette pratique contribue ainsi à faire de l'analyse statistique petit à petit un nouveau procédé de validation. A ce titre on constate une accélération des publications d'articles et d'études en l'espace d'une décennie, preuve que l'essor de la technologie et des réseaux d'échanges permet aux sciences dures comme sociales de favoriser la production de savoir (malgré une augmentation en parallèle des retraits d'articles). Désormais l'explosion des *data,* notamment par l'ouverture des « données publiques », des données ouvertes (*open data*) et le recours à la simulation – approche expérimentale - permet de découvrir des comportements encore non suspectés, comme en témoignent les dernières découvertes au *LHC* (grand collisionneur à hadrons du CERN) sur l'existence d'une possible particule encore inconnue et qui serait plus lourde que le boson de Higgs. Ainsi la possibilité de multiplier la simulation presque sans limite générant d'énormes quantités de données est en train de modifier l'approche de la recherche en général sous la forme d'un nouveau paradigme comme l'explique Chris Anderson : *« The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all»* (Anderson, 2008). En effet traditionnellement pour les chercheurs, les

hypothèses étaient le préambule de la réflexion, celles qui permettaient aux génies de s'exprimer au travers d'intuitions, de modèles (Galilée, Copernic, Newton, Einstein pour ne citer que les plus connus…). Le temps, avec la simulation venaient, en second lieu, confirmer au sens cartésien du terme si ces intuitions étaient fondées ou non. Désormais avec cette multiplication des simulations, on observe d'abord des comportements répétés appelé tendances qui permettent souvent de découvrir des hypothèses voire des modèles jusqu'alors jamais pressentis. Bien que cette nouvelle approche puisse étonner au sens scientifique du terme, il convient de la prendre en compte tant son efficacité semble se conforter d'année en année : *« Yet, science advances only if it can provide explanations, failing which, it becomes an activity more akin to stamp collecting. Now, there is an area where petabytes of information can be used for their own sake. But please don't call it science. »* (Pigliucci, 2009)

- Des conséquences pour l'économie

La profusion de *data* a également eu d'importantes conséquences dans le domaine de l'économie d'abord dans le *e*-marketing puis dans l'*e*-économie avec l'émergence de géants du numériques grâce à des *business models* centrés sur cette connaissance des clients autour des *data*. Cette réussite de l'économie numérique en l'espace d'une décennie attise les convoitises, comme en atteste le *boom* des nouvelles *start-up* du numérique créées avec pour seule finalité : devenir les Google de demain.

Dans l'*e*-marketing, l'apport des *data* en nombre a permis de mieux comprendre les attentes du client pour davantage le cibler avec des produits correspondant à ses attentes. Cette compréhension du client avec des techniques comme les *cookies*, qui suivent chacun de nos clics de souris lors des navigations sur les différents sites a permis de considérablement améliorer le ratio de passage d'actes d'achat par rapport au marketing classique. L'*e*-marketing est donc au cœur des enjeux économiques du numérique dans les sociétés libérales. Dans la même lignée, en l'espace d'une décennie, l'*e*-marketing a permis l'émergence de multinationales comme les GAFA(M) qui constitue les plus grosses capitalisations boursières mondiales (plusieurs milliards de $). Ce *e*-marketing généralisé à grande échelle a permis d'obtenir une connaissance des internautes suffisante au travers de leurs navigations régulières sur la toile pour les placer au cœur de leurs *business models* : *« Dans une étude*

publiée à l'automne 2014, l'agence en innovation Fabernovel s'est penchée sur les nouvelles règles que ces multinationales du numérique ont érigées. Elle a distingué un principe de base : le client est placé au centre du jeu là où, dans l'« ancienne économie », c'était le produit qui jouait le rôle principal. L'important pour l'entreprise est donc avant tout de répondre aux besoins des consommateurs et, ce faisant, de capter la base de clients la plus large possible. C'est en fidélisant ces derniers avec des produits ou des services dont le coût d'usage devient une question secondaire – la gratuité de certains d'entre eux est assumée – que l'entreprise pourra envisager de créer de la valeur. Il s'agit là d'un second point essentiel pour comprendre le modèle porté par les Gafa : la présence de ces nombreux consommateurs gravitant autour de l'écosystème de l'entreprise permet de rassembler des données en grande quantité. Des données dont la possession va constituer une richesse stratégique apportant à leurs propriétaires une connaissance essentielle des marchés mais aussi un savoir monnayable.»[19]
(Roché, 2016)

- Des conséquences sociétales

L'explosion des données a également profondément transformé notre société non seulement dans la gestion de l'information mais aussi dans les relations humaines et autres comportements sociétaux.

Cette profusion d'informations diffusées de plus en plus sur des supports numériques est en train de progressivement changer notre façon d'accéder à l'information. Ainsi l'essor du numérique en une génération remet en cause une culture qui s'est transmis durant plusieurs millénaires d'abord par l'oral puis par l'écrit. Cette transformation aussi rapide des vecteurs de l'information et de diffusion apporte un certain nombre d'avantages mais pose aussi des problèmes.
Tout d'abord, cette multiplication des accès à l'information s'accompagne d'une multiplication et d'une diversification des sources d'information, jusqu'à peu encore « propriété » exclusive des médias. Désormais chacun peut au travers des contenus qu'il diffuse informer les autres et devenir un informateur aux yeux du grand public. Cette

[19] Source : http://www.lemonde.fr/emploi/article/2015/03/25/le-modele-economique-reinvente_4601288_1698637.html

multiplicité des informateurs permet un accès à l'information souvent plus rapide et surtout de plus en plus difficile à cacher, comme en témoigne l'essor des lanceurs d'alertes ces dernières années. Dans un autre registre, la campagne de communication du président Donald Trump directement adressée au public démontre à quel point les médias classiques ont perdu de leur importance et peuvent être contournés avec une certaine efficacité.

Mais cet accès à l'information s'accompagne en parallèle d'une multiplication de la désinformation, qui ne peut être combattue qu'avec un sens critique que le média doit par définition produire sous forme de réflexion et de mise en perspective, au profit du commun des mortels, qui lui, ne l'effectue pas forcement soit par négligence, simplicité ou tout simplement par excès de confiance. Dès lors la recherche d'une sorte de vérité « absolue » par un travail en amont de vérification des informations s'efface de plus en plus au profit d'une vérité relative pondérée par la vitesse de communication et la puissance de l'informateur souvent résumée en nombre de *followers*, qui serait en quelque sorte un gage de confiance. Or cette approche et ce tempo médiatique, au mieux limitent et nivellent l'information, au pire faussent l'information en générale.

Aussi ce comportement entraîne un certain communautarisme de l'information, où chacun recherche dans cette pléiade d'informateurs l'information qui lui plaît, qui va en quelque sorte dans son sens au détriment d'une information construite. Ce comportement oblige à fidéliser le client plus versatile par évolution (qui peut facilement changer de média) en privilégiant des informations plus sensationnelles susceptibles de faciliter parfois l'essor du populisme. Ces changements fragilisent les médias classiques qui perdent progressivement leur âme et *in fine* nos démocraties.

Par ailleurs la quantité d'informations produites ne cesse de croitre d'année en année, générant une surinformation résumée par le terme d' « infobésité » qui nécessite de plus en plus de temps pour trier le bon grain de l'ivraie. Cette recherche de l'information dans ce déluge informationnel nécessite toujours plus de temps et de moyens. A ce titre le *data mining* constitue réellement une solution, ou pour le moins une perspective de recoupement.
De plus le développement des réseaux sociaux sans contrôle abolit progressivement la frontière entre vie privée et publique de chaque citoyen. Cet effacement des frontières bénéficie à de plus en plus d'*e*-entreprises qui voient dans cette omniscience, une valorisation

de leur richesse. Ce nouveau canal d'information est aussi en train de devenir une véritable source d'information voire de renseignement aussi appelé intelligence économique dans le monde de l'entreprise.

Enfin cet accès à l'information plus ou moins choisie rapproche virtuellement les internautes mais paradoxalement isole physiquement les gens et individualise davantage notre société, la rendant plus fragile notamment en terme de solidarité : Plus d'informations pour moins de démocratie. Quel paradoxe !

On assiste donc à un triple bouleversement en marche dans nos sociétés d'abord dans la manière de diffuser cette profusion d'informations, ensuite dans sa pertinence où quantité et qualité semblent parfois s'opposer, enfin dans la relation qu'elle instaure entre informateur et informé, relation qui tend à se virtualiser voire s'individualiser. Un des risques liés à cette évolution de l'information réside d'abord dans la virtualisation de la communication qui modifie les rapports humains et ensuite dans l'essor du populisme qui peut parfois fragiliser cette solidarité nécessaire à la cohésion des démocraties.

En conclusion, l'explosion des *data* est réellement en train de modifier notre rapport dans de nombreux domaines structurant nos sociétés. Ce chapitre tend à montrer que notre société se modifie plus rapidement que durant n'importe quelle période de l'histoire. Ces changements induisent des enjeux importants et nécessitent donc de traiter cette énorme quantité de données avec efficience, d'où le recours nécessaire au *data mining*. Utile aussi bien pour synthétiser, que pour trouver des comportements suspects, ou encore renseigner, le *data mining* nécessite d'être compris dans sa globalité pour en tirer son plein potentiel. C'est donc tout l'intérêt du deuxième chapitre de comprendre « comment gérer cette masse de données pour ensuite en tirer de l'information ? ».

2. …qui oblige à mettre en place des techniques adéquates pour en tirer profit.

Pour comprendre le *data mining* il convient de l'étudier à plusieurs niveaux. Tout d'abord, ce chapitre s'attachera à en décrire les grands principes généraux (cf. 2.1) au travers de ses caractéristiques, de son fonctionnement des méthodes structurantes et du rôle de l'humain dans cette science combinatoire. Les deux autres sous-chapitres décriront de manière détaillée les deux méthodes existantes aux finalités différentiées : celles descriptives (cf. 2.2) qui visent à extraire l'information pertinente recherchée par l'exploitation des données, et celles prédictives (cf. 2.3) qui visent davantage à prédire un comportement, détecter des signaux faibles, une information au sens d'une intelligence artificielle. Ce chapitre aborde bien le *data mining* d'abord sous sa forme littérale et théorique nécessaire pour permettre de comprendre ensuite les avancées possibles en général et plus particulièrement pour la Défense, objet du chapitre suivant.

2.1 Le *data mining* et son environnement

Cette sous-partie vise à comprendre les principes et fonctionnements généraux du *data mining* dans son aspect global comme technique, en le distinguant du *big-data*. Cette distinction entre deux termes voisins souvent confondus par facilité et abus permettra de mieux cerner le périmètre de chacun. Enfin comprendre le *data mining* passe par la connaissance de son écosystème à commencer par ses acteurs et en particulier le rôle singulier du *data scientist* dans cette révolution numérique.

- Principes et fonctionnement du *data mining* (processus itératif en 5 temps)

Répondre au « comment » passe indubitablement par un retour préalable sur les différents sens du *data mining*, donnés par des chercheurs ou autres spécialistes des nouvelles technologies. Ainsi une définition possible du *data mining* pourrait être *«l'analyse des données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns»*[20]. Pour Anand et Buchner, il *« offre des algorithmes et des outils pour la découverte de modèles non*

[20] Source : http://www.lebigdata.fr/data-mining-definition-exemples

triviaux, implicites, non connus, potentiellement utiles et compréhensibles à partir d'une grande masse de données ». Pour d'autres, il s'associe au *«Knowledge Discovery in Data »*, qui pourrait se traduire par découverte de savoir dans les données et qui résume bien la finalité première du *data-mining*.

Il convient ensuite de s'interroger véritablement sur les grands principes caractéristiques du *data-mining* pour en comprendre son fonctionnement avant d'en discerner les subtilités : cette science consiste à faire parler des données quel que soit leur format (formats qui seront détaillés au chapitre suivant) en essayant d'établir ou de trouver des corrélations (*patterns*) entre des données initialement sans valeur pour pouvoir *in fine* en tirer des informations. Le *data mining* selon Anand et Buchner *« offre des algorithmes et des outils pour la découverte de modèles non triviaux, implicites, non connus, potentiellement utiles et compréhensibles à partir d'une grande masse de données. Le data-mining n'est pas un nouvel outil magique qui résoudrait tous les problèmes d'extraction de l'information qui seraient soudainement apparus. Le data mining s'appuie à la fois sur les techniques statistiques, les réseaux de neurones, les techniques de visualisation... ainsi que sur des techniques spécialement développées pour parcourir d'immenses bases de données à la recherche de patrons fréquents. »* (S. S. Anand, 1998). Cette valorisation des *data* ne peut s'effectuer au préalable sans une capacité de stockage centralisée des données appelée *data warehouse*. Quant-au *data warehousing*, il s'agit du procédé consistant à centraliser la gestion mais aussi la recherche de données.

En matière de processus, le *data-mining* se décompose en cinq étapes clés qui correspondent à une description séquentielle débutant par la collecte de données et leur chargement au sein des *data warehouses* précédemment évoquées. Ces données sont ensuite stockées et retranscrites dans un système de bases de données multidimensionnel, avant d'être exploitées par les *data scientists*. Lors de cette étape d'harmonisation, il s'agit de faire converger deux types de données : des données brutes et des données qualifiées de « symboliques », provenant de la description des premières : contexte d'acquisition, géo-référencement, *et cetera*. Cette étape cruciale et complexe consiste à mettre au point une représentation sémantique de ces données symboliques dans le but de leur attribuer une signification scientifique, à l'instar du projet

CrEDIBLE[21] développé dans le domaine de l'imagerie médicale. Ensuite un logiciel applicatif réalise l'opération de recherche des corrélations suivant les entrées et sorties souhaitées par les utilisateurs. Pour finir le résultat est présenté dans un format exploitable de type graphique et tableau.

Dans ce phasage, l'étape de recherche de corrélation est bien primordiale et constitue le cœur du *data mining*. Elle peut se réaliser par cinq méthodes plus ou moins complémentaires, définies ci-dessous :

- ***Règle d'association*** *: chercher des patterns au sein desquelles un événement est lié à un autre événement.*
- ***Méthode d'analyse de séquence*** *– chercher des patterns au sein desquelles un événement mène à un autre événement plus tardif.*
- ***Méthode de classification*** *– chercher de nouveaux patterns, quitte à changer la façon dont les données sont organisées.*
- ***Méthode de clustering*** *– trouver et documenter visuellement des groupes de faits précédemment inconnus.*
- ***Méthode de prédiction*** *– découvrir des patterns de données pouvant mener à des prédictions raisonnables sur le futur. Ce type de data mining est aussi connu sous le nom d'analyse prédictive».* Cette technique s'effectue grâce à l'apprentissage supervisé.

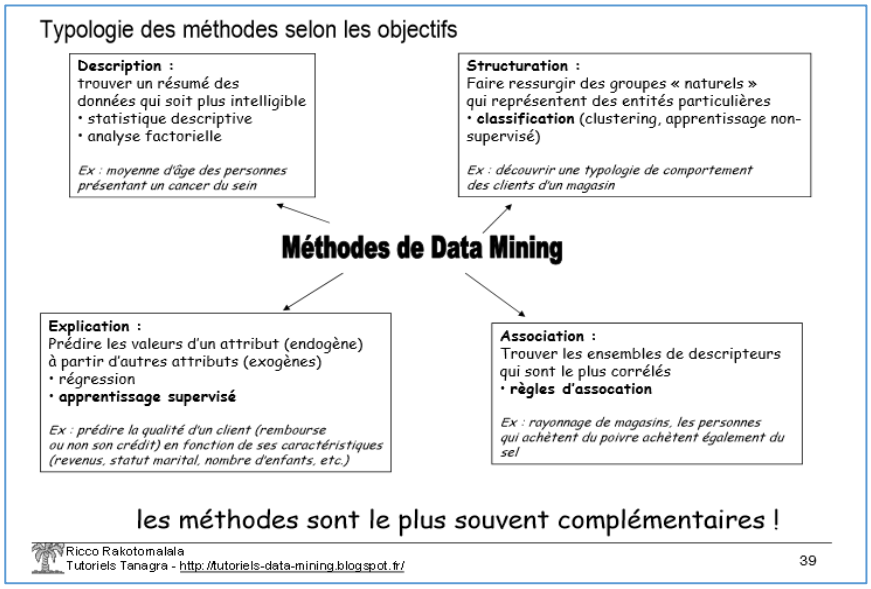


Figure 7: Typologie des méthodes selon les objectifs (Source : Ricco Rakotomaliala)

[21] Projet CrEDIBLE : fédération de données et de ConnaissancEs Distribuées en Imagerie BiomédicaLE. Ce projet de la mission pour l'interdisciplinarité du CNRS vise à étudier les verrous qui empêchent aujourd'hui de construire les systèmes de partage de données réparties hétérogènes dans le domaine de l'imagerie médicale. (Source : https://credible.i3s.unice.fr/)

Un des premiers défis réside dans la capacité à adapter le *data mining,* à le paramétrer pour lui permettre cette permanence de la manipulation des données. Cette phase d'adaptation passe par la recherche de nouveaux algorithmes capables de digérer cette profusion de données portée par l'explosion du web social notamment : *« Pour contourner les obstacles rencontrés dans la gestion des grandes masses de données, il faudra certes améliorer les technologies de stockage et de calcul, mais aussi inventer de nouvelles manières de manipuler les données* », annonce Farouk Toumani. Après avoir développé ces techniques véritablement au cœur du *data mining,* il paraît judicieux de comprendre ce qui le distingue du *Big data*.

- *Data mining* et *Big data* : Une complémentarité pour maîtriser cette explosion de données

Ces deux termes sont souvent utilisés et assimilés par souci de simplification sans réellement connaître leurs différences. Cette sous-partie va s'attacher à mieux comprendre ces deux termes qui sont complémentaires à bien des égards.

Les Big Data, littéralement les grosses données, est une expression anglophone utilisée pour désigner des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information. Sa finalité consiste donc à manipuler, broyer ces données pour les rendre accessible (indexation) et exploitable de manière aisée. Pour relever ces défis technologiques et méthodologiques, les Big data s'appuie sur de nombreuses techniques comme le *data warehousing* déjà développé précédemment, mais aussi en s'appuyant sur des architectures et modèles de calcul distribués indispensables pour obtenir une puissance de calcul suffisante pour digérer ses volumes de données en quasi temps réel. Ces techniques ne peuvent s'effectuer sans le recours à des logiciels incontournables dans ce milieu tel Map-Reduce ou Hadoop.

Le *data mining* constitue donc une capacité venant se greffer en complément du *Big data* pour extraire de la connaissance dans cette forte volumétrie de données. Quant aux données, elles sont digérées et traitées par le *Big data*. Pour le reste de ce mémoire, le terme de *Big data* fera référence à un volume de données ne permettant plus les usages traditionnels de base de

données mais nécessitant d'être manipulé avec des techniques et des moyens spécifiques et novateurs.

➢ La place du *data scientist*

Par ailleurs, on ne peut pas parler de *data mining* sans aborder la question des ressources humaines (RH) nécessaires, question qui reste centrale et particulièrement structurante. Dans cette sous-partie nous aborderons de manière synthétique le rôle, les missions et la place du *data scientist* dans la digitalisation future des entreprises.

Par définition, le *data scientist* est un « scientifique des données » qui doit comprendre les statistiques et le *data mining*. Pour cela il se consacre à la gestion, l'analyse des données pour les transformer en information, indicateurs et autres renseignements exploitables par les services et directions d'une entreprise. Ce métier relève donc d'enjeux à la fois fonctionnels et stratégiques pour l'entreprise, surtout dans un environnement concurrentiel où les décideurs doivent de plus en plus appuyer leur prise de décision sur des statistiques fiables. Il participe donc directement à la stratégie de l'entreprise ainsi qu'au processus de prise de décision. Par ailleurs son travail contribue à améliorer l'activité globale par la précision de l'analyse et la mise sur pied de modèle de prédiction. On retrouve les *data scientist* dans des secteurs variés comme la finance, l'informatique, l'assurance, l'e-commerce, la grande distribution. Ces missions, ici succinctement résumées, sont davantage détaillées dans la plupart des fiches métier consultables sur les sites spécialisés en ligne traitant des RH. En complément, un exemple de fiches de poste a été annexé en fin de mémoire (cf Annexe 1).

Concernant la place à accorder aux *data scientists* dans l'entreprise on observe deux tendances émergentes opposées :

- La première tendance, défendue par Michael Stonebraker[22], va vers une place croissante du *data scientist* au sein des structures stratégiques de l'entreprise et autres administrations, en lieu et place des traditionnels analystes d'entreprise. Leurs capacités à anticiper à prédire les tendances futures constituent une opportunité que les COMEX ne

[22] Michael Stonebraker, chercheur du MIT, spécialisé dans les bases de données, a reçu en 2015 le prix Alan Turing (financé par Google) qui récompense l'excellence en informatique, en reconnaissance de ses "contributions fondamentales aux concepts et pratiques qui sous-tendent les systèmes de bases de données modernes" (Source : http://www.zdnet.fr/actualites/le-dilemme-de-l-elephant-a-quoi-ressemble-l-avenir-des-bases-de-donnees-39823060.htm)

peuvent laisser échapper : *« Imaginons maintenant qu'au lieu d'embaucher une personne responsable de l'informatique décisionnelle, vous embauchiez un scientifique des données. Il créera un modèle prédictif des articles qui vont se vendre. Alors posez-vous la question : préférez-vous avoir un modèle prédictif ou un grand tableau qui vous indique ce que vous avez vendu ? Ce qui va se passer au cours de la prochaine décennie, selon moi, c'est que les scientifiques des données vont remplacer les analystes d'entreprise.»* (Barker, 2015)

- La seconde défendue par Pierre-Yves Baudot, cantonne le *data scientist* dans son rôle purement technique, en appui du management : *« il convient de ne pas accorder aux architectes de données davantage de pouvoirs qu'ils n'en revendiquent. Les analystes de données prétendent gouverner mais ne gouvernent pas »* (Baudot, 2015). Cette vision met en exergue l'importance pour le *data scientist* de travailler de pair avec un expert du domaine ou du métier pour orienter les recherches et définir les attendus.

Même si ces deux visions peuvent être contradictoires en apparence, elles restent en réalité recevables moyennant quelques adaptations comme la capacité à intégrer les spécificités et besoins métier. Au vu des arguments avancés par leurs défenseurs, le *data scientist* disposera d'un rôle central indéniable et restera donc au cœur des enjeux et préoccupations de demain. En effet, en son absence les données resteront muettes et desserviront les intérêts des entreprises. Ce poids croissant des *data scientists* dans les entreprises s'accompagne d'une offre existante encore trop limitée, qui génère de fait de fortes difficultés et contraintes de recrutement voire de fidélisation pour les directions des RH : *« À l'heure actuelle, il n'y a pas suffisamment de scientifiques des données, donc l'offre va être limitée par manque de personnes compétentes. Ce problème finira toutefois par être résolu et nous accéderons alors à un traitement analytique plus sophistiqué »*. (Barker, 2015).

Cette offre encore restreinte dans un marché très concurrentiel oblige les employeurs à une gestion réfléchie, individualisée s'il souhaite conserver et fidéliser ce genre de profils très recherchés. A titre d'indication, l'offre et la demande étant à sens unique les salaires d'embauches de jeunes diplômés s'envolent, avoisinant les 45k€ en début de carrière pour

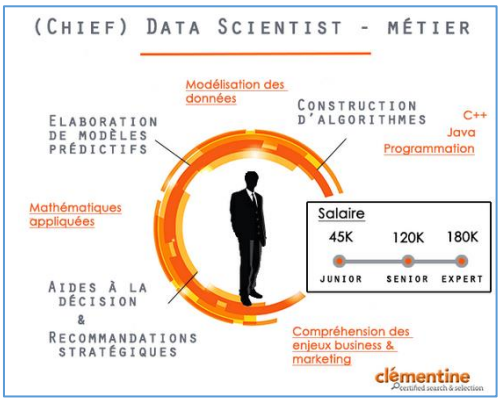


Figure 8 : Missions et salaires du data scientist (Source clémentine)

atteindre des sommets pour des *data scientists* seniors capables d'élaborer des algorithmes efficaces et pointus (cf Figure 8).

2.2 …pour répondre aux enjeux actuels : recherche de patterns (méthode descriptive)

Schématiquement le *data mining* s'appuie donc sur des *data scientists* mais aussi et surtout sur les *data* pour pouvoir en extraire des informations. Ce processus d'homogénéisation, de standardisation et enrichissement des données est complexe et mérite d'être détaillée pour comprendre ensuite le fonctionnement de la méthode descriptive et son cadre d'utilisation.

➢ Un processus d'homogénéisation et d'enrichissement des data nécessaire

Ce volume exponentiel de données qu'on regroupe sous le terme de *big data* se résumait souvent au travers des 3 V pour volume, variété et vélocité (Dumbill 2012). Désormais on leur adjoint un autre V (cf Figure 9) pour véracité. Comprendre cette règle des 4V permet de mieux paramétrer et utiliser les données :

- le volume de *data* en augmentation annuelle de plus de 50% oblige à une utilisation autre que les traditionnels outils classiques d'exploitation et de stockage. Le croisement de ces données entre elles étant à la base de la qualité de l'information générée, cette volumétrie des data devient explosive,
- la variété des données permet d'élargir le champ d'exploitation mais nécessite des outils d'harmonisation, des procédures de traitement et des ressources adaptés,
- la vélocité ou vitesse traduit le besoin en capacité de calcul pour traiter chaque jour davantage de données stockées mais aussi de données en direct sous forme de flux pour gagner du temps dans l'exploitation des informations et retarder l'obsolescence des informations dégagées,
- la véracité traduit la qualité des données et leur intégrité avec à la clé une réduction des bruits recherchée, enfin la visibilité comme gage d'efficacité dans la circulation des informations.

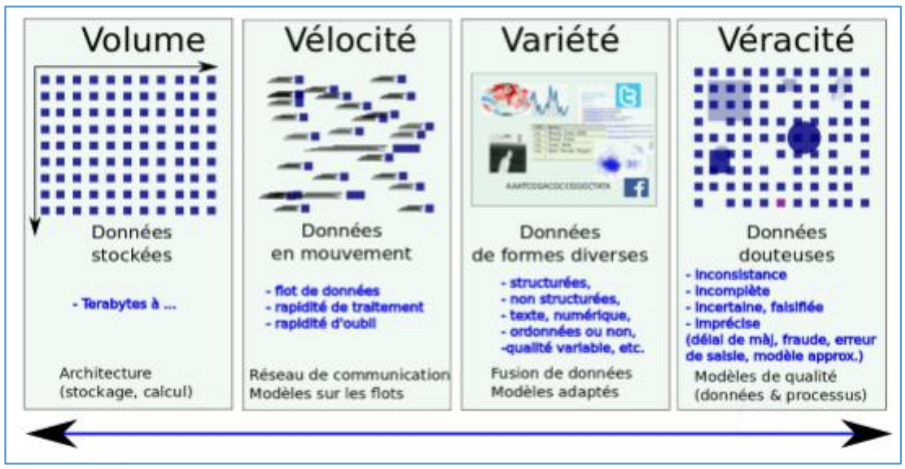


Figure 9: les 4V (Source : Picot-Clementé - Données massives, challenges et perspectives - 9/72015)

La multiplication des données s'est accompagnée en parallèle d'une prolifération des formats et des qualifications existants (résumée par le V de variété) : données brutes, désagrégées, hétérogènes (son, image, tableurs, texte…), métadonnées (cf. Figure 10: Exemple de métadonnées, *open data*… autant de types de *data* qui nécessitent d'être nettoyés, harmonisés au travers d'un processus particulier déjà décrit précédemment. Pour autant, il doit encore s'accompagner au préalable d'un travail spécifique de fiabilisation et sécurisation des données pour les rendre légitimes et espérer obtenir par le *data mining* des informations ou du renseignement faisant foi. C'est ce besoin de véracité qui est nécessaire en particulier pour les *open data.*

Tableau 1. Génération des métadonnées : un aperçu[76]

Service utilisé	Métadonnées générées	
Courrier électronique	• Nom, adresse de courrier électronique et adresse IP de l'expéditeur • Nom et adresse de courrier électronique du destinataire • Information de transfert des serveurs • Date, heure et fuseau horaire • Identifiant unique du message électronique et des messages reliés • Type de contenu et encodage	• Données de connexion du client de messagerie grâce à l'adresse IP • Formatage des entêtes du client de messagerie • Priorités et catégories • Sujet du message électronique • État du message électronique • Demande de confirmation de lecture
Téléphonie cellulaire	• Numéro de téléphone de chaque participant à l'appel • Numéros de série uniques des téléphones utilisés • Heure de l'appel	• Durée de l'appel • Emplacement géographique de chaque participant à l'appel • Numéros de cartes téléphoniques
Twitter	• Nom, emplacement géographique, langue, information biographique contenue dans le profil et adresse URL • Date de création du compte • Nom d'utilisateur et identifiant unique • Emplacement, date, heure et fuseau horaire du « Tweet » • Identifiant unique du « Tweet » et de celui auquel l'utilisateur répondait	• Identifiants des contributeurs • Décompte du nombre de comptes « suivant » l'utilisateur, du nombre de comptes qu'il « suit » et de son nombre de « Tweets » favoris • Statut de vérification • Nom de l'application envoyant le « Tweet »

Figure 10: Exemple de métadonnées

Une fois ces données harmonisées et donc exploitables il convient de les enrichir à plusieurs niveaux pour en tirer du renseignement encore appelé intelligence non seulement dans le monde économique mais aussi dans le monde anglo-saxon. Le schéma ci-dessous décrit ce processus d'enrichissement étape par étape, avec des données qui deviennent d'abord information en les contextualisant. Puis connaissance en ajoutant du sens. Présenter ensuite

ces connaissances en les rapprochant de nos intérêts permet d'en tirer du savoir qui deviendra de l'intelligence à condition de profiter aux différents services utilisateurs et demandeurs. Ce processus ne peut pas exister sans l'action humaine des *data scientist* et de leurs équipes. La sous-partie 3.2 traitant du renseignement cyber développera davantage les techniques d'enrichissement des données.

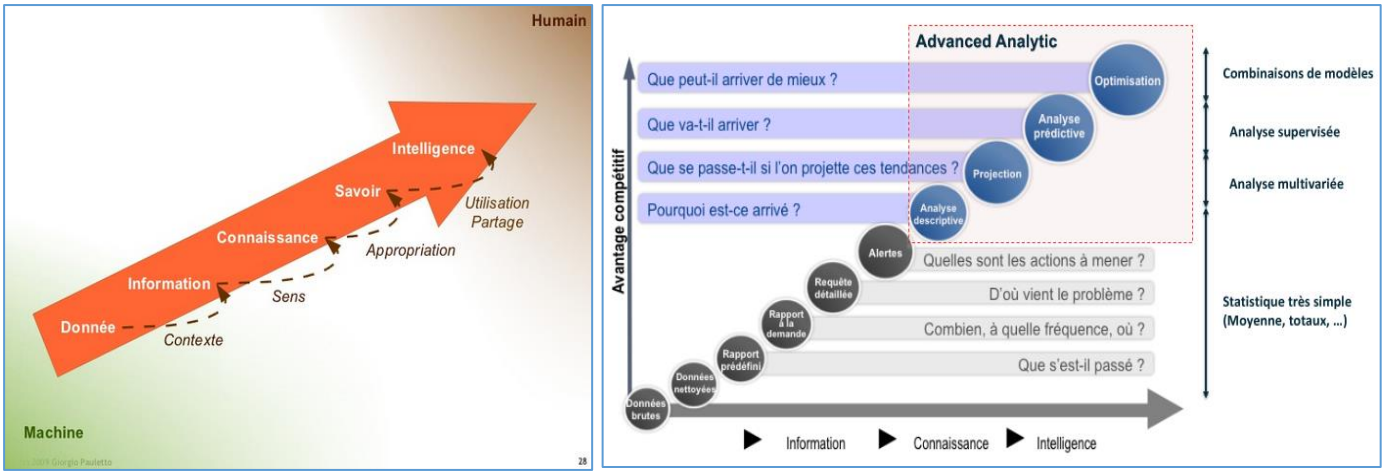


Figure 11 : Processus d'enrichissement de la donnée (Sources Giorgio Pauletto / Didier Gaultier)

Les deux principales méthodes permettant d'enrichir ces données homogénéisées sont soit descriptives ou soit prédictives. Les deux sous-parties suivantes présenteront ces approches spécifiques qui permettent de donner du sens à l'information.

- La méthode descriptive :

La méthode descriptive vise à mettre en évidence des informations présentes mais cachées par le volume de données. Cette méthode ne cherche à prédire de variable « cible » mais à décrire de façon simple une réalité complexe en la résumant. Elle permet de détecter des associations entre des objets comme des liens de causalités. Ce sous-chapitre assez technique s'appuiera beaucoup sur les cours et travaux réalisés par Stéphane Tuffery, éminent spécialiste dans ce domaine.

Une des techniques descriptives de *data mining* reste la classification qui vise à regrouper des objets en groupes, de manière à ce que ceux dans le même groupe se ressemblent le plus possible et ceux dans deux groupes différents se distinguent suffisamment. Cette technique est

particulièrement utilisée dans des domaines comme le marketing, le médical ou la sociologie afin de définir des groupes homogènes adaptés à une offre, un protocole thérapeutique, ayant une opinion particulière ou des attentes spécifiques. Cette technique permet alors de proposer des ciblages particuliers. La classification permet également de filtrer les variables dites discriminantes en homogénéisant des groupes d'individus. Par ce triage, cela permet également d'accroître la fiabilité des prédictions.

Une autre technique existe avec la recherche d'associations qui demeure très utilisée dans la grande distribution (analyse du ticket de caisse ou panier de la ménagère). Par exemple, elle vise à ajuster les options que l'on peut rajouter (package) dans les produits classiques pour accroître leur rendement. Cette technique nécessite des volumes importants de données pour trouver quelques règles intéressantes noyées parmi les règles triviales voire non utilisables.

La méthode descriptive s'appuie sur un certain nombre d'algorithmes suivant le modèle choisi pour l'étude des données. Le tableau ci-dessous détaille à titre d'information et de documentation l'ensemble des algorithmes utilisables dans le cadre de l'usage d'un modèle descriptif.

type	famille	sous-famille	algorithme
méthodes descriptives	modèles géométriques	analyse factorielle (projection sur un espace de dimension inférieure)	analyse en composantes principales ACP (variables continues)
			analyse factorielle des correspondances AFC (2 variables qualitatives)
			analyse des correspondances multiples ACM (+ de 2 var. qualitatives)
		analyse typologique (regroupement en classes homogènes)	méthodes de partitionnement (centres mobiles, *k*-means, nuées dynamiques)
			méthodes hiérarchiques
		analyse typologique + réduction dimens.	classification neuronale (cartes de Kohonen)
	modèles combinatoires		classification relationnelle (variables qualitatives)
	modèles à base de règles logiques	détection de liens	détection d'associations

En grisé : méthodes « classiques »

Figure 12 : Tableau des méthodes descriptives (Source : COURS DE DATA MINING - Stéphane TUFFERY – Octobre 2011

2.3 …et préparer les enjeux de demain : la prédiction (méthode prédictive)

➢ La méthode prédictive :

Cette méthode, appelée aussi apprentissage supervisé (réseaux de neurones) vise à expliquer une variable soit de manière qualitative (discrimination), soit de manière quantitative (régression). Enfin le *scoring* cherche à répondre à une problématique de manière binaire (OUI/NON ou risqué / pas risqué). Ces techniques peuvent prédire l'achat de produits, les impayés, la fuite d'un client (phénomène de *churn*) mais aussi la reconnaissance de tumeurs, des facteurs de décès pour une pathologie…

La méthode prédictive s'appuie sur un certain nombre d'algorithmes suivant le modèle défini pour l'étude des données. Le tableau ci-dessous détaille à titre d'information et de documentation l'ensemble des algorithmes utilisables dans le cadre de l'usage d'un modèle prédictif.

type	famille	sous-famille	algorithme
méthodes prédictives	modèles à base de règles logiques	arbres de décision	arbres de décision (variable à expliquer continue ou qualitative)
	modèles à base de fonctions mathématiques	réseaux de neurones	réseaux à apprentissage supervisé : perceptron multicouches, réseau à fonction radiale de base
		modèles paramétriques ou semi-paramétriques	régression linéaire, ANOVA, MANOVA, ANCOVA, MANCOVA, modèle linéaire général GLM, régression PLS (variable à expliquer continue)
			analyse discriminante linéaire, régression logistique, régression logistique PLS (variable à expliquer qualitative)
			modèle log-linéaire, régression de Poisson (variable à expliquer discrète = comptage)
			modèle linéaire généralisé, modèle additif généralisé (variable à expliquer continue, discrète ou qualitative)
	prédiction sans modèle		*k*-plus proches voisins (*k*-NN)

En grisé : méthodes « classiques »

Figure 13: Tableau des méthodes prédictives (Source : COURS DE DATA MINING - Stéphane TUFFERY – Octobre 2011

➢ La prédiction

Cette méthode prédictive se caractérise de manière schématique par la découverte de *patterns* de données conduisant à des prédictions raisonnables sur le futur. L'image qu'il convient d'associer est celle de la météorologie, qui au travers de données locales, d'algorithmes et de capacités de calcul colossaux, permettent d'anticiper la météo avec une plus ou moins grande fiabilité suivant la durée souhaitée. Elle présente donc un intérêt majeur pour toute entité souhaitant appréhender au mieux l'avenir et réduire au minimum l'incertitude. Or qui n'a jamais souhaité anticiper l'avenir pour gagner des marchés, pour conserver l'initiative dans des domaines très concurrentiels ou encore progresser en efficience ?

Figure 14 : Exemple de carte fournie par Predpol avec les zones ciblées de patrouilles (Source Predpol)

A ce jour, l'exemple le plus connu et médiatisé après la météo reste celui des patrouilles de police aux Etats-Unis qui utilise un logiciel appelé *« Predpol »* basé sur le *« predicive policing »* ou encore *« geographic profiling »*. Ce logiciel mêlant algorithme (secret) et statistique repose sur des bases de données étayées et suffisantes en termes d'antériorité pour permettre une vision assez large. Son fonctionnement ? : *« Predpol, par le biais d'une interface cartographique, affiche quotidiennement un ensemble de zones à risques (par exemple : 20 cases de 150m x 150m) pour lesquelles la probabilité qu'un délit se manifeste est importante »* (Ismael, 2014). Ce logiciel a par ailleurs été présenté comme une des 50 innovations de l'année 2011 par le *Time Magazine.* La raison évoquée ? : La baisse de la délinquance ; Avec pour en attester des premiers résultats positifs dans des villes comme Atlanta ou Los Angeles où une baisse de plus de 10% de la délinquance a été observée dans les quartiers où le logiciel était en expérimentation.

Mais est-ce dû réellement aux prédictions ou le simple fruit du hasard ?

Les études menées par Ismael Benslimane montrent paradoxalement que l'efficacité n'est pas si évidente qu'elle n'y paraît au premier abord. La comparaison avec les *opendata* sur la criminalité de ville de Chicago analysées via ses algorithmes propres lui ont permis d'obtenir des résultats proches de celui de *Predpol.* Ces résultats ne feraient en fait qu'illustrer la loi de Pareto déjà connue depuis le XIX[e] siècle : *« Ismael Benslimane a dû construire son propre modèle. […]. Il a développé un algorithme de prédiction aléatoire (où chacune des zones du territoire a chaque jour une chance sur l'ensemble d'être tirée au sort), un autre pondéré par le taux de criminalité (favorisant les zones avec une plus forte activité criminelle), et un dernier qui favorise les zones à plus haut risque sur les autres. Ce dernier algorithme obtient des scores de prédiction très proche de la courbe de Predpol et même meilleure si on élargit la carte du territoire. Le problème, souligne Ismaël Benslimane, c'est que lorsque l'on quadrille la ville, la plupart des délits ont toujours lieu dans les mêmes secteurs, suivant la classique loi de Pareto qui date du XIXe siècle : à savoir que 80% des délits à arme à feu ont lieu dans 20% du quadrillage»*[23]. Au final ce logiciel d'apprentissage automatique semble apporter non pas une plus-value prédictive mais davantage organisationnelle, en permettant de mieux patrouiller dans des zones sensibles déjà répertoriées par les polices. Il rappelle surtout tous les biais possibles dans la saisie des données et la mesure de la criminalité (notamment les effets paillasson, Barnum et cigogne qui rendent compte de la subjectivité des mesures), c'est surement ce qui le différencie à l'heure actuelle de la météorologie où les données sont moins contestables car standardisées. Néanmoins cet exemple montre bien l'utilité de données fiabilisées non sujettes à interprétation produites par ces patrouilles de police (qui ne peuvent se multiplier à l'excès avec des budgets de fonctionnement toujours contraints) qui gagnent ainsi en efficience en ciblant les quartiers les plus sensibles durant les différentes périodes de la journée. Cette optimisation de l'organisation des patrouilles de police a finalement contribué à diminuer les agressions, vols et autres comportement délictueux et donc

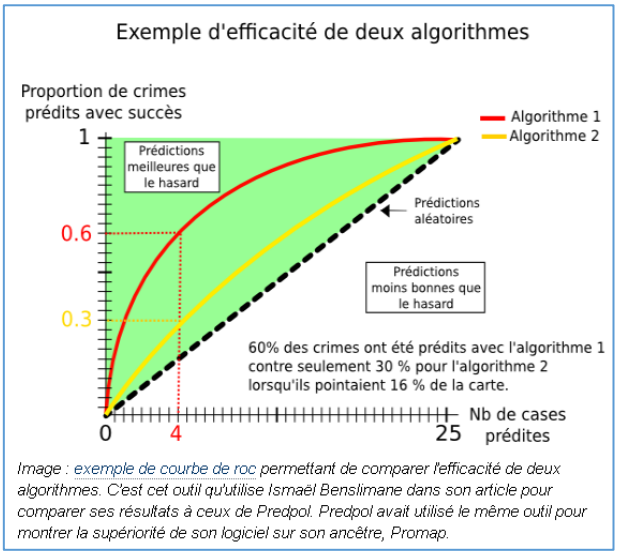


Image : exemple de courbe de roc permettant de comparer l'efficacité de deux algorithmes. C'est cet outil qu'utilise Ismaël Benslimane dans son article pour comparer ses résultats à ceux de Predpol. Predpol avait utilisé le même outil pour montrer la supériorité de son logiciel sur son ancêtre, Promap.

Figure 15: Exemple d'efficacité de 2 algorithmes

[23] Source : http://internetactu.blog.lemonde.fr/2015/06/27/police-predictive-la-prediction-des-banalites/

l'insécurité en général même si l'aspect prédictif avancé n'est pas si avéré en l'état actuel des recherches et nécessite par ailleurs d'autres réponses davantage sociétales : « *Au-delà des aspects les plus techniques, nous pouvons nous demander, si la démarche de Predpol ne désyncrétise pas la question de la criminalité en laissant penser qu'il suffit de prédire les délits pour en diminuer le nombre. Predpol et sa médiatisation véhiculent ainsi une idée répandue, "simple" et séduisante oubliant de facto les facteurs sociologiques amenant aux comportements délictueux. La question fondamentale des inégalités de répartition des richesses est, par exemple, rarement débattue. En effet, cette réflexion est beaucoup plus impliquante à long terme qu'un logiciel spectaculaire* ». (Ismael, 2014)

3. Une opportunité et une nécessité pour la Défense, mais avec des défis à relever

Après avoir détaillé le fonctionnement général du *data mining* et l'usage possible de la *data* au chapitre précédent, il convient de proposer dans ce dernier chapitre des pistes d'utilisation voire de réflexion pour la Défense, principal commanditaire de cette étude. Pour cela, il conviendra de les intégrer dans les enjeux futurs des armées souvent décrits au travers de documents de prospective, comme par exemple Action terrestre future destiné à préparer l'avenir de l'Armée de Terre. Cette anticipation se nourrira également des solutions déjà mises en place dans le secteur privé, secteur souvent plus réactif, innovant et entreprenant sur le sujet du numérique en raison de contraintes de sécurité et surtout de résilience moins poussées. L'apport du *data mining* pour la Défense, doit donc se concevoir à court terme comme une plus-value technique duale indispensable dans certains secteurs militaires déjà fortement influencés par le milieu civil (cf. 3.1), mais aussi à plus longue échéance comme le noyau, le socle de certaines capacités en plein essor, devenues incontournables car au cœur de la transformation digitale des Armées : le renseignement (cf. 3.2), la cybersécurité (cf. 3.2) et l'influence (cf. 3.4).

3.1 Le *data mining* transposable à court terme dans des secteurs universels voire duals.

Ayant fait ses preuves depuis une dizaine d'années dans certains domaines, le *data mining* permet d'envisager sa transposition à brève échéance dans un certain nombre de secteurs dépourvus de spécificité militaire particulière. Ainsi certaines pistes duales (cf. Figure 16) comme la logistique, les ressources humaines mais aussi des domaines plus proches du commandement (ou management dans le monde civil) comme les outils d'appui au commandement, de performances-synthèses peuvent légitimement être dotés de cette capacité pour gagner en efficience. C'est l'objet de cette partie de détailler ces possibilités et ces secteurs où le *data mining* va naturellement s'imposer.

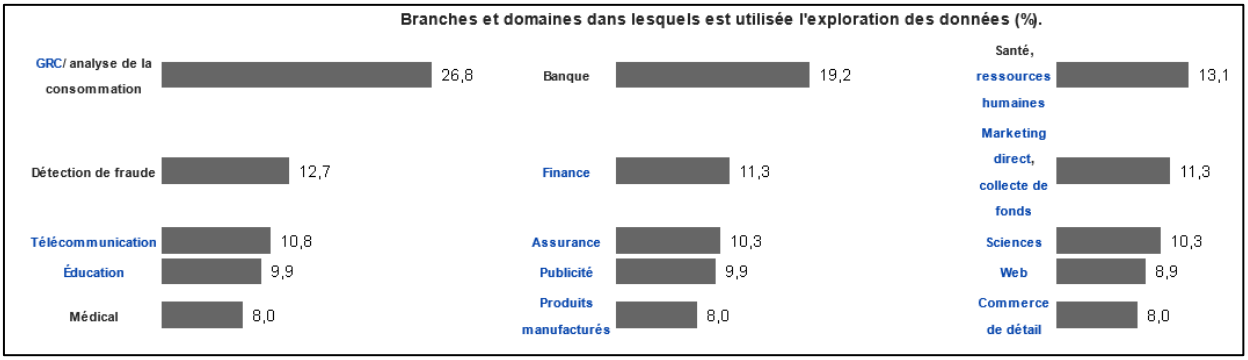


Figure 16: Branches et domaines utilisant le data mining

➢ <u>Les ressources humaines</u>

Apporter le *data mining* dans le monde des ressources humaines des Armées parait une évolution naturelle et incontournable pour tout simplement mieux connaître ce qui en fait leurs richesses : les femmes et les hommes qui les composent, pour encore mieux les gérer. En effet cette évolution permettrait d'une part de mieux comprendre certains phénomènes clés tout au long du cycle des RH : Le recrutement, la gestion des carrières et profils, l'évolution des parcours proposés, la formation, les départs et reconversions, les réserves…

Tout d'abord dans le recrutement, le *data mining* permettrait de mieux cibler les personnes susceptibles d'être motivées par un engagement sous le drapeau en les démarchant de manière plus directe et efficace que les campagnes de communication actuelle. En jouant à la fois sur une meilleure connaissance des profils sortants du système scolaire (via les *open data* fournies par le ministère de l'éducation nationale), du marché de l'emploi avec les organismes et entreprises « concurrents » des armées en termes de recrutement (2^{e} employeur national), et des besoins de recrutement en temps réels mais aussi à venir, les Armées pourraient accroître le taux de sélection, mais aussi diminuer le taux de dénonciation avant la fin des six mois de formation en renforçant la concordance entre attentes et réalité.

Cette approche très utilisée dans le marketing pourrait s'accompagner également de campagnes de communication mieux ciblées (et quand même indispensables) pour toucher le plus de candidats potentiels en quittant les supports traditionnels pour migrer vers des supports davantage en adéquation avec les profils recherchés. Ce changement de paradigme

est nécessaire et amorcé au sein de la DRHAT[24] à l'image du dernier reportage voulu non institutionnel pour mieux toucher l'audience cible en sollicitant une des stars de *Youtube* (Exemple : Reportage[25] « Je pars à l'armée » par TIBO sur *Youtube* vu par plus de 3 millions de *followers*).

Chaque année, le *data mining* permettrait d'automatiser et de faciliter le travail annuel dans la gestion du plan annuel de mutation (PAM) en prenant en compte davantage de paramètres pour répondre au mieux aux besoins de l'institution. La forte capacité de calcul inscrirait ce processus non seulement dans l'année du PAM en cours mais également sur les deux ou trois années suivantes pour davantage de cohérence sur le long terme et une meilleure lisibilité pour personnels étudiés.

Il apporterait une capacité de criblage supplémentaire et indéniable pour le gestionnaire dans la sélection et recherche de compétences rares et pointues. Cette capacité accrue de sélection permettrait, au travers des notations, qualifications, récompenses et autres paramètres ajustables, d'entériner de manière automatique et impartiale des processus de sélection davantage orientés sur dossier plutôt que sur concours pour gagner du temps, des ressources et donc de l'argent et ce en toute neutralité.

La *data mining* apporterait également une meilleure fidélisation de la ressource dans la durée. En effet par une meilleure connaissance des raisons des démissions, des départs anticipés, il permettrait de prédire les profils les plus enclins à quitter l'institution et les raisons possibles, fournissant au commandement un levier supplémentaire d'action pour retenir certains profils indécis et indispensables pour l'institution par le biais de mesures adaptées. Cette compréhension et anticipation accroitraient la fidélisation, améliorerait les indicateurs sur le moral en complément de ceux existant pour la chaîne de commandement, diminuerait les arrêts maladie pour le plus grand bonheur de notre système de santé. Ecouter et comprendre ses personnels est un acte fort et indispensable pour la chaîne de commandement. L'accompagner par le *data ming* accroitrait son efficience au demeurant.

A l'heure où les directions des RH des Armées s'inscrivent dans une gestion de flux, l'apport du *data mining* permettrait de mieux tester et de projeter les différents modèles envisagés sur le long terme (sur une à plusieurs décennies) afin de mieux répondre aux choix politiques souvent centrés sur le temps d'une mandature. Cet outil fournirait des projections qui viseraient à faciliter la gestion sans à-coup en évitant de « tirer constamment les bords »,

[24] DRHAT : Direction des ressources humaines de l'Armée de Terre
[25] https://www.youtube.com/watch?v=Ce4m_27G4Ts

procédé souvent destructeur en RH particulièrement dans un domaine régalien comme la Défense. Il permettrait également d'anticiper toutes les questions budgétaires sur le temps long à l'heure où les débats sont récurrents sur l'allongement du temps de travail.
Enfin avec un besoin croissant de réservistes opérationnels aux côtés des forces militaires traditionnelles dans des missions comme Sentinelle sur le territoire nationale, l'usage du *data mining* pourrait accroître l'efficacité des sollicitations en croisant les agendas électroniques des 40 000 réservistes. Trouver la ou les périodes idoines pour les réservistes par rapport au juste besoin des Armées, sans mettre en tension ou en déséquilibre l'activité des employeurs serait donc une des réponses possibles.

- Aide à la décision - accélération de la prise de décision

Vivant dans une société de l'immédiateté, les armées ne dérogent pas à la règle et se doivent de plus en plus d'intégrer en temps réel les éléments de situation pour permettre de proposer à l'échelon politique des réponses les plus appropriées et efficaces avec les moyens et forces à notre disposition tout en s'affranchissant du chaos informationnel croissant : *« Il (le milieu terrestre) impose un triple défi aux forces terrestres : […]*

- *un défi intellectuel, résultat de l'interconnexion de groupes humains et d'organisations entremêlés, qu'il faut appréhender et saisir ;*
- *un défi cognitif en raison de la masse de données générées par la nature même de l'environnement terrestre. Leur distribution par les technologies de l'information et leur amplification par la connectivité croissante des objets et des personnes, produisent un chaos informationnel. Il est un défi majeur au même titre que la viscosité et l'opacité du terrain ou la complexité des sociétés ».* (Etat Major de l'Arme de Terre, 2016)

Cette réponse attendue quasi-immédiatement et sans en altérer la qualité oblige donc le commandement militaire à disposer d'outils d'aide à la décision capables de les assister à tous les niveaux. Le *data mining* peut permettre de répondre à ce besoin que ce soit du niveau tactique en passant par le niveau opératif sur un théâtre comme la bande sahélo-saharienne (BSS) jusqu'au niveau politico-stratégique.

Ainsi au niveau tactique, le *data mining* peut apporter une réelle plus-value dans l'infovalorisation recherchée par les unités au sol. C'est d'ailleurs l'objectif du projet

SCORPION[26] de diminuer le brouillard de la guerre et les zones de friction pour apporter, en quasi-temps réel, l'information utile et nécessaire au chef tactique pour prendre la bonne décision à la fois, pour remplir la mission mais aussi préserver ses forces : *« Par ailleurs, le rythme élevé du tempo opérationnel imposera des interfaces homme-machine adaptées et performantes, alliant simplicité des procédures et intuitivité de mise en œuvre. Ce besoin militaire a déjà été exprimé. Il s'incarne progressivement par la numérisation des unités qui devra reposer sur des architectures de systèmes d'information et de communication performantes afin d'atteindre l'infovalorisation qui seule permettra le combat collaboratif ».* (Etat Major de l'Arme de Terre, 2016). Cet apport technologique doit cependant veiller à rester une aide subordonnée au chef tactique afin d'éviter l'écueil d'une technologisation à outrance qui déshumaniserait la nature même du combat. Cette vision clausewitzienne de la guerre a été perpétuée par le corps des *US Marines* qui ont su gérer et garder une certaine distance salvatrice à l'époque sur cette sur-technologisation issue de la *Revolution in Military Affairs* (RMA) telle que présentée dans La technologie militaire en question - le cas américain (Henrotin, 2008).

Au niveau opératif, une force de théâtre doit être à même d'appréhender la connaissance politique, ethnique, sociale et culturelle du théâtre tout en prenant en compte la sphère informationnelle inhérente à chaque pays : *« Les forces terrestres devront surmonter deux difficultés qui se combinent et troublent l'appréciation des situations opérationnelles. Il s'agit d'abord de la multiplicité des acteurs et des données de contexte à appréhender (plus d'information potentiellement utile) : variété des adversaires, cartographie des alliances et des réseaux (claniques, idéologiques, d'intérêt, d'influence, etc.), données sociologiques, codes culturels. Cette profusion perdurera et accentuera la complexité originelle des engagements terrestres, liée à la géographie, à l'évolution rapide des terrains, aux contraintes climatiques. À cette difficulté s'ajoutera le « vacarme informationnel » (trop d'information probablement inutile), résultat de la diffusion sans frontière des progrès technologiques. Conséquence de ce double phénomène, l'inflation de l'infomasse rend plus difficile le tri et la hiérarchisation des données. La compréhension partagée de cet environnement opérationnel compliqué et évolutif dans toutes ses dimensions, y compris humaine, est donc un élément-clé de supériorité qui doit permettre de décider d'une action militaire, de la planifier et de la conduire ».* (Etat Major de l'Arme de Terre, 2016). Cette

[26] SCORPION : Synergie du Contact Renforcée par la Polyvalence de l'infovalorisatiON

« compréhension » de l'environnement est désormais indispensable car susceptible de conférer à nos forces l'ascendant nécessaire sur l'adversaire, d'où sa définition comme un des huit facteurs de supériorité opérationnelle dans Action Terrestre Future : *« Fondée sur la conscience, l'analyse puis le jugement, la compréhension prolonge la connaissance pour lui donner une valeur réellement opératoire. Elle est l'aptitude à percevoir, interpréter et apprécier un environnement opérationnel complexe et évolutif en vue de fournir le contexte, la perspicacité et la clairvoyance requis pour la prise de décision ».*

Dans cet esprit, l'apport du *data mining* permettrait donc de pleinement intégrer tous ces paramètres complexes dans une zone d'opérations pouvant aller jusqu'à cinq pays dans le cas de l'opération Barkhane dans la BSS. Cet apport technique gommerait d'une part les biais propres à chaque culture humaine en évitant les préjugés par exemple, d'autre part elle permettrait d'inscrire cette compréhension dans sa globalité et sur le temps long : *« Elle (la compréhension) dépend d'abord d'un examen critique de notre propre système afin de révéler les biais culturels et intellectuels qui pourraient nuire à une analyse pleinement lucide. Elle requiert modernité et créativité pour employer des méthodes et adopter des points de vue originaux, destinés à produire des analyses alternatives sachant discerner l'immuable de l'évolutif. La véritable compréhension nécessite également une continuité géographique, temporelle pour inscrire la réflexion dans le temps long et construire une réelle mémoire des théâtres d'engagement, et enfin technique grâce à des moyens de fusion et de stockage de l'information. Autant que possible, elle sera enrichie des apports des autres acteurs puis partagée afin de faciliter la coopération. Enfin, pour être un atout décisif, elle devra s'exprimer de la manière la plus simple possible et primer sur d'autres compréhensions individuelles ou collectives concurrentielles. »* (Etat Major de l'Arme de Terre, 2016)

Enfin au niveau politico-stratégique, le *data mining* permettrait surtout un gain de temps nécessaire pour maintenir cette initiative des chefs mais aussi des dirigeants politiques face aux attentes voire « attaques extérieures » de tous genres telles celles des médias, mais aussi des organismes internationaux *et cetera*…Il contribuerait directement à accroître cette performance du commandement, indispensable pour s'attaquer aux vrais défis et enjeux de Défense. De même, dans ces sphères du pouvoir et de la réflexion, le *turn-over* du personnel reste assez élevé et limite l'instauration d'une mémoire (humaine) dans chaque service, ce qui reste préjudiciable pour l'efficacité des dits-services et bureaux et donc pour le fonctionnement global. Le *data mining* peut ainsi permettre de combler ce déficit grâce à cette

capacité d'analyse des données existantes et autres travaux déjà menés. Cette composante du *data mining* est appelée *DataDiscovery*. Une autre de ces composantes plus démonstrative, le *dataviz*, serait susceptible d'être utilisée dans ces services où les *reportings* et autres tableaux de bord sont monnaies courantes. En effet, grâce à ces outils, ces échelons de synthèse pourraient présenter des analyses dynamiques plus poussées, davantage visuelles et compréhensibles, rendant la performance encore supérieure.

- Autres domaines :

Le soutien est un des domaines les plus affectés par le *data mining*. Son recours apporterait aux armées d'une part de diminuer les stocks toujours couteux, même si la politique des flux poussés a déjà fortement contribuer à réduire ces derniers. Ensuite à l'heure des objets connectés et autres puces *RFID*[27] équipant de plus en plus de pièces de rechanges, il devient de plus en plus facile de suivre en temps réel l'approvisionnement mais aussi l'usure de ces dernières. Cette vision permet d'une part d'anticiper les pannes, de rallonger leur durée de vie, de détecter plus facilement des problèmes issus de mauvais paramétrages mais aussi des problèmes de conception relevant de la compétence du constructeur ou encore de conduite et ou d'entretien. Cette anticipation peut permettre une meilleure disponibilité technique opérationnelle (DTO) et donc de fait, accroitre la disponibilité des matériels dans les différents parcs d'entraînement mais aussi au sein des PSP des régiments. Elle permettrait également de pré-positionner au plus près les besoins de la force calculés au plus juste à l'instar d'Amazon avec l'approvisionnement de ces centrales : *«* Amazon *planifie et positionne ses stocks au plus près des demandes anticipées par l'analyse des traces des actes de consommation »* (Baudot, 2015). Toute cette gestion numérique dans la *supply chain* peut permettre au final aux Armées et plus particulièrement à l'Armée de terre d'améliorer sa préparation opérationnelle à un moment crucial dans le cadre de la mise sur pied et validation du modèle « Au Contact ». Dans le prolongement, le maintien en condition opérationnelle (MCO) pourrait être directement impacté par cet apport technologique en réduisant les temps d'immobilisation des véhicules.

Tous les services de contrôle, de la sécurité-protection (des emprises), aux finances en passant par les inspections d'Armées, tous ces secteurs devraient être intéressés par le *data mining*.

[27] *RFID : radio frequency identification*

Son usage pourrait mettre en évidence des tendances, des liens de cause à effet pas forcément évident en première approche ou encore dénicher des signaux faibles susceptibles d'alerter le plus tôt possible le commandement et autres autorités pour, comme dirait l'adage, prévenir plutôt que guérir. Son usage sous forme d'*image mining* (déjà utilisé par la police aux frontières dans les aéroports) permettrait d'améliorer la détection des profils inconnus et des comportements suspects ou inhabituels que des sentinelles se relevant régulièrement ne peuvent discerner dans le cadre de la sécurité protection des emprises militaires. De fait, cette plus-value améliorerait l'efficacité des gardes d'emprises et diminuerait la ressource en personnels mobilisés par le plan Cuirasse (protection des emprises militaires) pour les réinjecter plus facilement dans d'autres opérations ou entraînements. Cet usage amplifierait le contrôle en donnant une sorte de deuxième flair, il réduirait les fraudes et autres abus possibles. Il permettrait sur d'autres thématiques notamment financières d'ajouter plus de justesse et de transparence à chaque échelon dans le traitement des dossiers, ce qui contribuerait à améliorer indirectement le moral en évitant par exemple les déboires des trop-perçus apparus à la suite du dysfonctionnement du logiciel de solde Louvois.

Cette multitudes des usages est résumée par Pierre-Yves Baudot : *« Différents métiers sont affectés par la progressive distinction proposée entre la prise de décision et l'exécution, qu'il s'agisse des douaniers dont « le flair » est remplacé par un algorithme leur permettant de sélectionner les cibles de fouilles, des policiers dont les patrouilles sont déterminées par algorithmes (Benbouzid, 2015) ou encore des professionnels de la politique, dont le travail politique repose moins sur un charisme ou un savoir-faire que sur la précision du ciblage du porte à porte et de la personnalisation des mailings »* (Baudot, 2015). Le *data mining*, par simple déclinaison de ses applications du milieu civile vers le milieu de la Défense pourrait déjà rapidement apporter une réelle plus-value aux forces armées dans leurs fonctionnements organique comme opérationnelle à tous les niveaux : directement sur le terrain, dans un théâtre d'opération, ou au cœur du pouvoir.

3.2 Le *data mining* au cœur du renseignement numérique

Mais c'est sans aucun doute dans le domaine du renseignement que les possibilités du *data mining* semblent les plus prometteuses. En effet, dans cette société hyper connectée, le cyberespace constitue une base de renseignement évidente et incontournable à condition de disposer de cette capacité technique d'exploitation et de transformation des *data* en information puis renseignement (cf. 2.2). Cette sous-partie détaillera la méthodologie d'obtention du renseignement cyber et enfin l'enjeu lié à son intégration et fusion avec les autres sources de renseignement en temps réel pour gagner en efficacité.

➢ <u>De la donnée à l'intelligence (méthode d'enrichissement de l'information)</u>

Une fois ces données harmonisées et donc exploitables il convient de les enrichir à plusieurs niveaux pour en tirer du renseignement, encore appelé intelligence, non seulement dans le monde économique mais aussi dans le monde anglo-saxon. Le schéma ci-contre présenté dans le sous-chapitre 2.2 rappelle ce processus d'enrichissement, étape par étape, jusqu'à obtenir de l'intelligence.

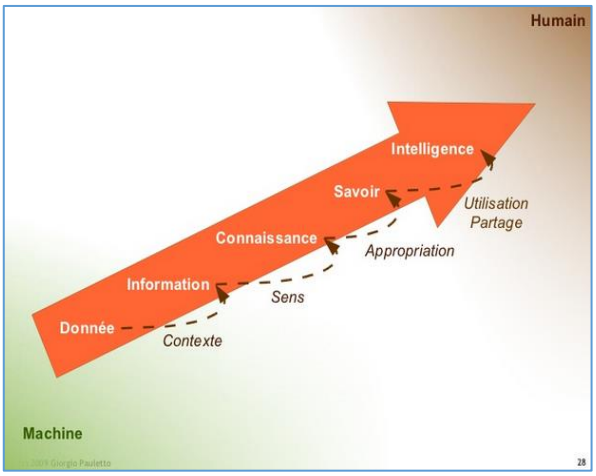


Figure 17 : Processus d'enrichissement de la donnée (Source : Giorgio Pauletto)

Pour permettre l'obtention de renseignement utile pour les armées, il est donc nécessaire de commencer par collecter des données. La figure ci-contre donne une idée des nouveaux champs de collecte possibles et à investiguer. On voit donc l'enjeu de stockage lié à l'appropriation de ces milieux, pour pouvoir ensuite en tirer après enrichissement, croisement et fusion, du renseignement cyber. Que ce soit du renseignement

Les nouveaux champs de collecte

Le Web 2.0:
- l'Internet tel que l'utilisateur le perçoit généralement. Consultation de pages, sites d'information spécialisés, blogs, sont autant de premiers éléments constitutifs de l'analyse.

Le *deep web*:
- le reste de l'Internet, non-indexé par les moteurs de recherche.

Les réseaux sociaux ;
Les messageries instantanées ;
Les champs clos de données (banques, santé, services particuliers) ;
Le cloud (informatique en nuage) ;
La téléphonie mobile ;
Les supports mobiles (Smartphone, tablettes) ;
Les objets connectés ;
Les métadonnées (sur divers supports, communications, images).

Figure 18: Les nouveaux champs de collecte (Boyer, 2014)

d'origine cyber (cf. 3.2) destiné à compléter les autres formes de renseignement obtenus dans un domaine ou secteur par exemple, en incorporant alors la sphère numérique ou que ce soit du renseignement plus technique qualifié de renseignement d'intérêt cyber (cf 3.3) destiné à être utilisé directement dans l'espace numérique, toutes ces formes de renseignement se nourrissent désormais des *data* présentes dans le cyberespace.

- L'espace numérique comme source d'information : le renseignement cyber

Thomas Rid, écrivain, chercheur et spécialiste sur les risques cyber, déclarait : *« Getting data is now easier, but not using them »*. Cet enrichissement des données constitue bien LA phase complexe pour obtenir du renseignement. Malheureusement la pensée stratégique militaire sur les rapports entre le cyberespace et le renseignement demeure peu développée comme le souligne Bertrand Boyer dans Cybertactique, conduire la guerre numérique : *« Pourtant, si l'on trouve une riche littérature prospective sur la « cyberguerre », on trouve assez peu de réflexion sur le rapport entre cyberespace et renseignement. […] la pensée stratégique a bien peu développé les liens entre renseignement et cyber»*.

Cependant, cette pensée stratégique existe, de manière embryonnaire et assez secrète pour éviter de dévoiler des techniques aux potentiels adversaires et surtout pour éviter une médiatisation inutile qui desservirait à tort l'institution, à l'image de l'affaire PRISM aux Etats-Unis. On peut cependant trouver quelques pistes en lisant cet ouvrage très pédagogique: *« En définitive, le cyberespace signe l'arrêt de mort du concept de « sources ouvertes » […] L'open source (OSINT) n'est donc que le socle qui permet de re-contextualiser le renseignement recueilli par ailleurs par des capteurs « nobles » humains ou techniques. Dès lors les sources ouvertes n'existent plus, car elles sont le produit d'une époque où l'on opposait l'ouvert et le fermé, l'information et le renseignement. Or, les nouveaux champs d'investigation ont profondément changé cette vision et ouvrent sur une nouvelle dichotomie. Ainsi, au concept ouvert-fermé, il convient de substituer celui de brut-traité. Le cyberespace impose à la fonction renseignement de redéfinir la nature même de son objet. Le renseignement n'est plus aujourd'hui, une information cachée (secrète), mais une construction fusionnant plusieurs bribes d'informations afin de dégager du sens. La difficulté n'est plus alors forcément d'avoir accès à une information, mais plutôt le traitement de celle-ci, sa cotation et sa remise en perspective. A la problématique d'accès, nous sommes*

aujourd'hui confrontés à la problématique de la compréhension. Plus que du renseignement caché, il faut aujourd'hui, à la fonction renseignement, compiler, analyser, donner du sens à la masse de données disponibles ». C'est toute l'utilité et l'intérêt du *data mining*.

Cette puissance du renseignement cyber peut se comprendre et s'illustrer au travers de l'opération de piratage russe *« Grizzly Steppe »* découverte et révélée le 29 décembre 2016 par deux agences américaines (le *Departmement of Homeland Security (DHS)* et le *Federal Bureau of Investigation (FBI)*). Cette opération cyber orchestrée par les services de renseignement russes visait à perturber les élections américaines en influant et favorisant l'élection de Donald Trump au détriment d'Hillary Clinton. Cette opération d'ingérence a réussi d'une certaine manière puisque Donald Trump est devenu le 20 janvier dernier le 45^{e} président des *USA*.

Cette attaque a débuté très en amont en collectant du renseignement sur des membres du parti démocrate au travers d'une campagne d'hameçonnage par mails (ou *fishing* en anglais) destinée à les convaincre d'installer un programme malveillant ou de fournir leurs identifiants. La faiblesse d'au moins un des destinataires a permis au virus de s'introduire dans le système d'information (SI) du parti démocrate. Cette première phase réalisée par le groupe APT29 a permis de glaner un certain nombre d'informations et de correspondances du parti démocrate. Une deuxième campagne d'hameçonnage orchestrée par le groupe APT28 a permis de poursuivre et compléter l'action entreprise par le premier groupe (APT29) en exfiltrant des informations sensibles sur les hauts placés du parti. Ensuite ces informations ont été exploitées et distillées à la presse à bon escient de manière à troubler la campagne présidentielle d'Hillary Clinton en influençant les opinions publiques. Au final, cette attaque l'a affaiblie en la faisant passer de favorite à perdante de l'élection, en l'espace de quelques mois.

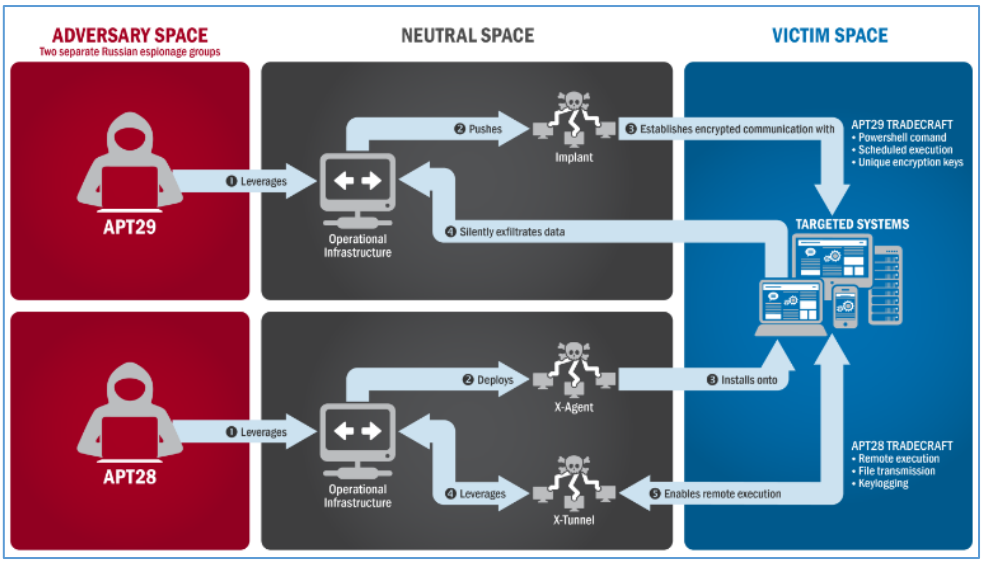


Figure 19: The tactics and techniques used by APT29 and APT28 to conduct cyber intrusions against target systems - Source : (NCCIC, 2016)

A travers ces deux schémas (Figure 19 et Figure 20), on voit bien le renseignement numérique obtenu à partir du cyberespace avec tout d'abord la connaissance des membres du partis grâce à leur empreinte numérique sur *Internet*, leurs faiblesses humaines et personnelles en ciblant les membres les plus vulnérables ou les moins sensibilisés à la sécurité informatique. Ensuite par l'action malencontreuse d'un des membres ciblés par l'attaque, un logiciel malveillant a

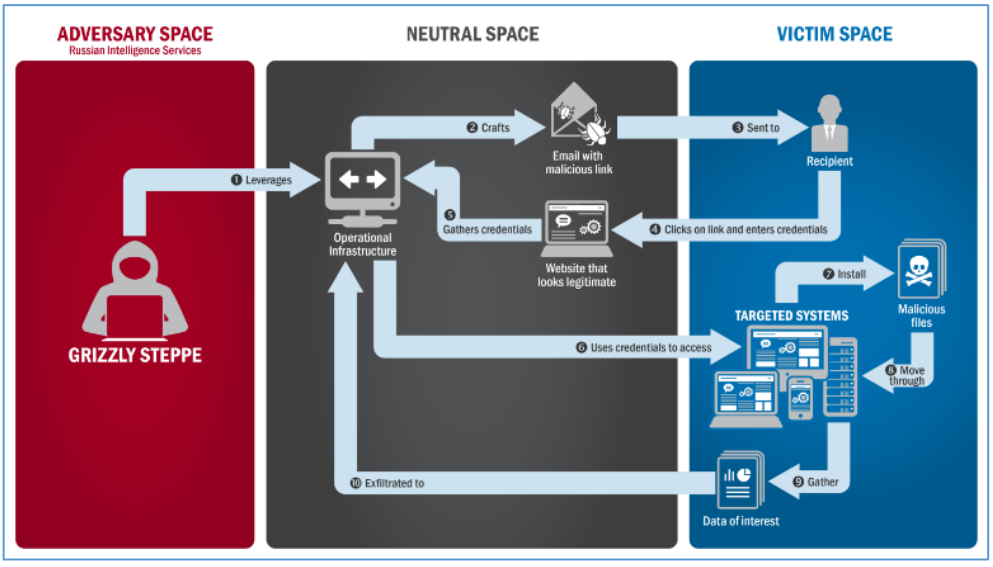


Figure 20 : APT28's use of spear phishing and stolen credentials - Source : (NCCIC, 2016)

pu être installé pour mieux connaître techniquement le SI du parti démocrate afin de faciliter la recherche de données sensibles puis leurs exfiltrations : Il y a donc recherche de renseignement d'intérêt cyber (RIC) dans ce cas-là.

Dans ces différentes étapes, le *data mining* a permis de gagner du temps en permettant une analyse rapide données et en faisant ressortir les vulnérabilités critiques.

Au final, ces actions russes se retrouvent dans la description des trois sous-domaines du renseignement de cyberdéfense, tels que le décrit Bertrand Boyer (cf. Figure 21):

- Renseignement à fin d'action,
- renseignement technologique,
- renseignement de situation.

➢ <u>Un besoin d'intégration du renseignement cyber avec les autres formes de renseignement</u>

Maintenant le renseignement cyber ne se suffit pas à lui seul. Pour devenir renseignement, les informations doivent au préalable être croisées voire fusionnées avec les autres renseignements traditionnels que sont les renseignements d'origine humain (ROHUM), électromagnétique (ROEM) et imagerie (ROIM) et leurs corollaires anglais, *HUMINT*, *SIGINT*, *IMINT*, si possible en temps réel. Ce besoin nécessite d'énormes capacités techniques et d'analyse d'où le recours nécessaire aux techniques du *data mining* mises en oeuvre par les *data scientists* mais aussi par des spécialistes du renseignement de manière à conserver et entretenir cette culture du renseignement sans l'automatiser pour éviter tout phénomène d'aliénation : *« Loin d'être exclusivement technique, la réponse doit impérativement intégrer les apports de l'analyse humaine créant de facto une similitude avec les modes opératoires du renseignement classique »* (Boyer, 2014). De la même manière, les

Sous-domaine	Objet	Description	Production / sortie
Renseignement à fin d'action	Contre quoi / pourquoi	Dans le cadre d'une préparation d'opération, il s'agit de recueillir le maximum de données exploitables sur la structure cible, sur le système d'information visé. Ces données peuvent être de nature technique (plan d'adressage, matériel, architecture...) mais également humaine (connaissance des administrateurs, évaluation des compétences des équipes de maintenance, plan de sauvegarde, utilisation du social engineering...)	Dossier d'objectif
Renseignement technologique	Avec quoi	Connaissance des outils d'attaque, évolutions protocolaires, nouveaux dispositifs de surveillance, vulnérabilités, exploits, équipements spécifiques...	Outils ; dossier technique ; manuel d'utilisation.
Renseignement de situation	Contre qui	Connaissance des structures et des acteurs du cyberespace qui peuvent entraver l'action. Connaissance du cadre juridique spécifique aux actions cyber dans les pays cibles et/ou origines d'attaque.	Dossier pays ; matrice des risques ;

Figure 21: Les 3 sous-domaines du renseignement de cyberdéfense (Boyer, 2014)

renseignements classiques se doivent de prendre en compte ce nouveau milieu omniprésent en l'intégrant dans chaque action de renseignement (cf Figure 22) :

Type de renseignement	Caractéristique	Conséquences
Renseignement pour le cyberespace	3 sous-domaines: à fin d'action; technologique; de situation. Le cyberespace constitue un milieu à part entière avec ses besoins propres en renseignement pour y conduire des opérations	La fonction renseignement doit intégrer les nouveaux besoins. Par ailleurs, la capacité offensive est directement impactée par l'atteinte des objectifs de renseignement.
Renseignement dans le cyberespace	Mise à disposition permanente d'une information de masse. Nouveaux champs de collecte du renseignement.	Disposer des outils permettant le traitement et le stockage de l'information de masse. Adapter les schémas de traitement et de recueil du renseignement aux nouveaux champs.
Renseignement par le cyberespace	Le cyberespace augmente la surface de vulnérabilité des cibles.	Besoin d'intégration du volet "cyber" dans chaque action de renseignement.

Figure 22: Impact du cyberespace pour la fonction renseignement (Boyer, 2014)

« L'émergence du monde numérique, sa prégnance et sa pénétration dans tous les aspects de la vie de nos sociétés supposent donc de repenser la place du renseignement. Impliquant la mise en œuvre de nouvelles modalités d'acquisition pour satisfaire de nouveaux besoins, l'espace cyber offre également des opportunités pour améliorer les méthodes classiques de renseignement. L'action « renseignement » peut se concevoir suivant trois axes distincts qui imposent chacun la création de nouveaux outils et de nouvelles procédures. La fonction renseignement se développe ainsi naturellement dans, par et pour le cyberespace. »

En guise de conclusion, *« le renseignement « dans » le cyberespace est donc un renseignement de masse qui s'oppose à la vision historique de la recherche classique qui visait à atteindre l'information secrète, celle qui était cachée. Aujourd'hui l'information n'est plus cachée dans un coffre mais diluée dans un océan de donnés. Le défi n'est donc plus la « collecte » mais bel et bien l'analyse et la capacité à donner un sens à la masse d'informations collectées. Dans ce contexte, plusieurs travaux font paraître tout l'intérêt de développer des outils de traitement statistiques automatisés afin de dégager des éléments d'anticipation, et constituer ainsi un véritable capteur d'alerte »*. (Boyer, 2014)

Sans même l'évoquer, la notion de *data mining* prend tout son sens à travers ce nouveau paradigme du renseignement, lié à la spécificité de ce milieu qu'est le cyberespace. Par ailleurs, cette vision développée par Bertrand Boyer converge avec la vision et les besoins de l'Armée de Terre développés dans Action Terrestre Future (cf Figure 23).

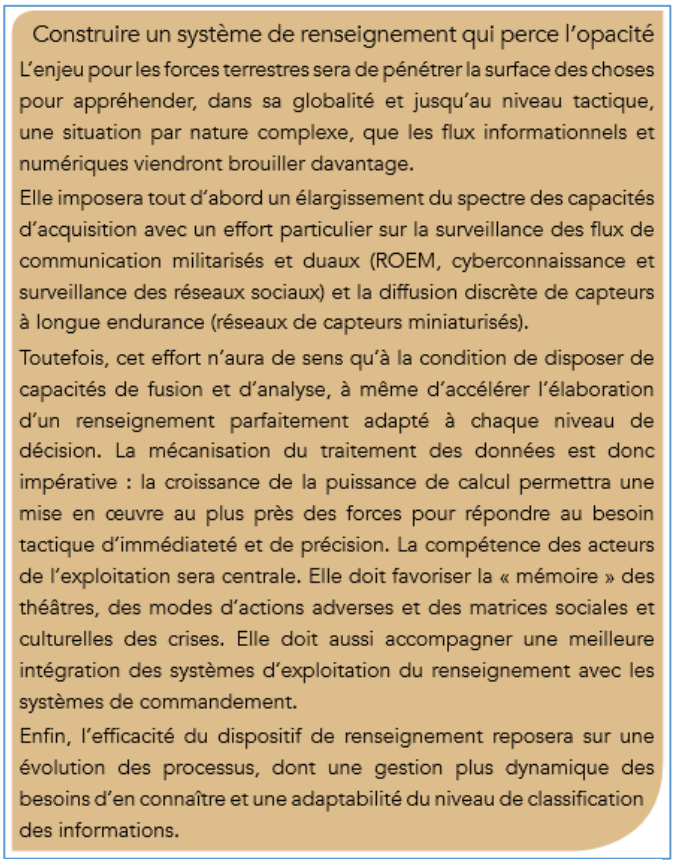


Figure 23: Construire un système de renseignement qui perce l'opacité (Source : ATF, 2016)

3.3 Le *data mining*, bras armé également de la cybersécurité

La cybersécurité peut se décomposer en deux volets : la cyberprotection qui veille à protéger nos systèmes d'informations (image de la muraille du château fort) et la cyberdéfense qui vise à combattre toute intrusion (image de combattants armés à l'intérieur du château fort). Le *data mining*, est déjà très utilisé en cybersécurité parfois à son insu, mais son usage ne va cesser de s'amplifier pour plusieurs raisons : il contribue directement à mieux contrôler la gestion de l'information sur tous les réseaux utilisés par les états-majors numérisés (presque la totalité), de plus il permet d'analyser les comportements de chaque utilisateur pour mieux détecter toute intrusion ou comportement anormal renforçant de fait la cyberprotection des systèmes d'informations. Enfin, il permet en analysant toujours davantage de logs[28] divers et variés de mieux comprendre certaines intrusions, de récupérer du renseignement d'intérêt cyber voire de paramétrer la détection de signaux faibles et contribue ainsi directement à renforcer nos capacités de cyberdéfense.

- Un outil de management de l'information dans tous les états-majors

Comme le numérique sert de support à tous les états-majors pour travailler au quotidien, son fonctionnement est régi par deux principes : le niveau de classification et le besoin d'en connaître (arrêté du 30 novembre 2011 portant approbation de l'instruction générale interministérielle sur la protection du secret de la défense nationale). Ces deux principes visent tous deux à empêcher les fuites d'informations en protégeant d'une part la sensibilité, d'autre part en limitant l'accès à l'information juste aux personnes concernées par le sujet. Ce besoin de préserver le renseignement est nécessaire et compréhensible car considéré comme sensible et contribuant directement à l'appréciation autonome de situation. Aussi, il permet également d'asseoir la souveraineté et l'indépendance de la France sur la scène internationale.

A ce titre, la fonction *IKM* (*Information and Knowledge Management*) ou CMI (Cellule de management de l'information) que l'on retrouve dans n'importe quel état-major, prend une importance croissante en raison de l'omniprésence du numérique mais surtout d'une menace croissante de fuite interne (*insider threat*) à l'image de Julian Assange qui a exfiltré plus de 11 millions de documents sensibles aux services américains : *« Enfin, les révélations en*

[28] Log : journal d'événement

cascades sur les programmes de surveillance massifs de l'Internet jettent un trouble nouveau sur ces questions et y mêlent le débat sur la protection de la vie privée. Les données librement diffusées, les communications les habitudes de connexions, autant de paramètres que nous croyons inviolables, se retrouvent au centre du débat qui in fine aliment la réflexion sur la nature du renseignement dans le cyberespace » (Boyer, 2014). On perçoit donc toute l'importance d'avoir des systèmes interopérables, c'est-à-dire qui parlent entre eux, pour détecter ces anomalies comportementales au travers du *data mining* : impression, téléchargement, copies de documents sensibles croisées avec les comportements des utilisateurs, leurs habilitations, leurs droits d'en connaître mais aussi les horaires de services, de présence, l'absence d'activité sur les autres ordinateurs d'une même pièce par exemple. Cette liste est bien évidemment ajustable en fonction des besoins et requêtes. Ces éléments diminueraient le risque de fuite sans toutefois l'annuler. Il renforcerait par exemple la sureté des systèmes d'information du Ministère et de ses services de renseignement mais aussi du gouvernement et des OIV. Cet apport technique permettrait par croisement des données d'améliorer les capacités en temps réel de contre-ingérence actuellement assurées par la DRSD[29]

- Un outil capable d'apprentissage comportemental en cyberprotection

Le *data mining* contribue également à renforcer la muraille protectrice des SI de la Défense en la rendant encore plus active en corrélant toujours davantage de logs générés par tous les différents systèmes de sécurité : *SIEM*[30], *IPS*[31], *IDS*[32], firewall, antivirus…En corrélant tous ces logs, le *data mining* apporte de l'interconnexion et de la compréhension aux systèmes de défense pour les rendre plus efficaces entre eux.

Cette compréhension peut se faire d'abord par une approche par scenario avec par exemple l'apprentissage des signatures d'attaques (*pattern matching*) permettant ensuite de les reconnaître ou encore de caractériser les techniques d'attaques dans leur ensemble. Cette technique nécessite la connaissance au préalable de la signature, ce qui peut parfois prendre

[29] DRSD : Direction du renseignement et de la sécurité de la Défense
[30] *SIEM: Security information management system*
[31] *IPS: Information protection system*
[32] *IDS: Information detection system*

des mois voire des années comme pour Stuxnet[33]. Par ailleurs cette signature reste passive au sens ou une simple modification ou évolution de la signature de menace, ne permettra pas de détecter cette nouvelle menace qui aurait juste « mutée ». Pour cela d'autres méthodes existent comme l'analyse de protocoles qui visent à détecter des attaques inconnues qui dégraderaient les performances des préprocesseurs, ou encore l'analyse heuristiques qui à travers une analyse intelligente (grâce au *data mining*) détecteraient des activités suspectes.

Cette compréhension peut aussi se faire sur l'approche comportementale en étudiant le comportement passé de l'utilisateur. En prenant le cas d'un employé travaillant sur son ordinateur de 9h à 17h tous les jours de la semaine hors week-end, si son ordinateur venait à exfiltrer des données un dimanche vers 20h, ce comportement suspect générerait de fait une alerte, puisqu'en dehors du comportement habituel de l'intéressé. Cette alerte permettrait à la permanence d'intervenir et d'arrêter ce vol de données. Cette stratégie n'est pas suffisante puisque contournable avec un peu de renseignement sur les pratiques de la cible. Elle peut se voir complétée d'une approche probabiliste relative au fonctionnement des logiciels en suivant les protocoles utilisés par exemple, ou encore par une approche statistique basée sur la connaissance de l'utilisateur (vitesse de frappe au clavier, sites les plus visités…) : *« Data Mining may be thought of as the most interesting one in accomplishment of intrusion detection and intrusion prevention system. In IDS and IPS, Data Mining used for to discover consistent and useful patterns of system features that describe user behavior»*. (Ashok, D, & Rohini, 2011)

Tout l'intérêt du *data mining* est alors d'arriver à agréger un maximum de logs souvent de formats différents en les croisant et combinant par des approches en temps réel, pour permettre d'affiner la détection et de diminuer le nombre de fausses alertes (faux positifs). L'objectif consiste à ne faire ressortir à travers ces filtres que les alertes les plus significatives et susceptibles de correspondre à une attaque réelle, sans en oublier. Ce volume d'alertes doit donc être réduit à un seuil acceptable pour permettre l'exploitation humaine. L'être humain devra alors apprécier grâce à son expérience métier et sa connaissance des systèmes la véracité de l'attaque. Ce recours au *data mining* pour la détection d'attaques nécessite tout un environnement adapté et une redéfinition des procédures, comme l'explique l'organisation MITRE[34] dans leur article *Data Mining for Network Intrusion Detection: How to Get Started*:

[33] Stuxnet : cyberattaque orchestrée pour retarder le programme iranien d'enrichissement d'uranium à des fins militaire

[34] The MITRE corporation : organisation américaine à but non lucratif qui gère les centres de recherche et de développement financés par le gouvernement fédéral et les agences.

"We have described our experiences with integrating data mining into a network intrusion detection capability. We believe that when starting such a project you should:

- *Choose your requirements carefully and be realistic.*
- *Assemble a team with broad, relevant capabilities.*
- *Invest in adequate infrastructure to support data collection and data mining.*
- *Design, compute, and store appropriate attributes with your data.*
- *Reduce data volume with filtering rules.*
- *Refine the overall architecture for your system, taking into account both automated*
- *processing and human analysis.*
- *Use data mining techniques such as classification, clustering, and anomaly detection, to*
- *suggest new filter rules.*

Make sure that automated data processing can be done efficiently". (The MITRE Corporation)

➢ <u>Un outil de renseignement d'intérêt cyber (RIC)</u>

Le *data mining* peut également servir de renseignement d'intérêt cyber (RIC), c'est-à-dire utiliser les données pour les enrichir en diversifiant les sources jusqu'à en faire du renseignement directement en lien avec le monde cyber. Pour faire le lien avec la sous-partie précédente, ce RIC peut se matérialiser par la mise en place de règles automatiques de filtrage, par exemple, par simple croisement de l'analyse du trafic entrant-sortant et de comportements suspects. C'est ce qu'avait réalisé et démontré Wenke Lee au travers de son modèle MADAM ID : *"This thesis describes a novel framework, MADAM ID, for Mining Audit Data for Automated Models for Intrusion Detection. Classification rules are inductively learned from audit records and used as intrusion detection models. A critical requirement for the rules to be effective detection models is that an appropriate set of features need to be first constructed and included in the audit records. A key contribution of the thesis is thus in automatic "feature construction". Using MADAM ID, raw audit data is first preprocessed into records with a set of "intrinsic" (i.e., general purposes) features. Data mining algorithms are then applied to compute the frequent activity patterns from the records, which are automatically analyzed to generate an additional set of features for intrusion detection purposes."* (Lee, 1999). Cet usage du *data mining* sous sa méthode prédictive permettra d'adapter en permanence et en quasi-temps réel des contre-mesures grâce au RIC obtenu. Il

permettra de maintenir cette avance technologique dans un domaine où cet avantage est régulièrement contesté voire remis en question même par des armées dissymétriques (Forces Armées Syrienne par exemple) ou asymétriques (type Daesh) préférant acheter voire sous-traiter ces compétences à des organisations criminelles transnationales voire spécialisées. *« Ainsi, l'avantage technologique occidental pourrait ne pas demeurer un postulat stratégique valide. Les révolutions des nano et biotechnologies, intelligence artificielle et sciences cognitives (NBIC) sont synonymes d'accélération mais aussi de privatisation et de marchandisation du progrès. Les États se retrouveront ainsi contestés par la dissémination des avatars guerriers de l'innovation et l'irruption de nouveaux acteurs (lanceurs d'alerte, pirates informatiques). La domination technique deviendra donc relative, dans les champs informationnel et cybernétique en particulier […] Est foudroyant celui qui parvient à frapper l'adversaire dans ses flux électromagnétiques, ses systèmes informatiques, ses capacités de navigation et de localisation, ses perceptions. »* (Etat Major de l'Arme de Terre, 2016)

C'est pourquoi ce RIC déjà très utile en défensif comme évoqué ci-dessus, peut aussi apporter une plus-value dans une approche active voire offensive pour gagner en foudroyance : *« il (le renseignement cyber) est un élément primordial de la constitution de l'arme numérique. En agissant sur la couche logique, l'attaquant doit contourner un certain nombre de difficultés techniques (pare-feu, antivirus, sondes, réseau…). Dans ce contexte, pour être efficace, l'arme doit connaître sa cible. Cette connaissance de la cible ne peut s'obtenir que par une démarche active de renseignement… »* (Boyer, 2014). Ce besoin complexe de connaissance du cyberespace se retrouve donc à la croisée du ROC et du RIC. *« Le renseignement spécifique au cyberespace, « pour » le cyberespace, présente de nombreuses particularités qui rendent son développement complexe. En effet, si l'objectif demeure, dans un premier temps, l'attribution d'une attaque (mode réactif) et l'analyse de la menace (mode anticipatif – décrit supra), […] Dans le cyberespace, un adversaire potentiel s'appuie sur deux capacités complémentaires : des outils et des structures. Les outils sont les constituants de « l'arme » numérique (charge utile, perforateur) alors que les structures permettent son déploiement, le recueil et le traitement des informations (vecteurs, serveurs de commande et de contrôle). Mais ces deux critères d'analyse ne suffisent pas à déterminer des cadres de recherche pour le renseignement : il faut lui adjoindre l'homme et les compétences, qui seuls permettent de faire un lien entre outils et structures. »* (Boyer, 2014)

3.4 Le *data mining* indissociable de l'influence et en appui des forces

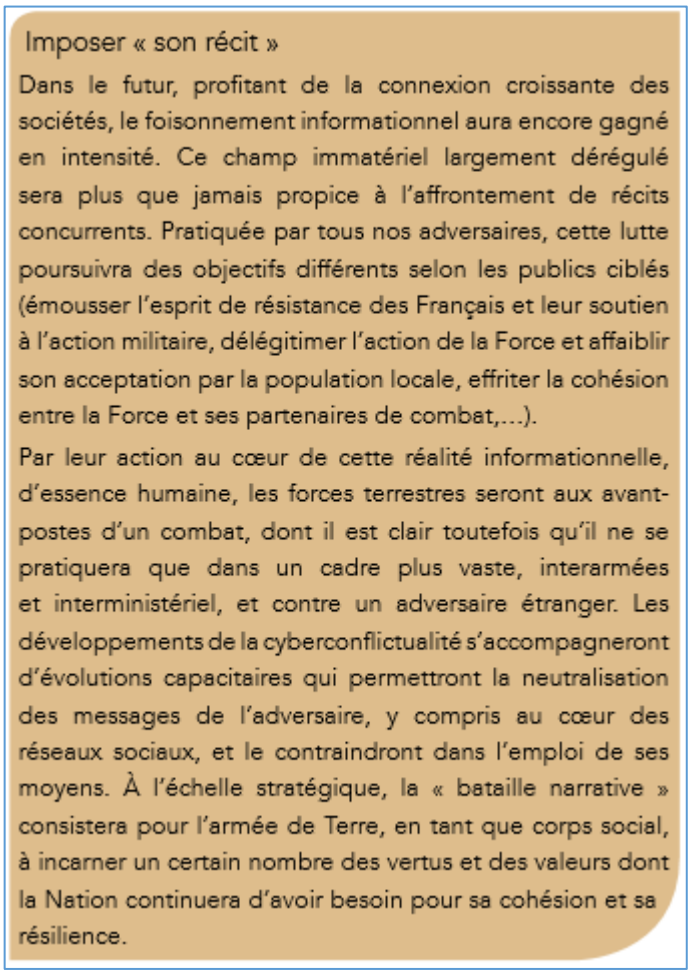
C'est également dans le domaine de l'influence que le *data mining* pourrait apporter le plus dans notre société bercée par l'immédiateté des faits et la crédibilité des *followers* (cf. 1.3). Les armées doivent donc investir ce domaine d'avenir dans sa composante numérique en l'intégrant pleinement, à l'image des forces terrestres qui en ont fait un facteur de supériorité opérationnel. Elle fait l'objet d'une sous-partie par son importance et les plus-values possibles dans le champ cognitif, domaine clé, d'intérêt et d'avenir pour légitimer l'action des Armées.

- Le *data mining* pour imposer son récit et protéger l'institution.

Cette maîtrise de l'influence constitue un changement radical de portage pour cette institution d'habitude peu prolixe dans ce domaine. La raison première réside dans la difficulté à garder un secret dans un monde hyper connecté. La question n'est donc plus de savoir si les faits vont sortir ou pas dans la presse mais plutôt quand ? C'est pourquoi il paraît légitime dans ce monde numérique d'être pro-actif et d'influer en permanence de manière à disposer d'une empreinte permanente positive (*e*-réputation) sur la Toile. Ce travail permettra en cas de coup dur d'atténuer le « choc » d'un incident desservant la cause militaire, par exemple en opposant une image favorable acquise sur le long terme : « *Depuis des décennies d'opérations menées sur des théâtres très divers, la maîtrise de l'information au sens large, dont les répercussions au sein de la société peuvent être redoutables, est devenue un enjeu majeur. Le moindre événement peut instantanément faire le tour du monde, sans recoupement, vérification ni analyse. La sensibilité immédiate et médiatique remplace alors, dans certains cas, la raison politique et peut parfois déterminer le succès ou l'échec d'une opération, indépendamment du sort des armes. Portés par l'hyperconnectivité et l'individualisation des sociétés, ces phénomènes échapperont demain à tout contrôle. La manière dont sera présentée l'action de la force sur le théâtre et surtout la manière avec laquelle elle sera perçue par un public large dans et en dehors de la zone des opérations sera dès lors un facteur déterminant de supériorité opérationnelle.* » (Etat Major de l'Arme de Terre, 2016)

Dans ce cadre, le *data mining* veillera à protéger l'image de l'institution en faisant référence à un certain nombre de valeurs socles en sélectionnant et diffusant toutes les données utiles

partageables et favorables à la force pour instaurer cette réputation dans la durée. Elle sera alors plus difficilement contestable puisque digérée et pérennisée régulièrement par les robots d'indexation des géants d'Internet. Par ailleurs, le *data mining* permettra de connaitre les info-cibles de manière individualisée et spécifique dans la durée pour adapter les messages au plus juste afin d'obtenir l'effet recherché sur les perceptions. Il s'agira ensuite d'*« imposer son récit »* (cf. Figure 24). Cette connaissance générale des tendances, des relais numériques, et autres avatars favorables ou défavorables à la force, pourront être utiles pour mettre en exergue les faits et informations que l'on souhaite voir imposer tout en masquant certaines, en les « diluant » dans le flot continu d'informations. D'où ce besoin d'une empreinte déjà existante et favorable pour imposer notre récit et pas celui que nos détracteurs et autres opposants souhaiteraient mettre en lumière : *« Afin d'éviter d'être inefficace, voire contre-productive, l'action d'influence sur la population devra procéder d'une connaissance à la fois fine et exploitable du tissu humain et de l'environnement culturel. Elle visera notamment à analyser et préparer les info-cibles. Cette connaissance s'exprimera au travers de quatre capacités :*


Imposer « son récit »

Dans le futur, profitant de la connexion croissante des sociétés, le foisonnement informationnel aura encore gagné en intensité. Ce champ immatériel largement dérégulé sera plus que jamais propice à l'affrontement de récits concurrents. Pratiquée par tous nos adversaires, cette lutte poursuivra des objectifs différents selon les publics ciblés (émousser l'esprit de résistance des Français et leur soutien à l'action militaire, délégitimer l'action de la Force et affaiblir son acceptation par la population locale, effriter la cohésion entre la Force et ses partenaires de combat,...).

Par leur action au cœur de cette réalité informationnelle, d'essence humaine, les forces terrestres seront aux avant-postes d'un combat, dont il est clair toutefois qu'il ne se pratiquera que dans un cadre plus vaste, interarmées et interministériel, et contre un adversaire étranger. Les développements de la cyberconflictualité s'accompagneront d'évolutions capacitaires qui permettront la neutralisation des messages de l'adversaire, y compris au cœur des réseaux sociaux, et le contraindront dans l'emploi de ses moyens. À l'échelle stratégique, la « bataille narrative » consistera pour l'armée de Terre, en tant que corps social, à incarner un certain nombre des vertus et des valeurs dont la Nation continuera d'avoir besoin pour sa cohésion et sa résilience.

Figure 24: Imposer son récit (Source : ATF 2016)

- *une capacité de discrimination des groupes humains afin d'éviter les amalgames, les erreurs d'appréciation, un message inadapté, ou une mauvaise utilisation des symboles et coutumes ;*
- *une capacité d'identification des info-cibles, correspondant à l'effet d'influence recherché ;*

- *une capacité de veille qui vise à évaluer l'opportunité d'une action d'influence sur ces info-cibles ;*
- *une capacité de coordination - déconfliction pour éviter les doublons voire des actions d'influence contradictoires.*

➢ L'influence comme facteur de supériorité opérationnel des forces terrestres grâce au *data mining*

Cette influence doit être mise en place à tous les niveaux du tactique au niveau stratégique, et partout où les forces ont des intérêts que ce soit en opérations extérieures comme sur le territoire nationale. Dans ces cas-là, l'influence n'est pas mise en œuvre de la même manière mais la finalité reste la même : *« Emporter l'adhésion des populations de la zone d'opérations sera certes primordial mais surtout assurera la légitimation de l'opération bien au-delà du théâtre »*. Il s'agit bien d'asseoir la légitimité de l'action de la Force militaire.

Ainsi au niveau tactique et opératif « *l'objectif premier et limité de ces fonctions est de protéger l'image de la force, soutenir les populations locales et préparer le retour à une vie normale le plus rapidement possible »*. Pour mener alors des actions d'influence efficace c'est-à-dire qui permettent de prendre la supériorité opérationnelle sur l'adversaire, il faut réussir *« la combinaison de capacités mises en œuvre par les forces terrestres selon la segmentation (artificielle) suivante : planifier et coordonner l'influence ; connaître le tissu humain de la zone d'opération ; agir sur la population ; tromper l'adversaire ; expliquer l'action de la force et enfin évaluer les actions d'influence […]. Il s'agira de ne plus opposer le champ de la coercition et celui de l'influence en considérant, à tort, qu'il faille agir dans l'un ou l'autre selon les circonstances, mais bien de synchroniser les actions dans un plan et une conduite communs afin qu'elles s'appuient mutuellement. »* (Etat Major de l'Arme de Terre, 2016). Le *data mining* contribuera directement à appuyer cette combinaison des effets en fournissant au commandement, cette capacité de connaissance des info-cibles, tant dans leurs environnements d'évolution, que dans leurs spécificités culturelles. Cette connaissance du théâtre d'opérations permettra d'inscrire l'action dans la durée pour « *obtenir l'adhésion de la population (au mieux) ou de neutraliser son opposition (au moins). Ces actions nécessiteront pour les forces terrestres la mise en œuvre de capacités de contrôle (surveiller, séparer, enquêter, discriminer et protéger), de capacités d'assistance (secourir, reconstruire,*

rétablir les flux de première nécessité, évaluer les risques) et des capacités d'appui aux structures régaliennes (suppléer provisoirement l'administration, reconstruire, contrôler l'action d'acteurs non militaires, former). L'adversaire, groupe humain à part entière, n'échappera pas aux effets des actions dans le champ des perceptions. Il conviendra alors d'être capable de mener des opérations militaires de déception (désinformation, rumeurs, manipulations), y compris dans le cyberespace. Dans ce champ de confrontation particulier, des capacités de détection des actions adverses devront permettre d'exercer une forme de téléneutralisation (brouillage du système de commandement adverse, déni de service, lutte informatique offensive). Sur ces dernières mesures, le *data mining* apportera une réelle plus-value en matière de renseignement et de cybersécurité, mais aussi d'évaluation de l'efficacité des actions menées (cf 3.1) : *« Il conviendra dès lors d'être capable en permanence d'évaluer l'image de la force (présence sur les réseaux sociaux), d'exercer une forme de contre-propagande utilisant ces mêmes vecteurs et de disposer en propre d'une capacité physique de communication opérationnelle (station de radio, chaîne de télévision, structure d'accueil des médias, imprimeries, relais d'opinion). Enfin, intervenant en aval, une capacité d'évaluation des actions d'influence aura pour fonction de mesurer, en toute indépendance, leurs effets sur les populations, sur les médias, sur les leaders et sur les forces hostiles, notamment grâce à des outils classiques de mesure de l'opinion (sondages, enquêtes, veille média...). Il s'agira de mettre en place une forme de cycle Observation, Orientation, Décision, Action de l'influence»*
(Etat Major de l'Arme de Terre, 2016).

- Le *data mining* comme outil de compréhension et de veille

Surtout le *data mining, « prendra toute sa part au traitement analytique d'une information qu'elle rendra intelligible malgré son volume : moyens d'acquisition divers et complémentaires, réseaux, analyse systémique, Big Data (avec une réflexion particulière sur la subsidiarité des traitements et la déconcentration des outils « d'intelligence déportée »), traitement automatisé et fusion avec une capacité de réponse à un besoin en temps réel, ingénierie des connaissances, capacité à détecter des signaux faibles (veille et surveillance de l'environnement). Enfin, l'articulation opérationnelle sera systématiquement envisagée dans l'optique d'un meilleur partage de la connaissance par la mise en réseau des unités. L'évolution institutionnelle poursuivra le même objectif de fluidité informationnelle grâce à*

la connexion des individus et des organisations. Dans les deux cas, la notion de niveau hiérarchique conservera sa pertinence pour réguler intelligemment l'information, éviter la saturation, absorber la diversité et donner du sens ». (Etat Major de l'Arme de Terre, 2016)
En détectant les signaux faibles, le *data mining* permet d'anticiper au plus tôt tous les problématiques et crises en fournissant en amont au commandement les éléments et le temps nécessaire pour préparer et adapter au mieux la riposte que ce soit en termes d'actions à mener mais aussi de communication avec un narratif ajusté. Cette avance permettra à la force de maintenir sa crédibilité et donc son influence sur le long terme.

En conclusion, ce chapitre a permis d'entrevoir les possibilités qu'offrait le *data mining* pour les Armées. Ces perspectives d'usage du *data mining* sont projetables sur deux horizons :

- l'un à court terme qui viserait simplement à décliner les pratiques courantes d'usage du *data mining* dans le monde civil dans de nombreux domaines en les adaptant aux structures militaires (RH, soutien, …),
- l'autre plus complexe qui consisterait à utiliser le *data mining* pour accroître l'efficacité de certaines fonctions sensibles et régaliennes telles le renseignement, la cybersécurité et l'influence.

Parfois ces fonctions ont leur pendant dans le secteur privé mais à une échelle de développement moindre en raison de la nécessité de résilience attendu des Armées et des moyens alloués en conséquence. Tout au long de cette analyse il est apparu intéressant de confronter cette vision prospective faite par les forces terrestres dans Action terrestre future, afin de valider certains concepts et pistes. Il ressort de cette vision un net intérêt pour le fait numérique et technologique qui devra occuper notre future, preuve que le *data mining* a bien toute sa part dans les missions futures au moins de l'Armée de Terre.

Conclusions

Ce mémoire a permis au travers de ces trois chapitres d'expliquer d'abord l'origine, les risques et les conséquences liées à cette explosion des données puis le fonctionnement et l'intérêt du *data mining* pour enrichir les données. Enfin le troisième chapitre permet de se projeter à plus ou moins long terme et d'entrevoir l'usage du *data mining* au service de la Défense.

Pour débuter, le premier chapitre a démontré que l'explosion des *data* est réellement en train de modifier notre rapport dans de nombreux domaines structurants nos sociétés. De la Science en passant par l'économie, notre société se modifie plus rapidement que durant n'importe quelle période de l'histoire. Ces changements ne sont pas sans risque et nécessitent donc de traiter cette énorme quantité de données avec efficience d'où le recours nécessaire au *data mining*.

Ensuite le second chapitre a permis de comprendre le processus technique d'enrichissement des données nécessaire pour en tirer du sens. Pour cela, le *data mining* peut s'appuyer sur deux méthodes différentes et complémentaires :

- l'une utilisant la statistique descriptive visant à détecter et décrire des tendances et autres liens de causalité,
- l'autre prédictive permettant de répondre à des problèmes,

Enfin le troisième chapitre a exposé les applications et autres usages possibles du *data mining* dans l'univers de la Défense. Cet apport est en marche et va s'inscrire dans la durée tant les bénéfices entrevus paraissent incommensurables. Ils concernent aussi bien des domaines classiques des services en cours de transformation digitale, que les domaines plus régaliens et sensibles comme le renseignement, la cybersécurité ou encore l'influence. Le recours au *data mining* va donc permettre de gagner en efficacité dans les secteurs traditionnels en suivant en parallèle l'évolution naturelle de nos sociétés, toujours davantage numérisées. Dans les secteurs régaliens, l'intégration du *data ming* va se poursuivre pour permettre de profiter pleinement de tout le potentiel offert par cette profusion de données circulant dans le cyberespace au travers d'une nouvelle forme de renseignement : le renseignement cyber (incluant à la fois le renseignement d'intérêt cyber que le renseignement d'origine cyber), qui viendra en complément des trois autres formes déjà existantes. Cette connaissance du

cyberespace incluant les réseaux sociaux et média sociaux impose également une stratégie d'influence globale, sur tout le spectre afin de garantir une capacité de réaction suffisante pour défendre les valeurs et intérêts de l'institution dans un monde hyper connecté et sans secret. Pour cela, la « grande muette » devra consolider son empreinte numérique de manière visible ou non sur la Toile en s'appuyant sur le *data mining* pour connaître son environnement, ses infos cibles, les signaux faibles, les narratifs et imposer son récit en toutes circonstances.

Cette optimisation amène de fait d'autres problématiques inhérentes à ce milieu et développées dans ce mémoire comme notre périmètre d'action difficile à délimiter en l'absence de frontières clairement matérialisées. Ce cinquième milieu est également un nouveau terrain de jeu pour affirmer sa puissance et contester celles de ces adversaires. L'approche juridique américaine du cyberespace basée sur le principe d'extraterritorialité va dans ce sens et traduit ainsi tous leurs intérêts à récupérer un maximum de données à travers le monde pour satisfaire leurs appétits économique, politique et sécuritaire. Il permet d'asseoir leur prééminence dans ce secteur après avoir perdu leur prééminence industrielle à la fin du XXe siècle au détriment du Japon puis de la Chine. A contrario, dans un mouvement de balancier, on observe les pays européens renforcer la protection des données à caractère personnel pour espérer préserver un minima de confidentialité, de vie privée.

D'autres enjeux davantage liés sur l'aspect technique du *data mining,* vont apparaître progressivement dans le temps : avec le besoin quasi immédiat d'infrastructures numériques capables d'accueillir cette croissance exponentielle de *data*. A moyen terme apparaît le besoin d'algorithmes d'apprentissage automatique (*machine learning*) toujours plus performants capables de traiter n'importe quel type de données pour la rendre exploitable et corrélable avec le reste des *data*. Cela passera sans doute par une standardisation des formats ou une optimisation des *data*. Enfin les enjeux liés à l'évolution du *data mining* nécessiteront de prendre en compte à plus long terme l'intelligence artificielle afin de faciliter l'émergence de nouveau patterns toujours plus intuitifs tout en veillant à respecter certaines règles évoquées précédemment comme l'éthique. La présence de « l'Homme prudent » au cœur du *data mining* doit rester le principal enjeu. En effet cette maîtrise des risques reste indispensable pour éviter toute dérive transhumaniste.

Bibliographie :

Amaël Cataruzza, D. D. (2014). *EPS : La balkanisation du Web : chance ou risque pour l'Europe ?* Délégation aux affaires stratégiques.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*.

Ashok, C., D, H. N., & Rohini, B. (2011, 08). Data mining techniques for intrusion detection and prevention system. *International Journal of Computer Science and Network Security*, pp. 200-204.

Barker, C. (2015, 07 29). *Le dilemme de l'éléphant : à quoi ressemble l'avenir des bases de données ?* Récupéré sur zdnet.com: http://www.zdnet.fr/actualites/le-dilemme-de-l-elephant-a-quoi-ressemble-l-avenir-des-bases-de-donnees-39823060.htm

Baudot, P.-Y. (2015, 5). Encore une révolution informatique? Open et big data dans les organisations administratives. *Informations sociales*, pp. 8-18.

Boyer, B. (2014). *Cybertactique - Conduire la guerre numérique.* Nuvis.

Bruno Agard, A. K. (2005). *Exploration des bases de données industrielles à l'aide du datamining – Perspectives.*

Etat Major de l'Arme de Terre. (2016). *Action terrestre future, demain se gagne aujourd'hui.*

Guyader, P. L. (2013). *Protection des données sur Internet.* Hermes-Science.

Henrotin, J. (2008). *La technologie militaire en question - le cas américain.* Economica.

Ifrah, L. (2010). *Que sais-je ? L'information et le renseignement par internet.* PUF.

Ismael, B. (2014, 06 18). Etude critique d'un système d'analyse prédictive appliqué à la criminalité : Predpol.

Lee, W. (1999). *A data mining framework for constructing features and models for intrusion detection systems (computer security, network security).*

Lejeune, Y. (2014). *Big fast open DATA - Décrire, décrypter et prédire le monde : l'avènement des données.* FYP Editions.

NCCIC. (2016). *GRIZZLY STEPPE - Russian Malicious Cyber Activity - Joint Analysis Report.* Federal Bureau of Investigation.

Paquienséguy, F. (2016). *Open data - Accès, territoires, citoyenneté : des problématiques info-communicationnelles.* Editions des archives contemporaines.

Pierre-Jean Benghozi, S. B.-F. (s.d.). *L'Internet des objets .* MSH.

Pigliucci, M. (2009). *The end of theory in science?* PMC.

Richardson, D. J. (2010). Filling the light pipe. *Science*, pp. 327-328.

Roché, E. (2016). Open data et business models. *LEGICOM*, 121-127.

Russo, J.-C. P. (2016, 6). D'abord les données, ensuite la méthode ? Big data et déterminisme en sciences sociales. *Socio.*, pp. p. 97-115.

S. S. Anand, A. G. (1998). Decision Support Using Data Mining. *Financial Times* .

The MITRE Corporation. (s.d.). *Data Mining for Network Intrusion Detection: How to Get Started.*

Articles:

Wastell, C., Clark, G., & Duncan, P. (2006). Effective intelligence analysis: The human dimension. Journal of Policing, Intelligence and Counter Terrorism, 1(1), 36-52.

Lim, K. (2016). Big Data and Strategic Intelligence. Intelligence and National Security, 31(4), 619-635.

Schoech, D., Fitch, D., Macfadden, R., & Schkade, L. L. (2002). From data to intelligence: Introducing the intelligent organization. Administration in Social Work, 26(1), 1-21.

Ball, M. G., Qela, B., & Wesolkowski, S. (2016). A Review of the Use of Computational Intelligence in the Design of Military Surveillance Networks. In Recent Advances in Computational Intelligence in Defense and Security (pp. 663-693). Springer International Publishing.

Bauman, Z., Bigo, D., Esteves, P., Guild, E., Jabri, V., Lyon, D., & Walker, R. R. (2015). Repenser l'impact de la surveillance après l'affaire Snowden: sécurité nationale, droits de l'homme, démocratie, subjectivité et obéissance. Cultures & Conflits, (2), 133-166.

Document numérique et société, May 2015, Rabat, Maroc. De Boeck, Open Data, big data: quelles valeurs ? Quels enjeux ? Collection "Information & Stratégie", pp.67-83, 2015, Information & Stratégie.

MARTIN-JUCHAT, M. F., MIÈGE, M. B., BOULLIER, M. D., & ROCHELANDET, M. F. Le cadre privatif: des données aux contextes.

Türk, A., & Piazza, P. (2009). La difficile quête d'un équilibre entre impératifs de sécurité publique et protection de la vie privée. Entretien avec Alex TÜRK; propos recueillis par Pierre PIAZZA. Cultures & Conflits, (76), 115-134.

Chartron, G., & Broudoux, E. (2015, May). Enjeux géopolitiques des données, asymétries déterminantes. In Document numérique et société (pp. 67-83). De Boeck.

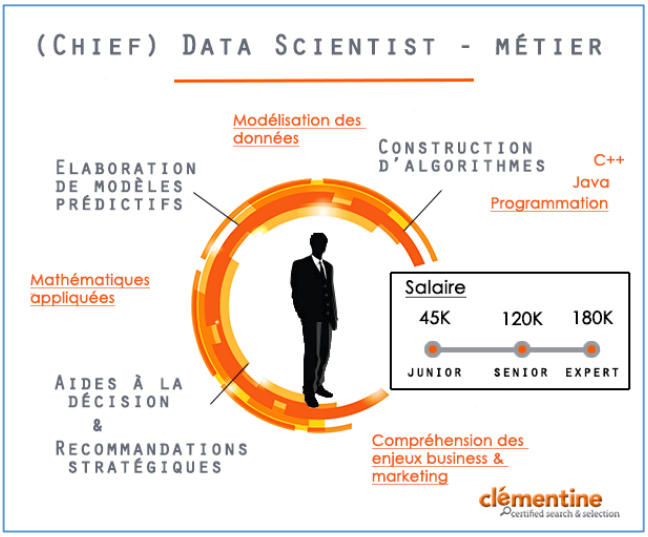
Annexes :

Annexe 1 : Fiche métier - data scientist/chef data

Le data scientist est responsable de la gestion et de l'analyse des données (big data). Il est en charge de la récupération et du traitement de millions de données. Au-delà de la fonction du « data analyst », il a le rôle de faire parler les données et de les mettre au service de la direction d'une entreprise. Ce métier relève d'enjeux à la fois fonctionnels et stratégiques pour l'entreprise.

Il s'agit d'un métier récent, s'exerçant dans des secteurs très variés, la finance, l'informatique, l'assurance, l'e-commerce, la grande distribution… impliquant des compétences techniques diverses selon l'environnement de travail. Les data scientists sont des profils rares et extrêmement recherchés. Certaines entreprises préfèrent combiner différents profils pour créer une équipe big data, en raison de la complexité technique des taches demandées pour ce poste.

Le métier de data scientist apparaît dans un environnement de concurrence accrue des marchés, demandant aux décideurs d'appuyer leur prise de décision sur des statistiques fiables.



Les champs d'analyse des données en entreprise sont très larges : données relatives à la production, l'analyse du marché et à ses prédictions, la connaissance le comportement client, la précision markéting, les résultats de l'entreprise… Le data scientist joue un grand rôle dans la création d'indicateurs précieux à tous les niveaux de l'entreprise. Il s'investit alors dans l'amélioration de l'activité globale grâce à la précision de l'analyse et en mettant sur pied des modèles de prédiction.

Certains considèrent que le métier de data scientist est l'évolution du poste de data miner.

Ses missions principales :

- Etude des données en possession de l'entreprise qui permettront de définir les données qui seront extraites et traitées, en accord avec les exigences de la direction.
- Récupération et analyse des données pertinentes liées au processus de production de l'entreprise, à la vente ou encore liées aux données client.

- Construction d'algorithmes permettant d'améliorer les résultats de recherches et de ciblage.
- Elaboration de modèles prédictifs afin d'anticiper l'évolution des données et tendances relatives à l'activité de l'entreprise.
- Modélisation des résultats d'analyse des données afin de les rendre lisibles et exploitables par les managers.
- Recommandations business auprès de la direction générale afin d'améliorer la prise de décision. Ce travail d'interprétation des données pourra également se faire au travers de la création d'un tableau de bord spécifique / logiciel sur mesure analysant les données traitées. La création de métriques d'aide à la décision pourra avoir une influence conséquente sur la stratégie de l'entreprise.

Ses missions connexes :

- Définition de solutions de stockage des données, en lien avec la direction des systèmes d'information
- Participation au recrutement d'experts Big Data pour compléter l'équipe qui travaille sur le traitement des données
- Recherche & développement relatif au traitement de grands volumes de données

Le chief data scientist manage une équipe de data scientists et est responsable des projets d'analyse des données et de la mise en place des outils de mesure. Il intervient directement dans la stratégie de l'entreprise et prend part au processus de prise de décision.

En tant que chief data scientist, vous managez une équipe de data scientists et vous êtes responsable des projets d'analyse des données et de la mise en place des outils de mesure. Vous intervenez directement dans la stratégie de l'entreprise et prenez part au processus de prise de décision.

Son profil :

De formation supérieure en école d'ingénieur, école d'informatique ou de statistique, il justifie également de 4/5 ans d'expérience à un poste d'analyse de données (data analyste ou similaire) ou dans un environnement datacenter (au contact de grandes structures de données). Il existe actuellement très peu de formations spécialisées sur ce métier.

Ses compétences :

- Solides compétences en programmation informatique (Java, C++…) et une bonne compréhension des structures de données.

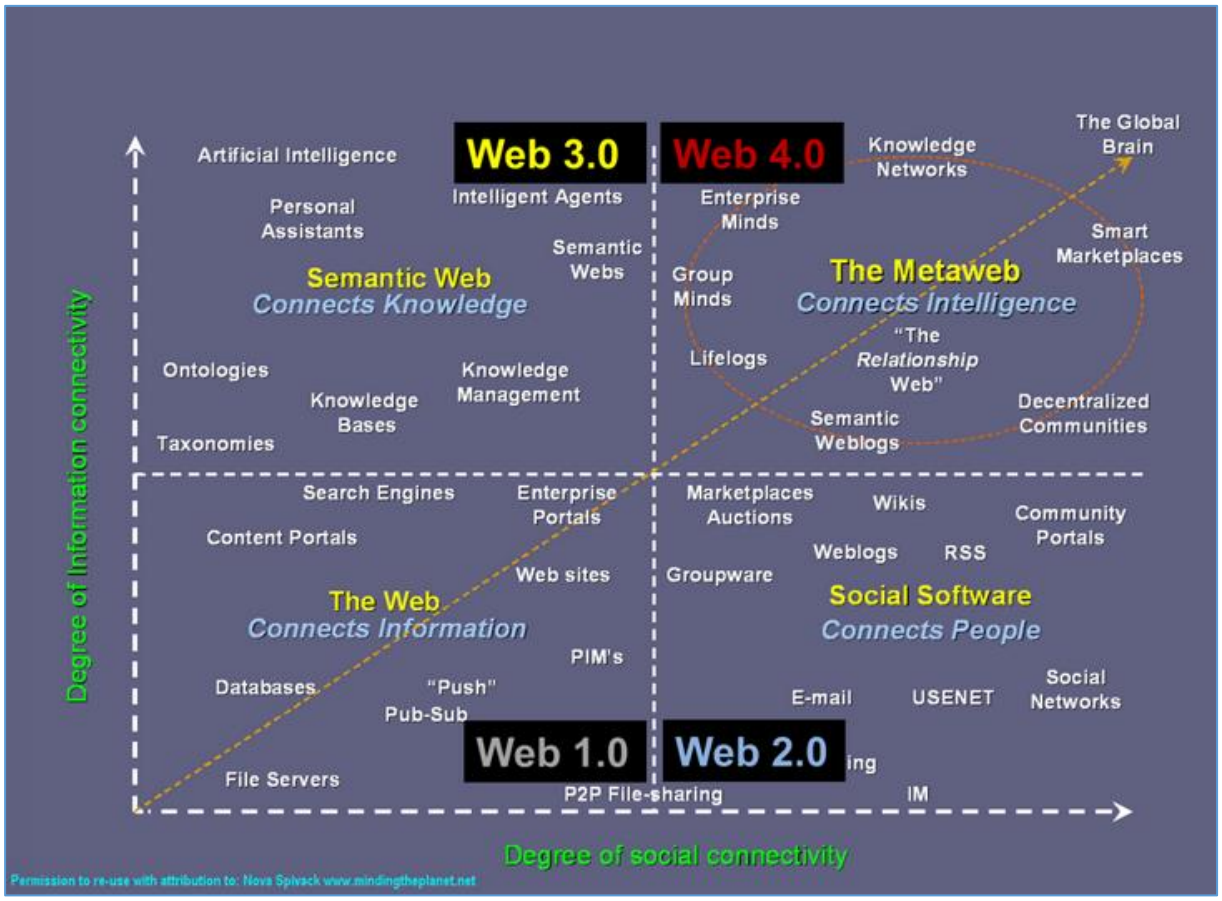
- Expertise en algorythmie et gestion des bases de données (NoSQL, Cassandra…)
- Maîtrise de l'architecture des bases de données décisionnelles (data warehouse)
- Mathématiques appliquées : Construire des algorithmes pour améliorer les résultats de recherches et de ciblage
- Statistiques : capacité à réaliser des analyses prédictives et statistiques à partir des différentes bases de données
- Des compétences en gestion de projet seront également appréciées.

Son salaire :

- De 45K euros pour un profil junior, jusqu'à 120K € /an
- Le chief data scientist (manageant une équipe de data scientist) peut gagner jusqu'à 180K € /an

La rémunération du data scientist dépend en grande partie de son expérience et de sa capacité à élaborer des algorithmes puissants et efficaces. Son expertise est alors reconnue lorsqu'il parvient à produire des modèles de données précis et utiles à la prise de décision.

Annexe 2 : Les différents Web : du 1.0 au 4.0



Annexe 3 : Quelques notions sur les octets

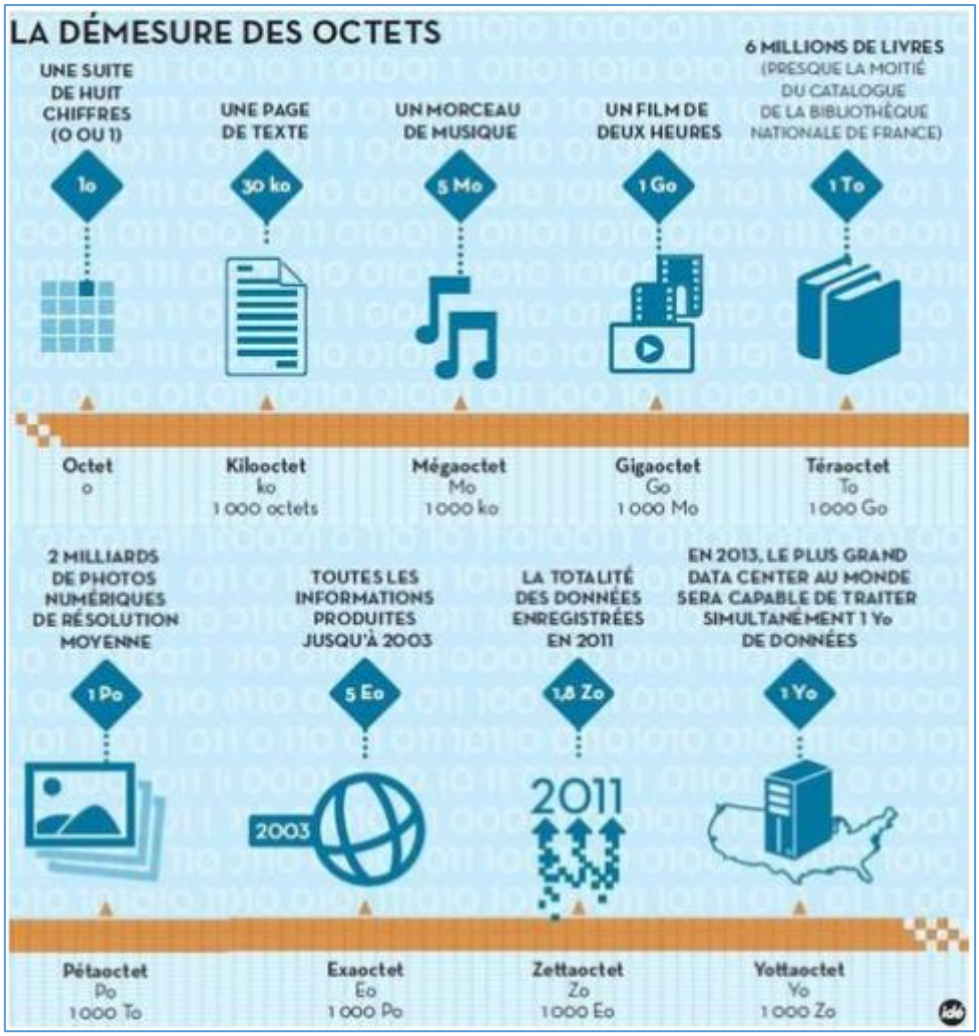


Figure 25: La démesure des octets (Source Idé)

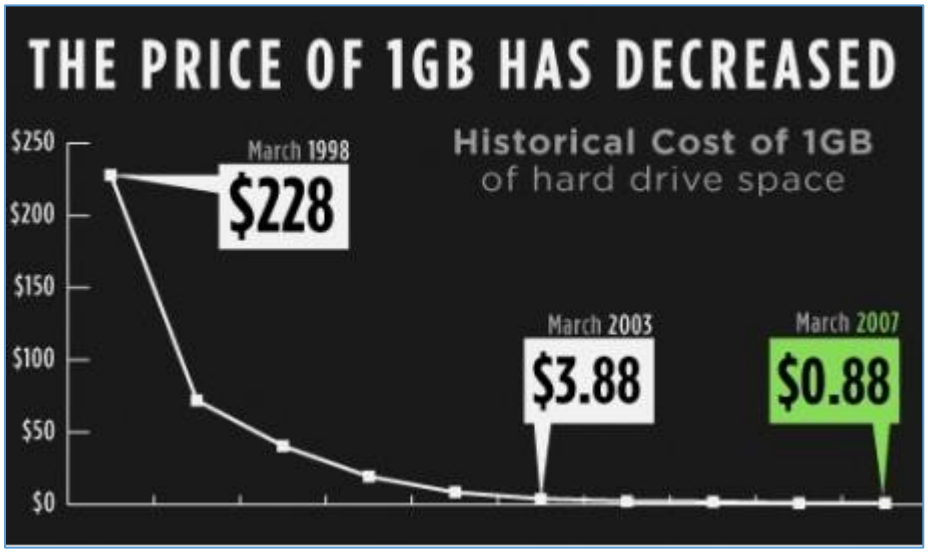


Figure 26: La baisse du coût du stockage (Source : Mozy)