DATA SCIENCE AU SERVICE DE L'ANALYSE COMPORTEMENTALE ET PREDICTIVE

Application à la criminologie

MEMOIRE DE FIN D'ETUDE 2016 - 2018

©DRAGAN JOVANDARIC
Expert en Ingénierie Informatique

Remerciements

Je souhaite avant tout remercier ma femme, Sophie, de m'avoir soutenu dans ce projet et d'avoir confiance en moi. Ensuite, je souhaite remercier mes enfants, Matthieu et Lucas de m'avoir transmis leur enthousiasme dans cette poursuite d'études.

En particulier, je tiens à exprimer tout ma reconnaissance à ma directrice de mémoire, Emna BAHRI, de m'avoir encadré, orienté et conseillé.

J'adresse mes sincères remerciements aux professeurs qui m'ont inspiré pour choisir ce thème, en particulier Noël BARON, qui a su transmettre sa passion d'analyse et profilage.

Mes remerciements sont également adressés à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé mes réflexions et ont accepté de me rencontrer et de répondre à mes questions durant mes recherches.

L'enseignement de qualité dispensé par le Master de Business Intelligence et Big Data a également su nourrir mes réflexions et a représenté une satisfaction intellectuelle.

En fin, mes remerciements s'adressent à Business et Décision, l'entreprise qui m'a accueilli et soutenu pendant cette période.

Data science au service de l'analyse comportementale et prédictive : Application en criminologie

Table des matières

Remerciements ..... - 1 -

Data science au service de l'analyse comportementale et prédictive : Application en criminologie ..... - 2 -

I. ∞ INTRODUCTION ∞ ..... - 6 -

Mots clé ..... - 7 -

I. Introduction ..... - 8 -

Keywords ..... - 9 -

I.1. Où en sommes-nous aujourd'hui ? ..... - 10 -

I.2. Introduction à l'analyse criminelle ..... - 11 -

I.2.1. Le comportement ..... - 11 -

I.2.2. L'attitude ..... - 12 -

I.2.3. La relation entre attitude et comportement ..... - 13 -

I.3. Influence et changement de comportement (persuasion) ..... - 14 -

I.3.1. Comment persuader autrui ? ..... - 14 -

I.3.2. Ecole de Yale ..... - 14 -

I.3.3. 5 W ..... - 15 -

I.3.4. La persuasion : approche cognitive : Modèle de Vraisemblance d'Élaboration ..... - 17 -

I.3.5. Nudge ..... - 18 -

I.4. Détection du comportement ..... - 19 -

I.4.1. Modèles normaux de développement ..... - 19 -

I.4.2. Détection de changements ..... - 20 -

I.4.3. Théories du changement comportemental ..... - 20 -

I.5. Rétrospective de l'analyse criminelle ..... - 22 -

I.6. RGPD et des nouvelles technologies – problème de profilage ..... - 24 -

II. ∞ ETAT DE L'ART ∞ ..... - 29 -

II.1. BIG DATA ..... - 30 -

II.1.1. Le marché des Big Data ..... - 32 -

II.1.2. De nouveaux outils pour la prise de décisions ..... - 33 -

II.1.3. Quels résultats espérer des Big Data ..... - 34 -

II.1.4. Le volume ..... - 35 -

II.1.5. Variété ..... - 35 -

II.1.6. Vitesse ..... - 35 -

II.1.7. Véracité ..... - 36 -

II.1.8. Valeur ..... - 36 -

II.1.9. Le défi de la définition de 5V de Big Data ..... - 37 -
II.2. Analyse des données ..... - 38 -
II.2.1. L'analyse par grappes ..... - 38 -
II.2.2. Analyse de régression ..... - 38 -
II.2.3. Test A/ B ..... - 38 -
II.2.4. Simulation Monte Carlo ..... - 38 -
II.2.5. Traitement de graphique ..... - 39 -
II.2.6. Machine Learning ..... - 39 -
II.2.7. Informatique cognitive ..... - 39 -
II.2.8. Deep Learning ..... - 40 -
II.3. Gestion de données ..... - 41 -
II.3.1. Modèle traditionnel ..... - 41 -
II.3.2. Modèles hybrides ..... - 41 -
II.3.2.1. Bit Torrent ..... - 42 -
II.3.2.2. Bitcoin ..... - 42 -
II.4. Maintien de l'ordre, Big Data et la marchandisation de la sécurité ..... - 43 -
II.4.1. Maintien de l'ordre et de la sécurité à la fin du 20ème siècle ..... - 43 -
II.4.2. L'impact de la technologie ..... - 44 -
II.4.3. Sources d'information ..... - 45 -
II.5. La police prédictive ..... - 46 -
II.5.1. Types de police prédictive ..... - 46 -
II.5.2. Efficacité de la police prédictive ..... - 48 -
II.6. Outils d'analyse ..... - 49 -
II.6.1. Anacrim ..... - 49 -
II.6.2. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) ..... - 50 -
II.6.3. Faception ..... - 53 -
II.6.4. BEWARE ..... - 54 -
II.6.5. Deep Science : Coban ..... - 55 -
II.6.6. Predpol ..... - 56 -
II.7. Prédiction et discrimination ..... - 58 -
II.7.1. Préjugés, la discrimination algorithmique et la stigmatisation ..... - 58 -
I.1.1. Faux positifs/négatifs ..... - 59 -
III. ∞ PROPOSITION ∞ ..... - 60 -
III.1. LE PROFILAGE AVANT LE RGPD : ENJEU DU BIG DATA ..... - 60 -
III.1.1. Obstacle RGPD (extrait d'article 4) ..... - 60 -
III.1.2. G29 ..... - 61 -
III.1.3. Droit d'opposition au profilage basé sur une prise de décision automatisée ..... - 62 -

III.1.3.1. Article 22 du RGPD ..... - 62 -
III.1.4. Les exceptions au droit d'opposition de la personne concernée au profilage automatisé – le temps perdu ..... - 62 -
III.1.4.1. Extrait d'article 22 a) du RGPD ..... - 62 -
III.1.5. Un droit national contraire en vigueur : ..... - 63 -
III.1.5.1. Extrait d'article 22 b) du RGPD ..... - 63 -
III.1.6. Le consentement explicite de la personne concernée ..... - 63 -
III.1.6.1. Extrait d'article 22 c) du RGPD ..... - 63 -
III.2. LE PROBLEME DES SOURCES DES DONNEES ..... - 64 -
III.2.1. Les " logs " (journaux de connexion) issus du trafic sur le site officiel de l'entreprise ..... - 65 -
III.2.2. Le contenu et les mesures de réputation issus des médias sociaux ..... - 65 -
III.2.3. La " third party" data ..... - 65 -
III.2.4. L'open data ..... - 65 -
III.2.5. Les données ouvertes en France : Open data Gouv ..... - 66 -
III.2.6. Grey data ..... - 66 -
III.3. PROPOSITION DE SOLUTION D'ANALYSE DES DONNEES ..... - 67 -
III.3.1. Répondre aux RGPD et profilage ..... - 67 -
III.3.2. Anonymisation ..... - 68 -
III.3.2.1. Les techniques d'anonymisation ..... - 69 -
III.3.2.2. Les techniques de pseudonymisation ..... - 69 -
III.3.2.3. Système cryptographique à clé secrète : ..... - 70 -
III.3.2.3.1. Fonction de hachage : ..... - 70 -
III.3.2.3.2. Fonction de hachage par clé, avec clé enregistrée : ..... - 70 -
III.3.2.3.3. Chiffrement déterministe ..... - 70 -
III.3.2.3.4. Chiffrement a trois niveaux ..... - 71 -
III.4. PROPOSITION DE REDUCTION DU TEMPS DE TRAITEMENT ..... - 72 -
III.4.1. Métadonnées ..... - 72 -
III.4.1.1. Où sont stockées les métadonnées ? ..... - 72 -
III.4.1.2. Les métadonnées EXIF : ..... - 73 -
III.4.1.3. Les métadonnées IPTC : ..... - 73 -
III.5. PROPOSITION DE METHODE D'ANALYSE ..... - 76 -
III.5.1. Pattern et profil ..... - 76 -
III.5.2. Méthode inverse : Surveillance de changement de pattern et des relations ..... - 77 -
III.5.2.1. Pattern ..... - 77 -
III.5.2.1.1. La dimension temporelle ..... - 78 -
III.5.2.1.2. La dimension spatiale ..... - 79 -
III.5.2.1.3. La dimension relationnelle ..... - 79 -

III.5.2.1.4. La dimension quantitative ..... - 79 -
III.5.2.2. La dimension relationnelle ..... - 79 -
III.5.3. Comment peut-on accélérer et améliorer le processus ? ..... - 81 -
III.5.3.1. Relations et pattern ..... - 81 -
III.5.4. Visualiser des données dans l'analyse criminelle ..... - 84 -
III.6. MISE EN PRATIQUE DES SOLUTIONS PROPOSEES ..... - 85 -
IV. CONCLUSION ..... - 88 -
Bibliographie ..... - 91 -

I. ∞ INTRODUCTION ∞

"Nul ne peut agir avec l'intensité que suppose l'action criminelle sans laisser des marques multiples de son passage. Par une action inverse, il a emporté sur son corps ou sur ses vêtements les indices de son séjour ou de son geste." E. LOCARD

Séduit par les histoires de Artur Conan Doyle et son personnage, Sherlock Holmes, je me suis intéressé par la méthode de recherche décrite dans son œuvre. Ces méthodes de déduction ont été reprises ou interprétées dans la cinématographie par de nombreuses œuvres qui décrivent le métier de profiler. Comme un petit garçon qui voulait devenir tout ce qu'il voit, j'ai souvent endossé ce rôle. Installé à Lyon depuis quelques années, j'ai découvert également que j'habite dans la ville de la première personne qui a créé le laboratoire de recherche scientifique en 1910, Edmond Locard. Lui-même, admirateur des histoires de Conan Doyle, il a mis en réalité le rêve de son écrivain préféré en formant ce laboratoire.

Le cas d'Edmond Locard nous montre que si on réfléchit et applique des connaissances de son domaine d'expertise, on peut améliorer la procédure de recherche d'un autre domaine car on apporte un autre point de vue et on voit des détails non pris en compte auparavant.

Une des motivations pour choisir ce thème pour mon mémoire prend racines dans mon enfance et d'autres dans le métier j'ai choisi de faire aujourd'hui, dans le domaine de traitement des données numérique. Je me suis interrogé après des évènements qui ont soulevé la question de la liberté et confidentialité des données, sur la manière de rendre notre quotidien plus sûr et de prime abord je n'ai pas trouvé la réponse. Je me suis dit qu'avec les NTIC (Nouvelles Technologies de l'Information et de la Communication) les outils de recherche ont été considérablement améliorés et qu'il ne reste pas grande chose à inventer, si ce n'est d'avoir une approche différente, un regard où un modèle qui peut nous permettre d'améliorer l'analyse et réduire considérablement le temps de traitement des données.

Dans ce mémoire, j'ai d'abord choisi de faire une rétrospective et de voir quelle était la raison et la nécessité d'établir un modèle de profilage criminel. Je pose des questions auxquelles des réponses existent, mais ces questions méritent d'être posées car elles nous amènent vers l'importance de l'information et sa source.

Ensuite, pour bien comprendre pourquoi on va s'intéresser à certains points qui proviennent de la science, de la nature humaine et de la psychologie, je m'intéresse à expliquer les modèles de comportements

humains bien ancrés en nous. On va parler des comportements, de leurs différents types et leurs définitions.

En parallèle, je vais expliquer la liaison entre le comportement et l'attitude, où on va trouver une forte interaction. Je vais également présenter l'attitude à travers un exemple simple et ensuite je vais faire l'introduction vers le changement du comportement et des différentes écoles de manipulation…

Pourquoi la manipulation ? Tout simplement car on se fait manipuler ou on manipule des personnes tout le long de notre vie. Pour comprendre comment on se fait manipuler, il faut d'abord voir de quels moyens on dispose et quelles techniques on utilise pour faire une modification de comportement.

Je vais également aborder le sujet de Nudge, une technique utilisée dans nos entreprises aujourd'hui, celle qui nous permet de modifier notre comportement.

Ensuite, je souhaite voir, une fois appris qu'est ce qui est le comportement et comment on peut le manipuler, comment peut-on détecter le changement du comportement et quelles sont des techniques pour le détecter. Egalement, je trouve intéressant de présenter les étapes du changement du comportement car une fois bien définies, on peut plus simplement détecter dans quel stade de changement on se trouve et cette approche peut nous orienter vers des actions à mener.

Tous ces concepts mis ensemble nous amènent vers la solution logique, la naissance de l'analyse criminelle, d'abord aux Etats-Unis, puis en Europe. Je souhaite montrer la vision et corrélation entre les premières data center utilisées pour centraliser des données et la naissance du Big Data.

Au final, à la fin de ce chapitre, je souhaite parler de RGBD et de l'impact qu'il va créer sur l'utilisation des données. Non seulement l'impact est énorme du point de vue de la collecte et de la vie privé, mais également sur le métier de profilage et la sécurité intérieure.

Commençons alors par une réflexion sur les données et leurs sources ainsi que quelques questions que je me suis posées avant de commencer la rédaction de ce mémoir.

Mots clé

PROFILAGE, COMPORTEMENT, ATTITUDE, INFLUENCE, ANALYSE, SOURCES D'INFORMATION, PREDICTION, DONNEES OUVERT, ANONYMISATION, METADONNEES

I. Introduction

"No one can act with the intensity that criminal action implies without leaving multiple marks of his passage, and by a reverse action he has taken away the signs of his stay or gesture on his body or clothes."
E. LOCARD

Seduced by the stories of Artur Conan Doyle and his character Sherlock Holmes, I became interested in the method of research described in his work. These methods of deduction have been taken and interpreted on big screen by numerous films and series that describe the profiling profession. As a little boy who wanted to become everything he saw, I often took this role. Based in Lyon since a few years, I also discovered that I live in the city where the first scientific research laboratory was created in1910 by E .Locard. Himself, a admirer of Conan Doyle's stories, put the dream of his favorite writer into the real life.

The story of E.Locard shows us that if we reflect and apply knowledge of our area of expertise, we can improve the search procedure of another area because we bring another point of view and we see details not considered until now.

One of the motivations to choose this theme for my final work is rooted in my childhood and others in the job I chose to do today, in the field of digital data processing. I asked myself, after events who raised the question of freedom and data confidentiality, how to make our daily life safer and at first I did not find the answer.

I told myself that with NICT (New Information and Communication Technologies) the research has been considerably improved and that there is not much left to invent, except to have a different approach, a model that can allow us to improve the analysis and significantly reduce the data processing time.

In this report, I firstly chose to do a retrospective and see what was the reason and the need to establish a criminal profiling model. I ask questions to which answers exists, but questions worth asking because they lead us to the importance of information and its source.

Then, to understand why we are going to be interested in some points that come from science of human nature and psychology, I am interested in explaining the models of human behavior well anchored in ourselves. We will talk about behavior and different types of behavior and definitions of behavior.

At the same time, I will explain the connection between behavior and attitude, where we will find a strong interaction. I will also present the attitude through a simple example and then I will make the introduction to behavioral change and different schools of manipulation.

Why manipulation? Simply because we are manipulated or we manipulate people throughout our lives. To understand how we are manipulated, we must first see what means are available and what techniques are used to change a behavior.

I will also speak about the Nudge, a technique used in today's business, one that allows us to change our behavior.

Then, I want to see, once learned what the behavior is and how it can be manipulated, how can we detect the behavior change and what are the techniques to detect it. Also, I find it interesting to present stages of behavior change because once steps are well defined, we can more easily detect in what stage of change we are and this approach can guide us to make the choice of actions to take.

All these concepts put together leads us to the logical solution, the birth of criminal analysis, always in the United States first, then in Europe. Here I want to show the vision and correlation between the first data center used to centralize data and the birth of the BigData.

Finally, at the end of this chapter, I want to talk about RGBD and the impact it will create on the use of data. Not only that the impact is huge from the point of view of collection and private life, but on the profiling business and homeland security too.

I hope that you will enjoy reading this document as much I enjoyed writing it.

Keywords

PROFILING, BEHAVIOR, ATTITUDE, INFLUENCE, ANALYSIS, SOURCES OF INFORMATION, PREDICTION, OPEN DATA, ANONYMIZATION, METADATA

I.1. Où en sommes-nous aujourd'hui ?

À présent nous sommes rentrés dans une époque où l'information a pris une place très importante dans la vie courante. Avec l'arrivée d'internet et l'expansion des réseaux sociaux proposés aux utilisateurs "gratuitement" nous sommes submergés d'informations personnelles et professionnelles (des photos, nos humeurs, opinions politique, nos envies …). Les objets connectés s'ajoutent à cette liste de sources que l'on doit prendre sérieusement en considération (état de santé et déplacement d'une personne etc.). Ces différentes plateformes se révèlent à la fois source d'informations, de manipulations de promotion ou de déstabilisation. Savoir comment trouver la bonne information, l'extraire et la présenter en fonction de nos besoins est un métier qui a commencé à se développer et qui devenu un acteur majeur sur le marché d'informations.

Trouver l'information, ou plutôt s'assurer qu'elle est pertinente, être capable de prédire le comportement d'une personne ou un groupe, connaitre les habitudes et presque savoir lire les pensées des personnes peut décrire en quelques mots les métiers des services de renseignements d'état. De nos jours les agences publicitaires, les sociétés privés et les marchands d'informations ne se privent pas de ces techniques avancées.

Le métier d'analyste comportemental est particulièrement pointu. Il s'agit d'essayer de déterminer en temps réel notre besoin et ce que l'on compte faire. Le moment venu où nous allons passer à l'acte (d'achat, de prendre une décision etc.) des solutions adaptées vont nous être proposé automatiquement afin de nous faciliter la tâche (au profit économique des vendeurs ciblés) pour être en mesure de réagir convenablement.

Les questions suivantes se posent naturellement : Comment peut-on se servir de l'analyse comportementale pour mieux protéger notre société ? Quels outils où méthodes peut-on développer pour y parvenir sans passer outre la loi ? Peut-on se servir de toutes informations recueillies et comment peut-on trier et séparer les bonnes et les mauvaises informations ? Peut-on réduire le temps de traitement des informations et obtenir le résultat pertinent ? comment être sûr que des informations recueillies ne s'opposent pas au GRPD (Règlement Général sur la Protection des Données) ?

Aujourd'hui avec le développement des nouvelles technologies et des métiers de data science il devient possible de répondre à ces questions et d'aller encore plus loin. Pour parvenir à proposer une solution, il est nécessaire de se familiariser avec des méthodes traditionnelles et de comprendre la situation actuelle incluant la dernière "épine", le GRPD.

Depuis plusieurs années, de nouvelles techniques et outils d'aide à l'enquête importés d'outre-Atlantique apparaissent dans le paysage juridique, notamment à l'occasion de crimes commis en série. Ces techniques largement médiatisées entrainent une fascination du public à travers des films ou romans policiers qui mettent en avant des métiers de "profileurs".

I.2. Introduction à l'analyse criminelle

L'analyse criminelle, utilisée depuis plusieurs années par les services policiers se définit comme "la recherche et la mise en évidence méthodique de relations entre des données de criminalité elles-mêmes d'une part, et entre des données de criminalité et d'autres données significatives possibles d'autre part, à des fins de pratiques judiciaires et policières".

Quant à l'analyse comportementale, elle constitue une forme particulière d'analyse criminelle qui nécessite le recours à des connaissances relevant des sciences humaines.

I.2.1. Le comportement

Dans le cadre de l'analyse comportementale et prédictif, la « Data Science » prend une place principale. Mais avant de commencer à analyser des données, nous allons prendre un moment pour déterminer les points, le sujet d'analyse de comportement…

Le comportement est défini comme la manière générale de se comporter dont agir et réagir. Il s'agit de la façon d'agir vis-à-vis des incitations et par rapport au milieu et à l'entourage dans laquelle la personne évolue.

Il existe plusieurs modes de comportement suivant les circonstances en question.

Le comportement conscient dont celui qui a lieu suite à un processus de raisonnement. Dire bonjour à quelqu'un que l'on connait lorsqu'on le rencontre dans la rue en est un exemple.

Le comportement inconscient fait de manière quasi automatique étant donné que la personne n'a même pas besoin de penser ou de réfléchir à l'action (se gratter après avoir été piqué par un moustique, par exemple).

Le comportement privé a lieu dans l'intimité de chez soi ou en toute solitude. Dans ce cas, l'individu ne risque pas d'avoir à être sujet aux regards indiscrets des autres.

Le comportement public est le contraire car il se déroule en la présence d'autres personnes ou dans des endroits partagés avec le reste de la société.

Du point de vue de la psychologie, le comportement est tout ce que fait un être humain face à son milieu. Chaque interaction d'une personne avec son entourage implique un comportement. Lorsque ce dernier présente des standards stables, on peut alors parler de conduite.

Il est possible de parler de *bon comportement* ou de *mauvais comportement*, suivant la manière dont les actions peuvent s'encadrer au sein des règles sociales.

I.2.2. L'attitude

Le concept d'attitude occupe une place centrale en psychologie sociale depuis les années trente et maintient encore aujourd'hui cette position. Dans cette discipline, le concept d'attitude n'est pas similaire à son sens commun, il est ici une évaluation plus ou moins favorable d'un objet donné. Ces attitudes peuvent concerner aussi bien des objets très vastes (l'Église, la pollution, le soleil…) que des objets très précis (la forme d'une bouteille d'eau minérale, l'utilisation de l'huile de cacao dans la confection du chocolat, l'attrait envers une tâche fastidieuse…).

Les attitudes n'étant pas des objets facilement accessibles, elles sont le plus généralement appréhendées de manière déclarative à l'aide d'une échelle de mesure : l'individu donne par écrit son appréciation sur l'objet en se positionnant sur une échelle d'intervalles en plusieurs points allant de "je n'aime pas du tout" à "j'aime tout à fait".

Le premier enjeu des études sur l'attitude a été de *prédire un comportement effectif* à partir d'une simple déclaration. Les études originales cherchant à attester cette relation entre attitude et comportement se sont heurtées à un obstacle : il était difficile de mettre en adéquation un comportement à son attitude adapté.

La recherche faisant référence en la matière a été réalisée par *Lapiere en 1934*. L'objectif de cette étude consistait à s'assurer que les personnes agissaient en accord avec leurs attitudes déclarées. Lapiere a voyagé accompagné d'un couple de Chinois à travers les États-Unis d'Amérique au début des années 1930 et s'est arrêté dans 66 hôtels et 184 restaurants. Un seul établissement a refusé d'accueillir ce couple d'asiatiques. Lorsque six mois plus tard, les établissements visités ont été contactés pour demander s'ils accepteraient d'ouvrir leur porte à des clients asiatiques, les réponses furent négatives dans 92% des cas. Bien que critiquable à bien des égards sur le plan méthodologique, cette étude est l'une des premières à souligner les difficultés à prédire le comportement à partir de l'attitude puisque le comportement effectif des établissements s'avérait différent de leur déclaration.

De même, *Corey* a confirmé ces résultats dans une étude de *1937* en constatant qu'il n'y avait pas de corrélation entre les déclarations des étudiants concernant le comportement de triche et le comportement effectif de triche. Les conclusions de ces études ont amené certains chercheurs à désespérer du concept d'attitude voire à proposer son abandon *(Wicker, 1969).*

Il faudra attendre 1977 pour que *Ajzen et Fishbein* démontrent que seule une attitude précise permet de prédire de manière effective un comportement précis. En effet, *Ajzen et Fishbein (1977)* estiment que la mesure de l'attitude doit correspondre aux mêmes éléments constituant le comportement évalué : **l'action, la cible, le contexte et la temporalité**. Ils démontrent ainsi un lien fort entre attitude et comportement.

Actuellement, le concept d'attitude est prééminent en psychologie en raison de la fonction qu'il occupe dans ses capacités présumées à diriger les comportements *(Petty et Cacioppo, 1996, p.7).* Ces applications du lien entre attitude et comportement trouvent écho dans les domaines de la santé, la prévention routière, l'écologie mais également dans le marketing, l'exercice du pouvoir, la propagande ou la criminologie.

Une des applications les plus évidentes du lien unissant attitude et comportement repose sur la manipulation de l'attitude pour voir apparaître un comportement désiré. En modifiant volontairement l'attitude d'un individu par le biais d'une manipulation, le comportement associé à cette attitude est ainsi plus susceptible d'être réalisé.

Alors que le *passage de l'attitude vers le comportement* abordé dans les paragraphes précédent est relativement commun, le fait que le comportement puisse modifier l'attitude est quant à lui plus surprenant. En effet, l'idée que la réalisation d'un comportement puisse amener l'individu à modifier son attitude est moins acceptable avec la représentation d'un être rationnel qui agit selon ses opinions et adapte donc son comportement à ses attitudes.

L'ordre est ici inversé : l'attitude qui était en cause devient effet et le comportement qui était l'effet vient prendre la place de la cause.

I.2.3. La relation entre attitude et comportement

La relation entre attitude et comportement occupe une place centrale dans la psychologie sociale. Cette discipline étudie les comportements, les états mentaux et processus mentaux chez l'Homme. Conformément à la méthodologie expérimentale, seules quelques variables manipulées font l'objet d'une modification. Quand toutes choses égales par ailleurs, ces seules variables sont responsables d'une modification de l'état initial, il est possible de parler d'une relation de cause à effet. L'introduction volontaire de ces variables est appelée *manipulation expérimentale*.

En psychologie sociale, l'établissement d'un *lien entre les attitudes et les comportements* permet différentes applications pratiques. Si l'on manipule l'attitude d'un individu, on peut modifier ses comportements : on parle alors de manipulation persuasive. Les manipulations comportementales, telles qu'elles sont utilisées par la théorie de la dissonance cognitive, proposent quant à elles un cheminement contraire : en modifiant un comportement, l'expérimentateur génère un changement d'attitude.

I.3. Influence et changement de comportement (persuasion)

Les tentatives de persuasion aboutissent, lorsque le récepteur du message persuasif modifie son attitude dans le sens défendu dans le message. En toute rationalité, on pourrait attendre que la modification d'une attitude entraîne une modification du comportement, comme vu auparavant. Ce raccourci, souvent pris notamment par les tenants de la persuasion technologique, s'accompagne aussi de la polysémie des termes attitude et comportements.

I.3.1. Comment persuader autrui ?

Si cette question occupe les psychologues sociaux depuis six décennies, force est de reconnaître que ce sont les Grecs de l'Antiquité qui ont scellé les premières pierres de l'étude de la persuasion et de la rhétorique.

Déjà Aristote avançait que pour bien persuader (bien éduquer, résoudre des conflits, échanger des idées ...), le rhéteur devait connaître et comprendre les caractéristiques de la source (ethos), du message (logos) et aussi l'état émotionnel de l'audience (pathos). Ainsi, avait-il souligné en son temps qu'une source persuadait d'autant plus facilement autrui, qu'elle était digne de respect.

Persuader revient à déplacer l'attitude initiale d'autrui vers la nôtre.

Les travaux réalisés sur *les origines de l'attitude* ont dégagé l'existence de trois dimensions :

- Cognitive (ce que je sais à propos de l'objet d'attitude),
- Affective ou évaluative (ce que je ressens),
- Conative (ce que j'ai fait).

Si le modèle tripartite de l'attitude *(Rosenberg, 1960)* est le plus cité, d'autres lui ont succédé qui reprennent à leur compte cette partition cognitive-affective-conative. Et pourtant, dans la plupart des modèles et théories de l'attitude, ces trois dimensions sont rarement mesurées. La plupart du temps, les auteurs se contentent de mesurer (via des échelles d'attitudes) le changement sur une seule dimension le plus souvent la dimension évaluative.

I.3.2. Ecole de Yale

La compréhension des caractéristiques et des processus sociocognitifs impliqués dans les phénomènes de persuasion s'inscrit dans une tradition de recherche remontant aux années 1940, connu sous le nom d'École de Yale.

C'est à l'université éponyme que *Carl Hovland* a initié le premier programme de recherches, dont l'ambition était d'appliquer au changement d'attitude les principes du conditionnement opérant.

Le raisonnement tenu par *Hovland* et son équipe peut être résumé comme suit : une attitude est acquise parce que l'environnement, le contexte renforcent son apprentissage. En modifiant l'environnement de

la communication de face à face, on rend dans le même temps l'attitude de la personne moins adaptée à la "nouvelle" réalité. Dans notre vie quotidienne, les modifications de l'environnement se produisent naturellement ; dans de tels cas, le changement d'attitude est spontané.

Mais on pourrait tout aussi bien imaginer de modifier artificiellement l'environnement pour en étudier les effets en termes de changement d'attitude. Ce qu'a fait Hovland dans une série de recherches qui partagent la même procédure expérimentale en trois temps.

Tout d'abord, on mesure l'attitude "spontanée" des individus à l'égard d'un objet social, puis on délivre un message persuasif qui défend une position allant à l'encontre de l'attitude initiale et enfin, on mesure à nouveau l'attitude afin de vérifier dans quelle mesure le message supposé persuasif l'a été (mesure de l'attitude finale).

C'est au cours du second temps que les théoriciens de la persuasion introduisent des manipulations expérimentales, ayant trait aux caractéristiques soit du supposé émetteur du message (son expertise dans le domaine, sa popularité, son attrait, etc.), soit du message lui-même (aspect répétitif du message, sa dimension potentiellement émotionnelle, la nature des arguments qui y sont développés, etc.), soit du récepteur (son niveau d'instruction, son degré de connaissance sur le thème, etc.), soit enfin sur le canal.

I.3.3. 5 W

En d'autres termes, l'École de Yale va œuvrer pendant plusieurs décennies à déterminer les caractéristiques des variables du fameux schéma de la communication dits des 5 W :

- Qui dit
- Quoi
- Par quel Canal
- À Qui
- Avec quel(s) effet(s) ?

Dans notre vie quotidienne, il est bien rare que nous ne fassions qu'une chose à la fois, dans un environnement serein. La plupart du temps, nous sommes entourés par des sources potentielles de distraction. En d'autres termes, notre attention peut être détournée au moment où nous prenons connaissance du message persuasif.

La question qui se pose alors est de savoir dans quelle mesure le fait d'écouter distraitement (ou pas) limite (ou pas) l'impact persuasif du message ?

La réponse dépend là encore des caractéristiques de la cible. Si cette cible ne partage pas le point de vue défendu dans le message persuasif, la distraction augmentera son impact : étant distraits, nous serons moins enclins à chercher des contre-arguments, nos ressources cognitives étant prises par ailleurs *(Festinger & Maccoby, 1964).*

Après avoir choisi la source du message la plus apte à induire un changement d'attitude, après avoir conçu ce message en tenant compte des caractéristiques de la cible, reste maintenant à choisir le canal le plus propice à sa transmission. Si tous les auteurs s'accordent à dire que le choix du canal se fait en fonction du canal privilégié par la cible de la persuasion, il n'y a finalement que peu de recherches qui ont exploré l'impact des différents canaux : écrits, visuels ou oraux.

Eagly et Chaiken (1984) montrent que même en prêtant attention aux messages diffusés à la télévision ou à la radio, nous ne comprenons finalement que peu d'informations (entre 30 et 40 %).

Les canaux et surtout les canaux électroniques mettent la cible en position passive, réduite à prendre connaissance du message diffusé sans réelle possibilité de répondre (ou réagir) aux arguments.

Si on considère que cela pourrait présenter quelques avantages, c'est finalement contreproductif : *Watts (1979)* montre que le changement d'attitude obtenu à la faveur d'un processus de réflexion s'avère sur le long terme plus efficace parce que plus durable que l'écoute passive d'un message persuasif.

Au-delà de la question de la pérennité du changement comportemental, une autre difficulté réside dans le constat que bon nombre de nos comportements relèvent d'habitudes, et sont réalisés la plupart du temps sans même qu'on y pense :

"Most of the time, what we do is what we do most of the time"*(Townsend & Bever, 2001).*

Pour professeur Verplanken, un des spécialistes du domaine, l'habitude "est une séquence apprise d'actions qui deviennent des réponses automatiques à des situations particulières, et qui est fonctionnelle pour atteindre certains buts ou états" *(Verplanken & Aarts, 1999, p. 104).*

Si la répétition du comportement est nécessaire à l'émergence d'une habitude, elle ne saurait être suffisante : la qualité qui définit une habitude est l'automaticité du comportement. C'est parce que la répétition du comportement dans un contexte consistant augmente peu à peu l'automaticité avec laquelle un comportement est réalisé dans ce contexte particulier ou tout autre contexte qui lui ressemble *(Verplanken, 2006 ; Wood & Neal, 2007).*

Par automaticité, on comprend un comportement observable réalisé de façon peu raisonnée, mentalement efficient, sans intention consciente, et difficilement contrôlable *(Bargh, 2004).*

Verplanken suggère que des habitudes peuvent être changées si on intervient au moment opportun, c'est-à-dire à un moment où la situation, le contexte change. C'est le cas par exemple, lors d'un déménagement, de l'arrivée d'un enfant dans un foyer, d'un divorce.

Le changement d'attitude, c'est-à-dire l'impact persuasif d'une communication est conçu par les tenants de Yale comme à un mécanisme en trois temps.

- Premier temps, il faut que la cible de la persuasion soit attentive au message
- Second temps, il faut qu'elle le comprenne
- Troisième temps, il faut qu'elle l'accepte, au moins en partie

La clé du changement d'attitude serait donc à rechercher dans l'environnement, le contexte de la communication.

I.3.4. La persuasion : approche cognitive : Modèle de Vraisemblance d'Élaboration

Dans les années 1980, un modèle alternatif du changement d'attitude s'est développé avec l'idée que le changement d'attitude (la persuasion) n'était plus à chercher dans le contexte, mais dans les mécanismes psychologiques liés au traitement de l'information, donc dans la tête de l'individu.

Cette approche cognitive du changement d'attitude est connue sous le nom de modèle de Vraisemblance d'Élaboration ou ELM *(Cacioppo & Petty, 1986).*

L'ELM propose qu'une même variable puisse influencer les attitudes de façons différentes. La même variable, selon le rôle qu'elle joue, peut augmenter ou réduire la persuasion. Elle peut le faire à travers plusieurs mécanismes.
Le modèle de la vraisemblance d'élaboration repose sur le principe "qu'il existe deux routes de persuasion, que l'individu peut emprunter selon les circonstances par exemple, et qu'emprunter l'une ou l'autre des routes produit des conséquences différentes ... par conséquent, le modèle de vraisemblance s'intéresse aux différents processus de persuasion qui peuvent opérer dans différentes situations" *(Petty, 1994, p. 3).*

L'ELM fait partie des modèles duaux de la persuasion dans le sens où le changement peut être la résultante de processus psychologiques différents. La personne traite le message persuasif de façon attentive (voie centrale) ou de façon plus superficielle (voie périphérique). Si ces deux voies débouchent sur un changement d'attitude, elles reposent en revanche sur des degrés différents d'élaboration du traitement de l'information.

Elaboration selon *Petty et Cacioppo (1986, p. 128)* représente la mesure dans laquelle une personne non seulement pense aux arguments contenus dans le message, mais élaborent aussi de nouvelles pensées. Ce travail cognitif d'élaboration peut être quantifié, notamment en ayant recours à la technique du listage de pensées, cette technique étant issue de la théorie des réponses cognitives *(Greenwald, 1968).*

Exemple : On demande à la cible de la persuasion de noter toutes les idées qui lui sont venues en tête pendant la réception du message. Puis on lui demande de noter si chacune de pensée est en faveur, en défaveur ou neutre par rapport au message persuasif délivré. Ne reste ensuite qu'à calculer un indice de "favorabilité" des pensées en soustrayant le nombre de pensées négatives (défavorables) au nombre de pensées positives (i.e. favorables), le tout divisé par le nombre total de pensées pertinentes (Cacioppo & Petty, 1981 ; Brinol & Petty, 2003).

L'indice de favorabilité a souvent été utilisé pour évaluer dans quelle mesure le changement d'attitude était relié à la quantité et à la nature du traitement de l'information du message persuasif.

I.3.5. Nudge

Les pères de ce concept sont Richard Thaler, expert en finance comportementale, et Cass Sunstein, directeur des affaires réglementaires de l'administration Obama. Tous les deux sont convaincus qu'il est possible d'orienter subtilement les choix individuels en faveur de l'intérêt collectif sans altérer le libre arbitre.

Selon la théorie des nudges, mêmes parfaitement informés, nous sommes tous des "pneumologues fumeurs en puissance", susceptibles de prendre des décisions contraires à nos intérêts et à ceux du collectif. Comment, dans ces conditions, guider l'individu vers une décision et donc le comportement souhaité ?

Toujours maître de nos décisions ? Pas si sûr… À l'heure des choix, tout individu se laisse influencer par de nombreux facteurs :

- L'impact de l'environnement : notre état d'esprit et nos choix sont influencés par nos sens. Une odeur de propre dans un vestiaire de piscine nous amènera à être plus soigneux et nous accorderons plus de crédit à un document bien présenté qu'à un texte mal mis en page.
- L'aversion à la perte : nous sommes plus sensibles aux pertes qu'aux gains. Ainsi, la "souffrance" à l'idée de perdre 10 € sera plus grande que la joie de gagner la même somme.
- Le statu quo : l'aversion à la perte nous amène également à préférer conserver quelque chose (une habitude, un privilège) plutôt que d'en changer, même si c'est dans notre intérêt.
- Le poids de la norme : nous avons tendance à vouloir nous aligner sur le comportement majoritaire.
- La superficialité : dans notre processus décisionnel, nous privilégions les informations qui nous reviennent en premier en mémoire et notamment les événements récents ou très médiatisés. De même, les décisions que nous avons déjà prises pèsent beaucoup dans nos choix, même si les paramètres ont changé.

Face à cette "rationalité limitée", le mécanisme du nudge est simple : pour amener quelqu'un (sans contrainte) à faire un choix, le contexte dans lequel celui-ci s'opère doit être modifié en s'appuyant sur tous les biais qui déterminent les processus décisionnels.

C'est ce qu'on appelle "l'architecture du choix", qui vise à créer un environnement mental incitatif.

Formalisées par *Thaler et Sunstein* en 2008, les techniques de nudging ont rapidement rencontré un large écho dans la sphère politique avec la création d'unités de recherche dédiées au sein des gouvernements américains et anglais. De nombreux autres acteurs se sont aujourd'hui mis à appliquer cette méthode : organisations non gouvernementales, associations, pouvoirs publics…et également des "criminels".

La plupart des initiatives visent à faire adopter des comportements durables ou responsables aux citoyens, avec parfois un impact bien plus puissant que les traditionnels bâtons et carottes.

I.4. Détection du comportement

Les comportements humains sont complexes. Il est difficile de déterminer quels comportements posent un risque pour soi-même ou pour les autres. Certaines personnes ne présentent aucun signe extérieur visible de leurs motivations internes. D'autres adoptent un éventail de comportements qui indiquent clairement leur intention de faire du mal. Et d'autres personnes encore adoptent un éventail de comportements qui semblent indiquer clairement leur intention de faire du mal, même s'ils n'ont pas du tout de mauvaises intentions. Il y a aussi un lien entre les intentions meurtrières et suicidaires, et il y a des questions contextuelles et des facteurs liés à la personnalité qui rendent impossible la prédiction d'actions que prendront les personnes.

Ce que l'on peut faire, c'est observer les signes souvent liés aux comportements à risque ou aux comportements menaçants.

I.4.1. Modèles normaux de développement

La plupart des comportements prennent un cours plus ou moins prévisible. Généralement, au fur et à mesure que les personnes acquièrent des expériences positives de la vie et mûrissent, leur comportement reflète une évolution vers des choix plus complexes, autonomes et autosuffisants. Bien que toute personne soit unique, ses comportements reflètent normalement des choix socialement acceptables qui lui apportent une satisfaction personnelle. Il arrive toutefois, de temps en temps, que les personnes et les jeunes s'éloignent de cette voie et en empruntent une autre de nature plus négative et destructrice. Nous devons remarquer ce changement de voie et tirer la sonnette d'alerte le plutôt possible et cela devrait nous permettre de détecter la cause du changement de leur comportement habituel.

I.4.2. Détection de changements

Les modèles normaux de développement sont différents pour chaque personne. Lorsque vous essayez de détecter des changements, vous cherchez des changements par rapport au modèle normal de développement de la personne concernée ou par rapport à son comportement de base. Lorsqu'un quelqu'un s'éloigne de son comportement de base normal, il faut avoir la possibilité de le détecter. Par exemple, une personne qui accepte normalement des changements dans ses activités quotidiennes mais qui soudainement devient enragé au sujet d'un changement serait en train de s'éloigner de son comportement de base.

Un tel éloignement devient troublant lorsque la personne adopte un comportement moins approprié en ce qui concerne son développement. Tout personne qui devient plus renfermé, agressif, anxieux ou rebelle que d'habitude, par rapport à son comportement de base, devrai attirer l'attention de son environnement.

I.4.3. Théories du changement comportemental

Les théories de l'engagement *(Kiesler, 1971 ; Joule & Beauvois, 2013; Joule & Beauvois, 1998)* se sont avérées efficaces pour obtenir qu'une personne fasse ce qu'on attend d'elle. Cette approche préconise de placer cette personne en situation d'acteur, en lui faisant réaliser des comportements peu coûteux dans un premier temps, avant de l'amener à réaliser le comportement-cible, généralement coûteux en termes d'effort, d'investissement, etc.

Contrairement aux théories de la persuasion, on n'a pas besoin de développer un argument convaincant pour produire du changement. Il nous suffit de placer la personne dans un contexte qui l'amène à produire d'elle-même les comportements que l'on attend d'elle. On peut penser qu'après avoir réalisé ces comportements, la personne ne modifiera pas ses attitudes dans le sens de l'acte (ou des actes) qu'elle vient de produire *(Joule & Beauvois, 1998 ; Fointiat, Girandola, Gosling, 2013).* Dans la lignée de ces travaux, certains auteurs ont jeté un pont conceptuel entre théories de l'engagement comportemental et persuasion en développant un nouveau modèle appelé la communication engageante *(Joule, Girandola & Bernard, 2007).*

Le principe en est simple : aux questions classiques du champ de la persuasion, telles que :

- Quel type d'information à donner ?
- Quels sont les meilleurs arguments ?
- Quels sont les médias, canaux, supports de communication les plus appropriés ? …

Il convient d'ajouter :

- Quels actes préparatoires obtenir ?

Cette question et sa réponse constituent la pierre angulaire du modèle de la communication engageante, en donnant au "récepteur" de la communication le statut d'acteur.

Le changement de comportement est compris comme linéaire et relativement constant. Pourtant, il est possible que le changement de comportement soit non-linéaire. Premièrement, parce que tout le monde n'est pas prêt à changer de comportement au même moment. Deuxièmement, on peut changer le comportement aujourd'hui et revenir à nos anciens comportements demain. Partant de ce double constat, *Prochaska et DiClemente (1982)* élaborent un modèle théorique dans lequel le processus de changement passe par différentes étapes.

L'étape de pré contemplation désigne un stade dans lequel les individus ne sont pas prêts au changement. Ils sont même résistants : pour eux, le problème n'existe pas ; ils ne peuvent donc pas envisager de changer de comportement.

L'étape de contemplation marque la prise de conscience par l'individu qu'un problème existe ; ils commencent à envisager qu'il faudra le résoudre à un moment ou à un autre, sans pour autant être prêts à enclencher une dynamique de changement. Pour dépasser la contemplation, les individus s'engagent dans l'examen des avantages et des désavantages liés à leur situation actuelle. Ils peuvent alors prendre la décision de faire les premiers pas vers le changement : c'est le passage à l'étape de préparation.

L'étape de préparation marque le moment où les intentions et l'action se rencontrent : les efforts engagés ici ne sont cependant pas suffisants pour contrôler les nouveaux comportements. Il convient donc de consolider la décision et de maintenir les premiers 'petits' changements.

L'étape de l'action, c'est sans doute la phase la plus active du changement : les individus réorganisent leurs attitudes, leurs croyances, restructurent leur environnement pour résoudre le problème. L'énergie et les efforts requis au cours de cette phase pour enclencher un changement sont tels qu'ils sont visibles par des observateurs, et sont donc susceptibles d'être reconnus et valorisés par les proches.

Maintien de l'action est la dernière étape : les changements comportementaux sont visibles ; il s'agit donc maintenant de les maintenir dans le temps. Cette phase n'est pas statique : l'individu peut "rechuter" dans ses anciens comportements, par exemple s'il se retrouve dans un contexte favorable à la résurgence des comportements abandonnés.

Ce modèle implique qu'à chaque stade de progression vers l'adoption d'un nouveau comportement, l'individu peut faire une rechute et revenir à un stade antérieur. En d'autres termes, *Prochaska, DiClemente et Norcross (1992)* préfèrent à une vision linéaire du changement de comportement une vision cyclique, en spirale.

I.5. Rétrospective de l'analyse criminelle

L'analyse criminelle trouve son origine en Amérique du Nord dans les années 60 avec le développement de la criminalité organisée. A cette époque, des méthodes utilisés par des services de renseignement se font désapprouvées car ils montrent leur inefficacité et l'élaboration de nouveaux programmes de renseignement se met en route.

Ainsi, un ensemble de méthodes analytiques normalisées pour élaborer des hypothèses, reconstituer le déroulement des faits, déterminer si des infractions avaient été commises par le même auteur, comprendre le fonctionnement des réseaux de malfaiteurs et étudier l'ampleur et les caractéristiques des activités criminelles ont été mises au point. L'ensemble de ces techniques normalisées ont constitué ce qui a été convenu d'appeler l'analyse criminelle.

Les origines de l'analyse comportementale contemporaine ou "profilage criminel" remonte également dans les années 60 aux Etats-Unis. A cette période, une série d'attentats frappait les salles de cinéma de New-York depuis 17 années.

Vu que les méthodes traditionnelles se sont révélées inefficaces, la méthode du *Docteur A. BRUSSEL*, un psychiatre spécialisé en criminologie, a été appliqué. Docteur Brussel parvint très rapidement à dresser le portrait psychologique de l'auteur des attentats et il formule des suggestions d'enquête. Sur cette base, les enquêteurs réussirent à arrêter l'auteur dans les deux mois suivants. Depuis, le FBI s'est consacré au développement des sciences comportementales comme outil d'aide à l'enquête.

Toutefois, les origines de cette technique ne doivent pas être uniquement recherchées outre-Atlantique. En effet, à la fin du 19ème siècle, la France a connu une affaire criminelle, l'affaire "Vacher", qui semble révéler les origines de cette même technique "à la française".

Contrairement à l'analyse criminelle, le constat de l'absence d'une définition reconnue de l'analyse comportementale s'est rapidement imposé.

En premier lieu, il existe une composante technologique qui constitue un moyen mis au service des enquêteurs pour effectuer des rapprochements judiciaires et pour les aider à détecter un phénomène de sérialité.

Vu que cette composante s'avère insuffisante, elle est complétée par la composante comportementale, destinée à apporter les connaissances nécessaires aux enquêteurs dans le domaine du comportement humain aux fins de leur fournir des orientations d'enquête.

Aux USA, pour faire face à phénomène de taux de criminalité qui augmente, le programme VICAP (Violent Criminal Apprehension Program) a été établi. Il s'agit d'un logiciel informatique dans lequel sont entrées les caractéristiques précises de tous les meurtres commis dans tous les Etats américains. Ces données

sont ensuite traitées informatiquement par les agents du FBI, et le logiciel VICAP effectue des recoupements avec d'autres meurtres commis sur l'ensemble du territoire américain. Depuis sa mise en place, le programme VICAP a contribué à l'arrestation de nombreux tueurs en série. En effet, ça ressemble aux débuts d'utilisation du "Big Data".

D'autres banques de données informatisées centralisant les prélèvements, échantillons et indices de différentes natures relevés dans le cadre d'enquêtes criminelles existent aussi. Depuis 1969, le FBI a créé au sein de son département de recherche une unité spécialisée d'étude du comportement criminel ("Behavioral Sciences Unit") à laquelle peuvent avoir recours l'ensemble des enquêteurs américains.

Dans le tableau suivant, on repère des statistiques élaborées en fonction d'identification de suspect *(Jean-Paul Brodeur CICC)* :

Facteurs opérant dans l'enquête d'identification

Facteurs déterminants	Identification : suspect 1 (n = 153)	Identification : suspect 2 (n = 28)	Identification : autres suspects (n = 12)
Témoin oculaire	22,5 %	22 %	–
Aveux spontanés	20,5 %	3,7 %	–
Informateur de police	12,5 %	26 %	33 %
Patrouilleurs	10,6 %	–	–
Victime / covictime	10,6 %	–	–
Dénonciation par un proche	3,3 %	15 %	16 %
Dénonciation par un parent	2,6 %	–	–
Dénonciation par un conjoint	0,66 %	–	–
Enquête de routine	2 %	–	–
Surveillance physique et électronique	1,3 %	–	–
Instigation	1,3 %	–	–
Parade d'identification par des photos	0,66 %	–	–
Section renseignement	0,66 %	–	–
Police scientifique	0,66 %	–	–
Intervention extérieure	0,66 %	–	–
Autres	1,3 %	–	–

Tableau 1 : statistiques en fonction d'identification de suspect

Or, aujourd'hui les données utilisées ne proviennent pas seulement de ces sources "officielles" mais plus souvent des données dites « open source », la tâche devient encore plus difficile car la quantité d'informations est telle que le traitement des données prend de plus en plus de temps.

Avec l'arrivée du Big Data en particulier, les manquements dans le traitement des données de masse et le tri des données se voit améliorer mais ne donne pas pour autant des résultats dit "just in time".

I.6. RGPD et des nouvelles technologies – problème de profilage

Le Règlement Général sur la Protection des Données (RGPD) s'appliquera en mai 2018 à toute entreprise qui collecte, traite et stocke des données personnelles. C'est-à-dire, toute donnée dont l'utilisation peut directement ou indirectement identifier une personne. Ce règlement a donc un impact direct sur la manière dont les données personnelles sont collectées et stockées.

Il est important de s'intéresser dès à présent au détail des obligations qui s'imposeront lors de la collecte des données. Et plus particulièrement des données collectées en ligne.

Ce type de collecte est qualifié de "collecte directe" par le règlement.

En effet, la GDPR intègre une différenciation entre la collecte directe et la collecte indirecte. La collecte directe représente les données collectées directement, alors que la collecte indirecte représente l'achat ou location d'un fichier de données personnelles. Quelle que soit la nature de la collecte, les exigences du RGPD sont les mêmes.

Quelques changements sont donc à anticiper afin de respecter le règlement à partir de mai 2018.

I. Recenser les données collectées et identifier celles à caractère personnel
II. Limiter la collecte des données personnelles au minimum nécessaire
III. Obtenir et conserver le consentement éclairé et informé des individus

En vertu du GDPR, les "entreprises" ont l'obligation d'obtenir un consentement explicite, éclairé et accordé librement. Cette notion implique une plus grande transparence sur l'utilisation et la raison de la collecte ainsi qu'une plus grande obligation d'information sur les conditions particulières du traitement des données personnelles. Il est aussi obligatoire d'être en capacité de prouver que le consentement a bien été donné par l'utilisateur.

IV. Sécuriser et protéger les données collectées

Pour une "entreprise" le prix d'une donnée personnelle va possiblement augmenter car elle sera dans l'obligation de documenter toutes les mesures et procédures utilisées afin d'assurer la protection des données stockées. Les obligations de sécurité impliquent de se protéger contre les risques de perte, de vol, de divulgation ou de compromission.

La collecte des données personnelles se devra désormais d'être justifiée, transparente et sécurisée dans un souci de respect de la vie privée des citoyens européens.

Le profilage est l'un des grands enjeux du Big Data associé aux algorithmes prédictifs. Il désigne l'analyse et la prédiction du comportement, des achats, du rendement, des déplacements, des préférences… d'une personne physique sous la base des données collecté des sources divers et variées.

Le RGDP encadre cette technique susceptible de porter gravement atteinte aux droits fondamentaux. Ainsi, une mesure de profilage doit s'accompagner d'une étude d'impact.

Le Règlement européen relatif aux données personnelles autorise le profilage mais l'encadre strictement.

Le RGDP définit le profilage comme "toute forme de traitement automatisé de données à caractère personnel consistant à utiliser ces données à caractère personnel pour évaluer certains aspects personnels relatifs à une personne physique, notamment pour analyser ou prédire des éléments concernant le rendement au travail, la situation économique, la santé, les préférences personnelles, les intérêts, la fiabilité, le comportement, la localisation ou les déplacements de cette personne physique".

Si le RGPD bouleverse la façon d'appréhender la gouvernance des données à caractère personnel, la question du profilage demeure un enjeu important, si ce n'est vital, pour les nombreuses sociétés effectuant ce genre de traitement.

Les traitements non automatisés semblent de facto exclus car non mentionnés par le texte. La notion n'en reste pas pour le moins étendue car le profilage inclut des activités comme la détection de fraudes, le marketing comportemental ou les méthodes de sélection dans l'octroi de crédit ou en matière d'assurance dans l'appréciation des risques.

Le Groupe de travail du G29 (Réunion des autorités de contrôle des différents pays de l'Union Européenne comme la Cnil en France) a publié le 17 octobre dernier des lignes directrices (guidelines) sur le sujet qui visent à rassurer le secteur du marketing notamment.

Ces lignes directrices sont divisées en 5 sections :

- Définition du profilage et des prises de décision automatisées au regard du RGPD
- Dispositions spécifiques sur les prises de décision automatisées tels que définies à l'article 22 du RGPD
- Dispositions générales sur le profilage et les prises de décision automatisées
- La question du profilage des mineurs
- L'évaluation de l'impact sur la protection des données à caractère personnel

La prise de décision automatisée et le profilage vont nous particulièrement intéresser. Effectivement, les deux dernières sections étaient des points qu'il fallait éclaircir et ce sont les dispositions spécifiques et générales relatives à la prise de décision automatisée.

Le G29 différencient plusieurs types de profilages associés ou non avec une prise de décision automatisée:

1. Le profilage général non associé à un processus de décision purement automatisé et les processus de décision basée sur du profilage avec intervention humaine sont tout à fait possibles sans autre exigence que de répondre aux principes posés par le Règlement. (Traitement licite basé sur

l'intérêt légitime du responsable de traitement, information, finalité précise, modalités d'exercice des droits et *opt out* (se désengager) mentionnés)

A titre d'exemple sur l'intérêt légitime, fondement du traitement : "l'existence d'un intérêt légitime devrait faire l'objet d'une évaluation attentive, notamment afin de déterminer si une personne concernée peut raisonnablement s'attendre, au moment et dans le cadre de la collecte des données à caractère personnel, à ce que celles-ci fassent l'objet d'un traitement à une fin donnée."

A noter que la prévention de la fraude ou la prospection commerciale peuvent répondre à cette notion d'intérêt légitime.

2. En revanche, les processus de décision entièrement automatisés basés sur du profilage sans intervention – humaine, susceptible de produire des effets juridiques ou affectant les personnes de manière significative de façon similaire font l'objet de conditions plus stricte car vus, par le législateur et le G29, comme possiblement néfastes aux droits et libertés des personnes. Il est important de préciser que rentre également dans cette catégorie des processus où l'intervention humaine serait insignifiante et sans conséquence sur la décision finale.

Spécifiant l'interdiction de principe de ces processus de décision automatisée ayant des effets juridiques, le groupe de travail rappelle les exceptions possibles, définies par le RGPD en son article 22, lorsque le traitement :

- Est nécessaire à la conclusion ou à l'exécution d'un contrat entre la personne concernée et un responsable du traitement"

Le G29 fournit des exemples d'utilisation pour améliorer les prises de décisions et réduire les erreurs nées du facteur humain, permettre une réponse plus rapide ou simplement accéder à un service qui ne peut être humainement réalisable, le G29 rappelle les responsables de traitement à leur obligation *d'accountability* (montrer comment le principe de responsabilité est mis en œuvre et à le rendre vérifiable) et de justification du caractère nécessaire de l'opération pour l'exécution du contrat.

Ce qui devra la plupart du temps amener le responsable du traitement à envisager une analyse d'impact car c'est désormais à lui d'analyser et puis de décider s'il met en place ou non le traitement sans recourir à une autorisation de la CNIL.

- Est autorisée par le droit de l'Union ou le droit de l'État membre auquel le responsable du traitement est soumis et qui prévoit également des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée"

Cette disposition est intéressante car laisse la possibilité aux Etats membres d'édicter leur propre règle, notamment en matière de prévention de la fraude ou d'évasion fiscale, là où le RGPD cherchait une unification de la réglementation au sein de l'Union.

- Est fondée sur le consentement explicite de la personne concernée.

Bien sûr, le RGPD ne vient pas définir cette notion de consentement explicite sauf à préciser que cela résulte d'une déclaration écrite ou de tout autre acte non équivoque impliquant le consentement.

Le G29 reste quant à lui silencieux sur le sujet renvoyant à de futures lignes directrices et à la notion de consentement simple.

Le consentement sera requis lorsqu'il y aura une décision entièrement automatisée sans intervention humaine et avec conséquence juridique possible.

Dans ce cas, les Responsables de traitement, s'appuyant sur le consentement pour établir leur traitement de profilage et décision automatisée, devront bien sûr démontrer que les personnes concernées ont bien été informées et ont compris ce à quoi elles consentent afin de justifier d'un choix éclairé. Comme tout traitement de données personnelles, il est nécessaire pour le Responsable de traitement d'évaluer les conséquences probables et l'impact de la conduite de son activité basée sur ses intérêts légitimes en les confrontant aux droits et libertés des personnes concernées.

Le G29 énonce des critères pour cette évaluation préalable :

- Le niveau de détail du profil (exemple d'une personne ayant fait l'objet d'un profil au sein d'une catégorie large en tant que "professeurs d'anglais natifs vivant à Paris", ou bien segmentés et ciblés plus finement)
- L'exhaustivité du profil (si le profil ne décrit qu'un petit aspect du sujet de données, ou dresse un tableau plus complet)
- L'impact du profilage (les effets sur la personne concernée)
- Les garanties visant à garantir l'équité, la non-discrimination et l'exactitude du profilage

S'appuyant sur un exemple, celui d'une pharmacie en ligne se livrant à des activités de prospection fondées sur les achats de médicaments et d'autres produits, le groupe de travail met en avant le danger d'un résultat trop précis pouvant entrainer la révélation d'informations relevant de la catégorie et des dispositions spécifiques relatives aux données sensibles (données de santé révélée par le profilage trop ciblé ou exhaustif).

En conclusion il n'est donc pas exigé, de recueil du consentement, hormis pour les cas de décisions automatisées ayant des effets juridiques ou similaires, sans intervention humaine. Il y a lieu de veiller néanmoins à ce que la conduite des opérations de profilage non associées à un processus de décision

automatisée et celle des processus de décision basée sur du profilage avec intervention humaine induisent un résultat respectant une certaine granularité, un niveau de détail qui resterait adéquat et proportionné à la finalité définie, le Responsable de traitement n'étant tenu que de respecter les principes d'information et d'exercice des droits des personnes concernées y compris le droit d'opposition.

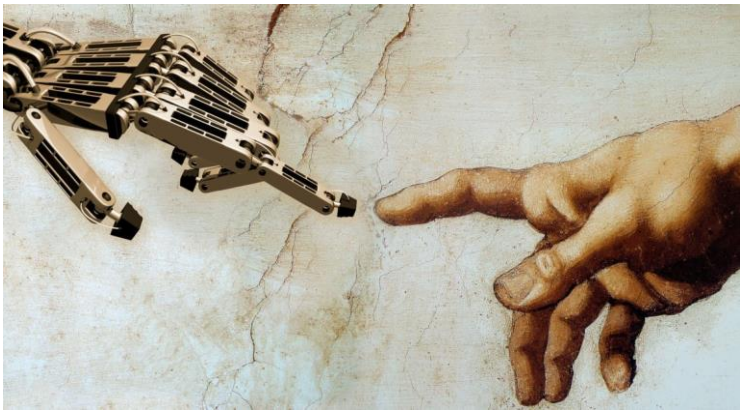
A noter que le droit d'opposition au traitement dans le Règlement Européen fait explicitement référence aux traitements intégrant du profilage et particulièrement en matière de prospection commerciale, laissant la possibilité pour les personnes concernées de s'opposer sans motif. Cette opposition peut aller dans ce cas de prospection jusqu'à l'effacement, et pourra conduire à la suppression des profils créés ou tout au moins à leur anonymisation.

En considérant des points importants que RGPD nous impose, on se trouve rapidement restreint dans le procédé de la recherche des informations, le stockage, le croisement des données et l'utilisation des algorithmes prédictifs dans le but d'une action préventive. D'autant plus si nos actions peuvent mener à utiliser ces informations pour potentiellement condamner une personne. En quelques mots, on se trouve dans une quasi impasse judiciaire et technologique.

Avoir le consentement d'une personne de se positionner comme une cible de profilage me semble peu probable. C'est pour cela que les sources d'informations doivent être bien choisies et que la manière de récupérer des informations du type Big Data doivent également être bien réfléchies. Une autre question surgit : comment faire pour ne pas se noyer dans la quantité des données recueillies et comment réussir à trouver de bonnes informations à temps ?

Pour répondre à ces questions on va s'intéresser d'abord aux techniques existantes et aux nouvelles technologies utilisées aujourd'hui et enfin au data science comme moyen de l'analyse comportementale.

II. ∞ ETAT DE L'ART ∞



Après avoir vu ce qu'est le comportement et comment on peut le modifier et influencer, il est intéressant de voir comment on peut appliquer le Big Data et le mettre au service de l'analyse et prédiction du comportement. Pour comprendre ce phénomène et voir plus précisément comment cette technologie est appliquée dans le service de police et renseignement national et international, on va s'intéresser à expliquer ce phénomène de Big Data.

Premièrement, je compte en parler des origines du Big Data avec la part de marché qu'il prend aujourd'hui tout simplement pour se rendre compte que les données qu'on propage sont par tout autour de nous et en grande quantité. De plus, le marché crée autour de ces données démontre un besoin qui ne cesse de croitre.

Pour bien comprendre les fondements du Big Data, je parlerai de cinq principes qui le définissent plus précisément ainsi que le défi qu'il nous pose.

Avant de commencer l'analyse de quoi que ce soit, il faut comprendre comment on peut faire cette analyse et quelles principes-algorithmes nous avons à disposition. Pour y parvenir, des différents types d'analyses utilisées aujourd'hui vont être présentés. Suite logique s'impose : si des principes d'analyse sont connu, quels modèles peut-on appliquer pour faire l'analyse ? La réponse à cette question se trouvera à la suite.

Ensuite, maintenant qu'on connait les bases du Big Data, il plus facile de parler d'application et d'impact sur l'analyse criminelle avec les débuts d'application des algorithmes prédictives. On va également parler des sources différentes de données qu'on peut utiliser pour faire cette analyse.

Avec l'apparition des premiers logiciels d'analyse des données et leurs applications dans la police on démontre le besoin justifié et exponentiel ainsi que ses inconvénients. Ensuite, je vais parler des logiciels

le plus utilisés dans le service de police en USA, Europe et France, pour démontrer que le besoin est mondial et applicable partout. Comme les algorithmes que l'on utilise ne sont pas infaillibles, la question de la discrimination dans l'analyse prédictive se pose naturellement.

Avec des logiciels utilisés aujourd'hui, on peut voir qu'ils ont tous un point commun, que par la suite, et avec la mise en application de la loi sur la vie privé et informations, GDPR, nous mets en difficulté d'exploitation des mêmes données. Dans la troisième partie de ce mémoire, je vais essayer de traiter le sujet du temps de traitement du profilage qui est étroitement lié avec la quantité des données et proposer des solutions qui peuvent être appliqués en amont de cette analyse "condamné" à "prendre" du temps.

II.1. BIG DATA

Le développement d'internet et la multiplication des objets connectés à travers le monde s'accompagnent d'une croissance exponentielle des données créées sur internet. La multiplication des moyens de communication et d'échange n'y est pas étrangère. En effet, les différents écrans nous suivent partout, tout au long de la journée.

En 2018, il y avait près de 23 milliards de terminaux connectés dans le monde et ce chiffre devrait s'élever à 75 milliards en 2025, si l'on en croit cette étude de *Statista (tableau 2)*. Outre les smartphones, tablettes et télévision connectées, les nouveaux objets connectés, tels que les voitures, les appareils électroménagers ou encore les montres connectées qui déferlent sur le marché devraient remonter une quantité phénoménale d'informations dans les années à venir.

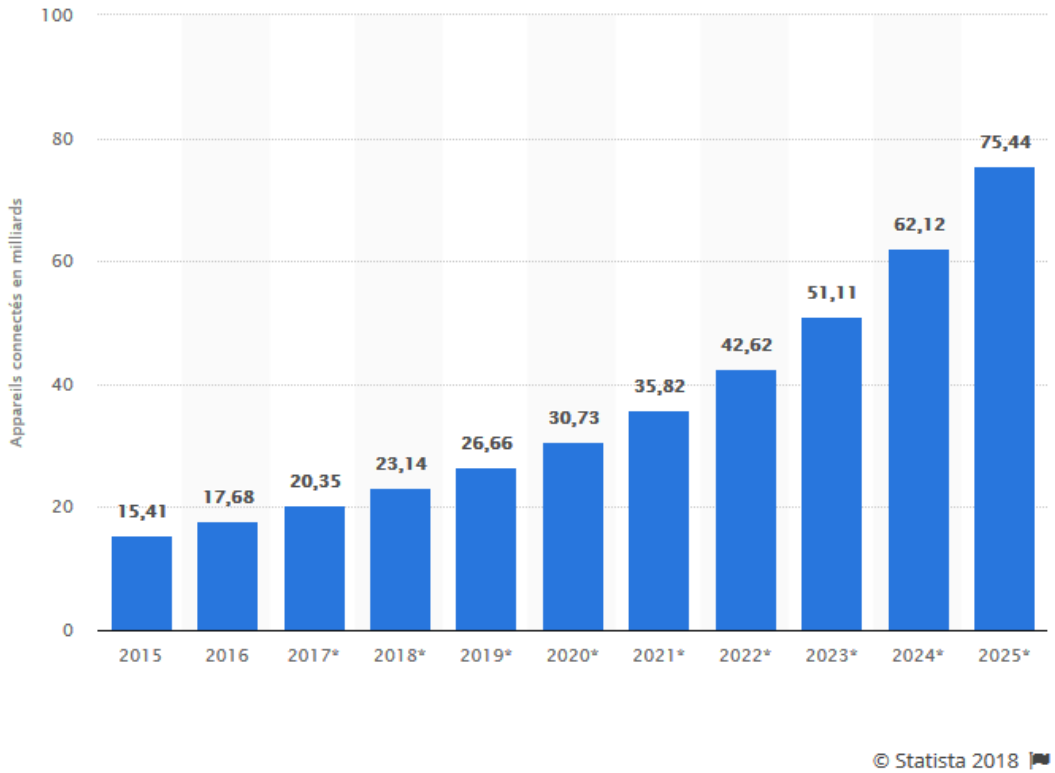


Tableau 2 : appareils connectées/an

Cette statistique représente le nombre d'appareils connectés (Internet des objets : IoT) dans le monde entre 2012 et 2025. En 2017, la source a prévu que 20 milliards d'objets connectés seraient en circulation dans le monde.

Les informations disponibles sur internet ne sont plus seulement volumineuses, elles sont également très diverses. Cela peut passer par des vidéos, de la musique, des photos… Elles ne sont pas structurées au sens où elles ne se présentent pas sous la forme de lignes et de colonnes comme aime à être structuré le web. Ces données doivent donc être structurées avant d'être analysées et exploitées par les technologies actuelles. À titre d'exemple, chaque jour produit plus de vidéos que les cinquante premières années de la télévision, cela représente donc une masse d'informations à structurer gigantesque *(Thomas H. Davenport).*

À l'heure actuelle, les Hommes sont les principaux responsables de ce volume d'informations. En effet, toutes leurs interactions avec les nouvelles technologies génèrent des données : téléchargement d'un fichier, consultation d'une vidéo, coup de téléphone, envoi de SMS, utilisation de GPS… ce n'est pas tant les interactions en tant que telles qui génèrent autant de données, mais c'est surtout l'ensemble des informations annexes (les métadonnées) et des communications " cachées " entre différents serveurs (publicitaires par exemple) qui ont lieu au même moment qui génèrent un flux impressionnant de données.

L'ensemble de ces milliards de données, c'est ce que l'on appelle communément les " Big Data ". Les premières entreprises à avoir compris leur intérêt sont les géants du web actuel tels que Google, Yahoo, Microsoft, Facebook ou bien Amazon.

Du fait de leur succès ou de leur volonté de vouloir gérer une quantité très élevée d'informations, des entreprises ont dû apprendre à maitriser ces Big Data, car les outils et méthodes traditionnelles ne leur suffisaient plus.

II.1.1. Le marché des Big Data

Le marché du Big Data est très récent ; il a émergé vers la fin des années 2000 aux États-Unis. Ce marché, poussé par les géants du net américain, Google en tête, représentait déjà un chiffre d'affaires de 3,2 milliards de dollars en 2010, mais ce n'est rien à côté de ce qu'il est devenu aujourd'hui.

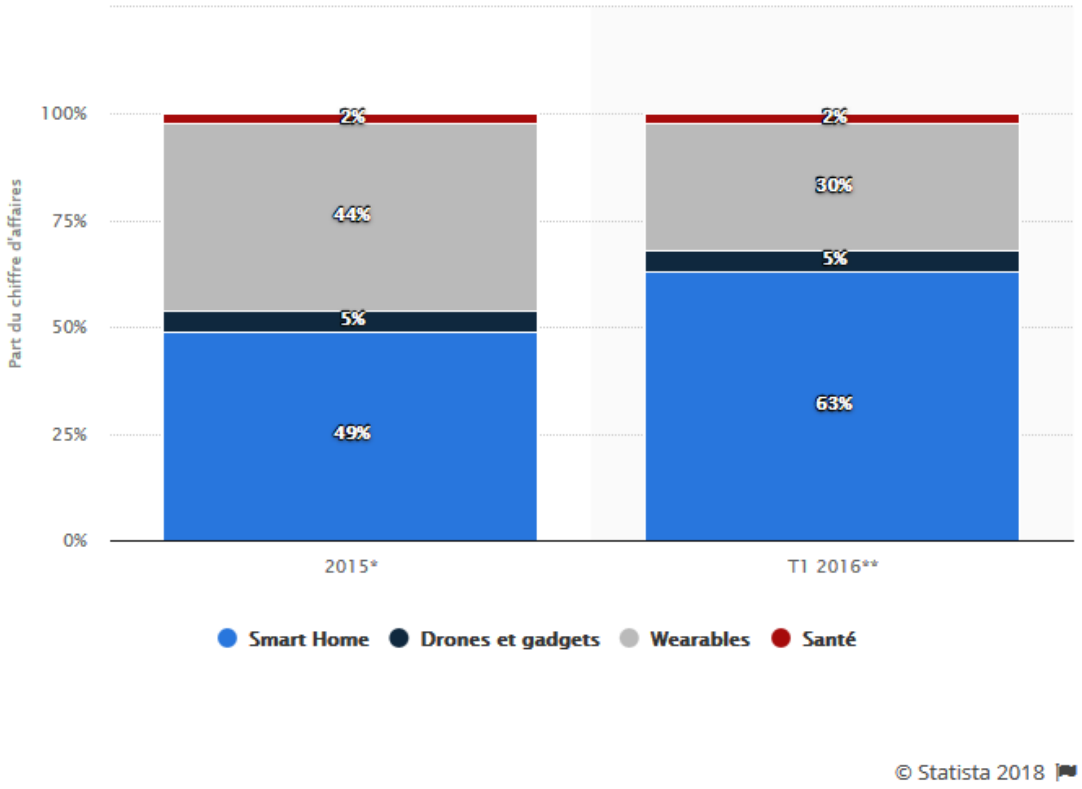


Tableau 3 : CA objets connectés en France

Tableau 3 montre la répartition du chiffre d'affaires du marché des objets connectés en France en 2015 et durant le premier trimestre de l'année 2016, par segment. Le segment « Smart Home » représentait près de la moitié du chiffre d'affaires de ce marché en France durant l'année 2015.

Tableau 4 représente le volume des ventes de smartphones, tablettes, ordinateurs portables et wearables en France de 2011 à 2015, ainsi que les estimations pour l'année 2016. En 2014, il s'est vendu près de 6,2 millions de tablettes en France.

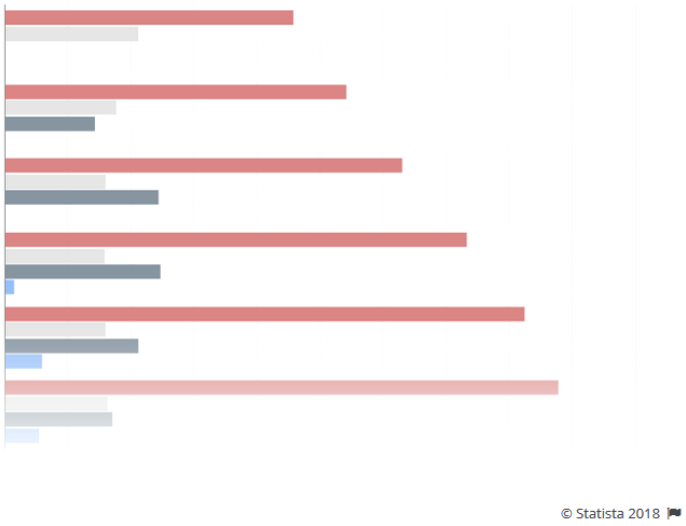


Tableau 4 : ventes de smartphones, tablettes, ordinateurs portables et wearables en France de 2011-15

Bien entendu nous retrouvons de grands noms de l'informatique tels que Microsoft, Cisco, Oracle, IBM, mais des entreprises spécialisées dans le Big Data ont également émergé telles que WibiData, Hadapt ou encore Domo. Toutes ces entreprises, pour l'essentiel basées aux États-Unis, ont effectué des levées de fonds impressionnantes, par dizaines de millions de dollars, pour pouvoir assurer le développement de leur entreprise, preuve que les investisseurs croient en ce marché du Big Data.

Une autre entreprise bien connue, IBM, n'a pas hésité à dépenser près de 20 milliards de dollars dans des rachats dans le secteur de l'analytique afin de s'assurer une place sur ce marché.

II.1.2. De nouveaux outils pour la prise de décisions

Le cœur de la survie d'une entreprise repose sur les décisions que prennent les dirigeants. Avant la démocratisation d'internet dans les entreprises, les décisions étaient prises en fonction de paramètres très génériques sur le secteur dans lequel évoluait l'entreprise. Il était très dur de connaitre les stratégies des concurrents et voir ce qui se faisait dans le monde. Aujourd'hui cela a bien changé et il serait inconsidéré de prendre une décision sans avoir récupéré au préalable des éléments très précis sur l'état du marché dans le monde et avoir analysé finement la concurrence. Le genre d'applications qui offrent déjà ce genre de service sont appelées Informatique décisionnel ou Business Intelligence.

Désormais, la moindre information se doit d'entrer en ligne de compte dans les solutions de Business Intelligence. Celle-ci peut être sous forme structurée ou non, mais elle se doit d'être analysée et remontée afin de faciliter la prise de décisions. Elle doit également être " fraîche " c'est-à-dire qu'il s'agit d'analyser en quasi temps réel l'état du marché et de l'entreprise. Les décideurs doivent disposer de l'ensemble des

informations disponibles dans l'entreprise et à l'extérieur de celle-ci afin de prendre des décisions en corrélation avec l'état réel du marché.

Les éditeurs actuels de solutions de Business Intelligence ont de plus en plus de mal à suivre la cadence et à répondre aux attentes des entreprises dans ce domaine, d'autant plus que le Big Data revient généralement moins cher sans toutes les contraintes que présentent ces anciens systèmes.

Les géants du web ont déjà positionné leurs produits, BigQuery chez Google ou Redshift chez Amazon, et proposent des solutions de Business Intelligence à la sauce Big Data à des tarifs compétitifs en comparaison aux solutions traditionnelles (500€ par To/mois)

II.1.3. Quels résultats espérer des Big Data

Après avoir vu ces cas d'utilisation selon les services de l'entreprise, il est temps de voir quel impact pourrait avoir concrètement la mise en place du Big Data au sein d'une entreprise. Les technologies du Big Data permettent aux entreprises d'atteindre des objectifs variés, mais il est important pour elles d'en cibler un en particulier lors du déploiement d'une infrastructure Big Data.

Le premier d'entre eux est la réduction des temps d'exécution et d'analyse ; en effet, dans certaines entreprises, il peut exister des processus mettant énormément de temps à être traités du fait de la taille des données ou de la complexité du calcul. La technologie des Big Data permet, grâce à son architecture distribuée (qui fonctionne sur plusieurs machines en parallèle), de gagner énormément de temps de traitement et donc de pouvoir complexifier encore plus les algorithmes de calcul en y incorporant de plus en plus de paramètres.

Le deuxième objectif est la réduction des coûts. Selon une estimation de Thomas Davenport, les coûts de stockage pour une année de 1 To de données sont de 37 000 dollars avec un SGBD (Système de gestion de base de données) classique, 5 000 dollars avec un système de stockage et seulement de 2000 dollars avec un cluster Hadoop. S'il est possible d'économiser sur le coût du matériel, la mise en place d'une solution Big Data exige tout de même l'embauche de " scientifique des données " ayant un salaire assez élevé et d'autres frais annexes.

Enfin le développement de nouvelles offres et l'amélioration des offres existantes représentent l'aboutissement des objectifs, car il permettra à l'entreprise de dégager de nouveaux bénéfices et lui permettra de se détacher de ses concurrents via une offre unique. Google est le parfait exemple d'une entreprise ayant bâti l'ensemble de ses offres sur les technologies du Big Data. Cependant le développement de nouvelles offres implique une certaine dose d'innovation et donc des coûts supplémentaires. On ne peut pas attendre du développement de ces offres une rentabilité à court terme, mais à long terme ils peuvent assurer une source de revenus supplémentaire pour l'entreprise.

II.1.4. Le volume

Le volume se réfère à la grande quantité de données générées. L'océans de données se remplit de plus en plus rapidement, maintenant que " l'Internet des choses " est là pour lier ensemble des milliards de dispositifs - la télé, des réfrigérateurs, des dispositifs de sécurité, les thermostats, les détecteurs de fumée - qui produisent et partagent des données. De nombreux ensembles de données sont tout simplement devenues trop grandes pour stocker, traiter et analyser avec la technologie de base de données traditionnelle.

De nouveaux paradigmes tels que MapReduce, No SQL bases de données, flux et mémoire de traitement ont été introduits pour fonctionner de manière optimale sur les systèmes distribués. Ces solutions évolutives horizontalement sont en mesure de traiter plus de données en ajoutant des composants des produits de base. Sur le plan économique, cela est une bien meilleure solution que super-type d'ordinateur de solutions (mémoire partagée par exemple). Mais même avec ces développements, la loi de Moore *(Moore 1965)* nous dit que concernant le CPU, la capacité double tous les 18 mois, et la capacité de stockage est doublée tous les 14 mois.

Par conséquent, les données deviennent exponentielles et incompressibles. Les seules options pour relever ce défi sont de devenir plus intelligent sur les données à analyser, ou de créer des algorithmes plus intelligents.

II.1.5. Variété

La variété fait référence aux différents types de données. Le Big Data aborde des défis au-delà de l'analyse des données avec des structures claires. Tous nos bavardages sur les médias sociaux, la musique en streaming et les vidéos, nos e-mails et ainsi de suite n'ont pas de structure prédéterminée claire (souvent improprement appelées données non structurées).

Les outils Big Data sont en grande partie indifférents sur les structures spécifiques, ce qui permet la collecte et l'analyse des différents types de données, telles que les messages, les conversations sur les médias sociaux, les photos, les données des capteurs, des enregistrements vidéo ou la voix, etc…

La combinaison de connaissances de ces types de données avec l'analyse de données classique peut aboutir à une compréhension plus profonde du comportement des (groupes de) personnes, les systèmes, les maladies, les processus, etc.

II.1.6. Vitesse

La rapidité fait référence à la fois au taux croissant de collecte de données et à la demande croissante de connaissances en temps réel et des réponses. Ces systèmes viennent avec leurs propres défis. Du côté de la collecte, les décisions sont prises en temps réel : certaines données sont conservées et certaines données sont rejetées. Du côté de la décision, plusieurs exemples existent où l'analyse de données a été

prise hors contexte, ou elles étaient tout simplement fausses, et aux décisions ultérieures en temps réel qui ont donné lieu à de graves instabilités dans un écosystème.

L'exemple le plus célèbre est probablement celui de 6 mai 2010 et du krach de Wall Street lorsque l'indice Dow Jones a perdu 10 pour cent de sa valeur en cinq minutes, pour récupérer la majeure partie de sa valeur peu après. L'accident a été attribué à des algorithmes de trading automatique défectueux, la prise de décisions en temps réel sans intervention humaine.

II.1.7. Véracité

La véracité fait référence à l'exactitude et la précision des informations. À l'ère du Big Data, les perceptions changent. Durant la dernière décennie, la gestion des données de base, la qualité des données et la gouvernance des données ont été au cœur des efforts de la plupart des grandes organisations pour obtenir un aperçu de leurs données. Bien que la qualité des données reste une question importante, l'omniprésence des données ne permet pas de gérer et de contrôler la qualité de toutes les sources. Cela est souvent déjà vrai dans les grandes organisations, mais devient particulièrement pertinent lorsque les données d'autres organisations sont incluses dans une analyse. Pourtant, nous avons besoin de savoir comment fiabiliser les ensembles de données. Cela peut souvent être réalisé par le traitement d'une quantité croissante de données, suivant le mantra " quantité sur la qualité " *(Halevy et al., 2009),* soit en augmentant la taille de l'échantillon pour réduire les erreurs statistiques ou en ajoutant des sources indépendantes pour réduire les erreurs systémiques. Bien que cela puisse sembler une solution straight-forward, des statistiques avancées et des techniques de modélisation sont souvent nécessaires pour estimer correctement l'impact de la combinaison des ensembles de données *(Hair et al., 2006).* L'ajout d'ensembles de données augmente également la complexité. Il y a des limites sur le nombre de sources que l'on peut intégrer et comprendre.

II.1.8. Valeur

Les quatre premiers V se concentrent sur les éléments techniques du Big Data. L'un des traits distinctifs du Big Data est qu'il est principalement tiré par l'activité *(Zikopoulos et Eaton 2011),* et non par la technologie. On constate que la valeur produit par des données devient importante.

Le Big Data offre une organisation différente. Elle permet d'offrir de nouveaux produits et services et de faire des choses qui étaient jusqu'à présent impossibles. Elle donne la chance à tous les secteurs et industries.

Beaucoup de gens associent le terme Big Data avec des entreprises qui veulent vendre à des clients toujours plus de choses en apprenant tout ce qu'il y a à savoir sur eux. Mais c'est juste un revers de la médaille. Il y a aussi la valeur de la société dans Big Data.

Nous pouvons l'utiliser pour sauver des vies humaines en optimisant les voies de soins *(Groves et al. 2013).*

Nous pouvons améliorer de façon plus rentable la maintenance et la planification, par exemple en analysant les données recueillies avec des capteurs sur les ponts et en dehors des installations à terre *(Feblowitz 2013).*

Nous pouvons augmenter les rendements agricoles à partir des informations obtenues à partir des terres agricoles de balayage *(Kaloxylos et al.),* par exemple avec des satellites ou des drones.

Tous ces exemples ont déjà été mises en œuvre et sont en cours d'élaboration en continu.

Et bien sûr, nous pouvons utiliser ces données dans le but d'améliorer notre sécurité en faisant la prédiction et plutôt prévention sur des actions illégales.

II.1.9. Le défi de la définition de 5V de Big Data

Bien que les 5 V soient la définition commune de Big Data, il existe plusieurs alternatives. L'un est de l'Institut McKinsey Global dans leurs études de marché 2011 sur Big Data *(Chui et al 2011).:*

" Big Data se réfère à des ensembles de données dont la taille est au-delà de la capacité des outils logiciels de base de données typiques pour capturer, stocker, gérer et analyser. Cette définition est volontairement subjective et intègre une définition en mouvement définie sur la taille d'un jeu de données pour être considéré comme Big Data - à savoir, nous ne sommes pas de défini Big Data en termes d'être plus grand qu'un certain nombre de téraoctets. Nous partons du principe que, comme la technologie progresse au fil du temps, la taille des ensembles de données qui se qualifient comme Big Data permettra également d'augmenter ".

Avec l'affirmation selon laquelle les capacités sont nécessaires au-delà des outils généralement disponibles, cette définition classifie Big Data en tant que science plutôt que l'ingénierie. Ceci est conforme à la façon dont la plupart des praticiens Big Data se voient : comme des " scientifiques de données " qui, selon Thomas Davenport dans la Revue de Harvard business, est le travail le plus sexy du 21 st siècle *(Davenport et Patil 2012).*

Le mot " grand " dans Big Data peut induire en erreur. De nombreuses nouvelles applications associées à Big Data ne sont pas sur la modification ou l'interprétation des quantités énormes de données - le soi-disant " grand, les données en désordre " - mais de la combinaison intelligente des quantités limitées de données pour la personnalisation.

En second lieu, il y a des développements étroitement liés au Big Data qui agissent comme des catalyseurs, tels que l'informatique en nuage et l'Internet des objets (IOT).

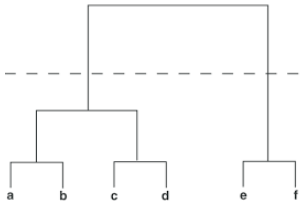
Parfois, ces développements sont confondus avec le Big Data, par exemple cloud computing vise à transformer l'architecture *(Abadi 2009),* alors que le Big Data vise à transformer le processus décisionnel.

II.2. Analyse des données

L'analyse Big Data est similaire à l'analyse des données traditionnelles, car elle repose sur l'application des méthodes statistiques appropriées pour extraire des caractéristiques ou des connaissances à partir de données. Les progrès technologiques et le dédoublement des données, cependant, ont donné lieu à plusieurs techniques nouvelles, ou améliorées, d'analyse. Nous commençons par quelques-unes des méthodes d'analyse des données traditionnelles *(Snijders 2011)* appliquée au Big Data avant d'aborder quelques-unes des nouvelles méthodes plus avancées.

II.2.1. L'analyse par grappes

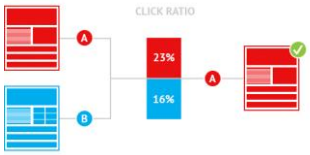
L'analyse par grappes, c'est à-dire la différenciation et la classification des objets en fonction des caractéristiques particulières. L'identification des corrélations, dire des changements dans les variables sont liées les unes aux autres, que ce soit par des relations de dépendance stricte ou indéterminé.



II.2.2. Analyse de régression

Analyse de régression, une méthode mathématique pour déterminer les corrélations.

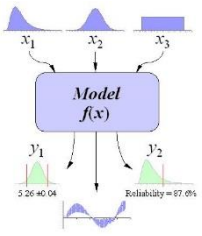
II.2.3. Test A/ B



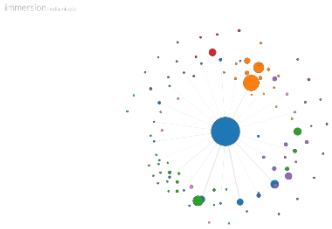
Un test A/ B, tests d'hypothèses en appliquant différentes conditions à différents groupes et la comparaison des résultats.

II.2.4. Simulation Monte Carlo

Simulation Monte Carlo, l'analyse du comportement d'un système par échantillonnage aléatoire de prédéfinie distributions de probabilité.



II.2.5. Traitement de graphique

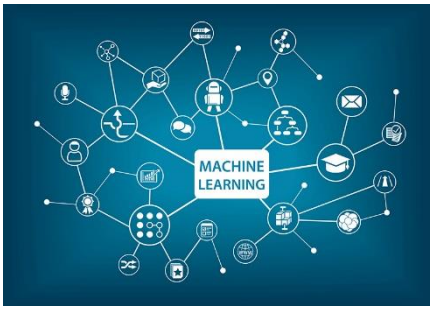


Certaines méthodes traditionnelles, comme le traitement de graphique, deviennent plus puissant grâce à de nouveaux modèles de programmation. Le graphique est constitué de nœuds et des relations entre les objets. Il est généralement difficile à analyser quand ils deviennent trop grands.

Cependant, les extensions de MapReduce permettent d'analyser de grands graphiques *(Lin et Schatz 2010)* par exemple à déterminer le chemin le plus court entre deux nœuds ou d'identifier les communautés dans les réseaux sociaux.

II.2.6. Machine Learning

Avec le développement des nouvelles technologies et le mis en place des algorithmes d'apprentissage tels que *la forêt aléatoire (Breiman, 2001)* ou de *machines à vecteurs de support (Cortes et Vapnik 1995)*, Il est permis aux ordinateurs d'être formés, plutôt que programmé, soit surveillé ou non surveillé.



En cas d'apprentissage supervisé, par exemple les entrées sont fournies en même temps que la sortie désirée, alors qu'avec l'apprentissage non supervisé, les algorithmes sont censés identifier des modèles cachés ou inconnus dans les données.

II.2.7. Informatique cognitive

Une variation sur le thème de l'apprentissage de la machine est informatique cognitive, rendu célèbre par ibm lorsque Watson - un super-ordinateur capable de répondre à des questions posées en langage naturel. Bien que Watson n'emploie des réponses basées sur des règles prédéfinis dans l'intelligence artificielle (combinée à des systèmes de récupération de l'information de qualité supérieure et le traitement du langage naturel), il montre que les capacités cognitives des ordinateurs peuvent déjà supérieurs à ceux des êtres humains, au moins à certains égards.

II.2.8. Deep Learning

On peut dire que la forme la plus avancée de l'analyse Big Data est l'apprentissage en profondeur, une relance de ce qu'on appelait autrefois les réseaux de neurones. L'un des principaux bénéfices de l'apprentissage en profondeur est la capacité d'apprentissage non supervisé ou semi-supervisé, permettant l'extraction de fonction automatisée ou sans minimum d'intervention humaine, ce qui est essentiel lorsque les ensembles de données croître de façon exponentielle.

Deux entreprises, Vicarious et Numenta, font des progrès significatifs répliquant le néocortex, la partie analytique du cerveau humain, avec des techniques d'apprentissage en profondeur.
Google a récemment acheté une société appelée Deep Mind faisant des recherches similaires.

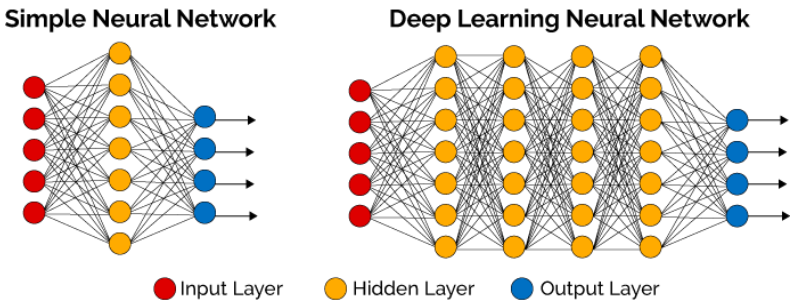


Tableau 5 : réseau de neurone

Les développements ci-dessus ont suscité un regain d'intérêt " singularité technologique " - le moment où les systèmes intelligents dépasseront la capacité intellectuelle humaine. Selon cette hypothèse, ces systèmes seront en mesure de créer des systèmes plus intelligents encore, ce qui provoque un effet d'inspiration, l'accélération des progrès au-delà de la compréhension humaine et la civilisation changer d'une manière qui est fondamentalement imprévisible de l'état avant la singularité.

II.3. Gestion de données

Les principaux défis de travailler avec le Big Data comprennent la représentation des données (comprendre la signification des données), la redondance de l'information (est la cohérence des données et que fait-on garder), la gestion du cycle de vie (Quelle est la fiabilité données et combien de temps peut-on garder), le traitement distribué (comment obtenir une réponse dans un délai raisonnable), confidentiel (qui on peut partager les données avec), l'évolutivité (comment faire face à la croissance exponentielle des données) et de la coopération (qui est responsable de quoi, qui doit investir, et comment faire face aux risques).

II.3.1. Modèle traditionnel

Le champ traditionnel face à ces problèmes est connu comme la gestion des données de base, comprenant les processus, la gouvernance, les politiques, normes et outils toujours de définir et de gérer les données critiques d'une organisation pour fournir un point de référence unique *(Loshin 2010).*

L'objectif principal de la gestion des données de base est sur la bonne intégration des sources de données et d'assurer la qualité des données avant l'analyse. La gestion des données de base peut être un processus de temps.

Avec la tendance croissante des cycles de livraison courts, par exemple grâce à des méthodes telles que manifeste le développement agile, une approche complémentaire gagne en popularité dans le monde du Big Data.

Plutôt que la définition des normes dès le départ, les données provenant de diverses sources (parfois à l'extérieur de l'organisation) sont réunies sans changer leur structure d'origine dans un " data lake ".

Les activités liées à la qualité et la normalisation des données sont reportées à la phase d'analyse. Les bénéfices de cette approche sont la facilité de mise en œuvre (ne sont pas traitées de manière globale les questions de normalisation et de la qualité des données) et flexibilité, les normes et les exigences de qualité peuvent différer d'une analyse à l'autre. Les inconvénients comprennent la nécessité de revoir les données de définitions et la qualité des données pour chaque analyse, qui peut conduire à un double travail et à la confusion.

II.3.2. Modèles hybrides

Les modèles hybrides apparaissent le plus probable, où un data lake contient une ou plusieurs couches de qualité et normalisation des données, en combinant les points forts de la gestion des données de base classique avec la souplesse d'une approche du Big Data.

En plus de ce changement d'évolution de l'approche de gestion des données de base traditionnelles, le Big Data conduit également à des solutions de gestion de données peu orthodoxes.

II.3.2.1. Bit Torrent

La popularisation de ces solutions a commencé avec le protocole Bit Torrent, qui est devenu très populaire pour télécharger de la musique et des films. Le système fonctionne sans ordinateurs centraux ou des partenaires intermédiaires. Le téléchargement de données avec un torrent se résume à obtenir les morceaux de centaines d'autres utilisateurs du réseau. La redondance et les incitations fournies par ce réseau pair à pair leur permet de fonctionner de manière fiable, même si la fiabilité de chaque fournisseur de données individuelles est limitée.

II.3.2.2. Bitcoin

Le traitement des transactions a récemment subi une transformation similaire avec l'introduction de Bitcoin *(Nakamoto 2008).* La confiance n'est pas assurée par un intermédiaire mais elle est établie par une solution mathématique qui assure la confiance à l'intérieur du réseau lui-même.

Dans sa forme la plus simple, tous les participants du réseau ont un registre de toutes les transactions de tous les autres participants sur leur système que l'on appelle la Block Chain.

La Block Chain est mise à jour avec chaque transaction, ce qui permet des transactions sécurisées, même quand il y a des partenaires peu fiables dans le réseau. En d'autres termes, le réseau assure la confiance qui supervise et vérifie la bonne exécution des transactions.

La façon dont les réseaux torrent et la Block Chain ont changé le traitement des données a entraîné et continue d'entraîner des changements fondamentaux aux applications et modèles d'affaires. Ils sont en quelque sorte le fondement de la révolution Big Data, qui touche toutes les industries basées sur l'information. Mais ils sont aussi que le début ; les nouveaux développements sont déjà sur la voie à l'adoption généralisée.

II.4. Maintien de l'ordre, Big Data et la marchandisation de la sécurité

La sécurité (sociale, juridique et autre) a été l'une des grandes réalisations de l'époque moderne. Récemment, la multiplication des solutions technologiques est non seulement en train de modifier le visage de la police, mais apporte également des changements sociaux plus larges qui influencent la compréhension de notre société de ce qui constitue la sécurité.

Les pages suivantes ont pour le but de lier l'évolution récente de la police dont la sécurité avec les possibilités et les risques des nouvelles technologies et des solutions Big Data.

II.4.1. Maintien de l'ordre et de la sécurité à la fin du 20ème siècle

Dans les années 70 et 80, l'augmentation des taux de criminalité et une érosion globale des mécanismes de contrôle social traditionnels qui sont caractéristiques des sociétés individualisées ont mis une énorme pression sur les budgets. Cela a ensuite conduit à une demande croissante de services de police efficace. Ce changement a été couplé avec de plus en plus d'appels à la participation communautaire, la participation et la collaboration de la police pour atteindre l'objectif de maintenir la sécurité, en particulier à l'échelle locale.

Ce fut aussi la période où les criminologues ont commencé à identifier un nouveau problème : le sentiment d'insécurité en tant que phénomène indépendant des statistiques de l'insécurité objective. Bien que de nombreux risques liés à la violence ont disparu pendant cette période, les gens ont exprimé la crainte grave encore du crime qui existe à l'extérieur la possibilité d'être victime d'un acte criminel.

Le processus de décentralisation, cependant, coexiste avec l'émergence de la " police mondiale ". Par exemple, les bases de données partagées et les échanges de personnel, la surveillance et les réseaux mondiaux de police sont déployés pour améliorer la poursuite de la criminalité transnationale. Dans ce contexte, les agences Europol et Interpol sont des exemples clés d'organisations de sécurité visant à surmonter les problèmes qui se posent lorsque la police est confinée aux frontières nationales. En dépit de cette portée mondiale, les forces de police en Europe continuent d'être fragmentés en raison de facteurs géographiques, fonctionnels, historiques, culturelles ou administratives.

En plus de l'interconnexion potentielle des fonctions et des compétences, une question importante se dégage autour de la question de l'échange d'informations. La sécurité est une zone significative de la souveraineté que les Etats membres hésitent à déléguer au sein de l'Union Européenne (UE).

Cela signifie que les informations recueillies par certaines forces de sécurité sont souvent inaccessibles à d'autres organismes, augmentant ainsi l'effort nécessaire et la diminution de l'efficacité. Le partage des données entre les pays d'UE peut apporter des informations précieuses aux nations.

Alors que les entreprises privées se développent et le nombre d'entreprises de sécurité augmentent, il y a une quantité croissante de renseignements liés à la sécurité répartis sur un certain nombre d'élargissement des organisations. Cette répartition des ressources à travers une large gamme d'acteurs peut donner accès à de nouvelles sources de données, mais il laisse aussi des organisations sans système centralisés.
Sans cadres juridiques clairs, le partage de protocoles, les technologies de partage de données et la complexité des nouveaux systèmes de partage de données public-privé pourrai aggraver l'accès à des informations utiles.

Les principales tendances qui peuvent être observées dans des services de police et de sécurité depuis les fin des années 90 sont une augmentation d'échange entre la défense, la police et l'intelligence ; l'accent sur les données et la prévision (par opposition à la prévention); et une dépendance croissante des solutions technologiques pour accroître l'efficacité.
Ces évolutions s'intensifiées après des événements frappants, notamment l'attaques du 11 novembre contre les tours jumelles aux Etats-Unis.
Il s'agit d'un phénomène nouveau dans lequel la sécurité est renforcée pour faire face à une menace mondiale qui se manifeste localement, quels que soient les frontières géopolitiques.

II.4.2. L'impact de la technologie

La technologie a été au centre de la plupart des développements liés à la police de ces dernières années. La technologie a toujours été présente dans la police : des photos des suspects ont été utilisés par la police rapidement après l'invention de la photographie. Les CCTV caméras à des fins de sécurité ont été introduites dans les années 1950, par exemple.
Les bases de données ont également été utilisées depuis longtemps dans les services de police, d'abord sous la forme physique et les plus récemment sous forme numérique.

Les années 80 marquent le début de l'utilisation de différentes sources de données grâce à l'ordinateur. En utilisant des statistiques et des chiffres pour produire des cartes de la criminalité et géolocaliser différentes catégories d'événements sur ces cartes, la police a commencé à se rendre compte que les modèles prédictifs pourraient voir le jour, l'intelligence décisionnelle pourrait être suscitée pour rendre les services de sécurité plus efficaces et sûrs.

II.4.3. Sources d'information

La liste des sources d'information peut comprendre :

- Statistiques (par exemple, les systèmes de police prédictifs, casier judiciaire) ;
- Capteurs CCTV, caméras intelligentes, détection d'activité inhabituelle, des appareils intelligents de la ville, etc. ;
- Des documents biométriques (des empreintes digitales ADN, etc.) ;
- Cartographie : Géolocalisation des mineurs et les personnes âgées, les piétons flux, trafic heatmaps, etc. ;
- Plaques d'immatriculation des dossiers ;
- Appareils mobiles : Automobile Systèmes de navigation, téléphones mobiles ;
- E-services publics et privés : e-shopping, l'activité des médias sociaux, e-mail, e-gouvernement;
- " Internet des objets " : appareils intelligents, wearables et domotique, etc. ;
- Citoyens et dossiers clients : les dossiers financiers, les enregistrements de noms de passagers, les casiers judiciaires, les dossiers de véhicules, les informations fiscales, le paiement d'informations, etc.

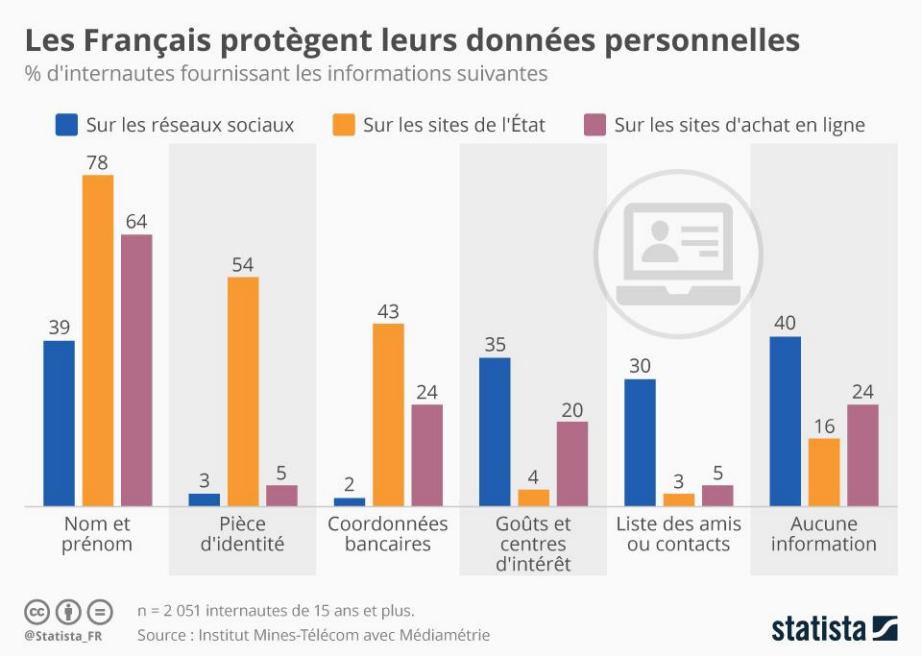


Tableau 6 : Les Français et la protection des données

Ces nouveaux types de technologies de surveillance qui emploient cette logique préventive sont connus comme la surveillance préventive. La surveillance préventive peut être définie comme "la collecte et le traitement systématique ou ciblé des données des entités, qui sont utilisés pour faire des prédictions sur les dommages futurs sur la base des profils avec l'objectif principal d'intervenir avant que le mal ne soit fait".

II.5. La police prédictive

Depuis les années 1990, l'évolution des services de police axée sur le renseignement en Europe et les théories de plus en plus populaires de la prévention du crime de la situation ont conduit à l'élaboration de l'analyse de point d'accès et plus tard à l'analyse hotspot prospective. Cela a ensuite conduit au développement d'une des premières applications de police prédictive au Royaume-Uni, connue sous le nom ***Promap***.

A peu près en même temps que *Promap* a été développé dans le Royaume-Uni à la fin de la première décennie du 21ème siècle, IBM élaborait l'un des premiers services de police des applications prédictives aux États-Unis, en utilisant les bases de données des crimes passés et des données telles que les temps typiques de la journée et les types de temps en corrélation avec la criminalité pour identifier les tendances et cartographier la prévision.

La police a démontré qu'à la suite des compressions budgétaires, la police prédictive devrait devenir l'avenir de la police.

À la suite de l'émergence de méthodes analytiques plus accessibles, du développement d'algorithmes plus complexes et d'efforts pour rendre la police plus productive et moins chère, les analystes de la criminalité génèrent maintenant des prédictions sur l'endroit où le crime est susceptible de se produire et où les suspects sont susceptibles de se trouver.

Avec le logiciel qui devient plus intelligent et plus puissant et la capacité de stockage de base de données quasi illimité, il est maintenant de plus en plus possible de faire des prévisions sur la base des quantités énormes de données structurées et non structurées provenant de sources différentes.

II.5.1. Types de police prédictive

Une distinction peut être faite entre les deux types de police prédictive :

- Cartographie prédictive
- Identification prédictive

Le première type fait référence à l'application de l'analyse prédictive pour prédire quand et où un crime peut avoir lieu à un niveau global.

La deuxième identification est de type prédictif, où l'analyse est faite au niveau individuel ou de groupe. Cela peut se concentrer sur la prévision des délinquants potentiels, l'identité des délinquants, des comportements criminels et les victimes potentielles de la criminalité.

Le type le plus couramment utilisé est la cartographie prédictive. Illustrations de ce type de logiciel de police prédictive comprennent, mais sans s'y limiter, le système ***Predpol*** aux US et Royaume-Uni, système de sensibilisation à la criminalité ***CEMFA*** aux Pays-Bas, ***Precobs*** en Allemagne et en Suisse, et ***Keycrime*** en Italie, comme ***Anacrim*** en France.

De nouvelles applications expérimentent avec des combinaisons de différentes sources de données. *(Bogolomov et al. 2014),* par exemple, présenter un type de police prédictive dans laquelle ils utilisent des données extraites à partir de téléphones mobiles et les données démographiques.

Une autre tendance émergente qui est l'utilisation de méthodes de hotspot, qui sont liés aux données des médias sociaux de Twitter et Facebook, par exemple, de faire des prédictions. Dans ce cas, l'algorithme recherche pour une utilisation de langue particulière qui indique une plus grande chance de la criminalité dans une certaine zone.
Par exemple, si les gens parlent de sortir, d'aller dans les pubs, et se saouler, ce sont des indicateurs qui sont identifiés par les modèles d'exploration de données. A partir du moment où les données sont collectées, la GPS balises dans les tweets permettent de visualiser les menaces et les points critique pour les crimes potentiels.

Un exemple du second type de police prédictive, identification prédictive en utilisant Big Data, est une application utilisée dans le US appelé **Intrado Beware**, ce qui a été vendu aux services de police américains depuis 2012. C'est une application mobile qui fonctionne au sein de Motorola Solutions new Intelligent Data Portal (IDP) plate-forme, une application mobile, basée sur le cloud qui rassemble des informations contextuelles des bases de données publiques et commerciales existantes *(9-11.com Magazine 2014).*

Selon la société qui a conçu la technologie :
"Accès par un navigateur (fixe ou mobile) sur tout appareil compatible Internet y compris tablettes, smartphones, ordinateur portable et les ordinateurs de bureau, Intrado Beware trie et scores des milliards de documents commerciaux publiquement disponibles dans en quelques secondes - alerte répondeurs à des situations potentiellement dangereuses en cours de route, ou à l'emplacement, une demande d'assistance 9-1-1 (2012). Intrado"

Après avoir analysé l'information des médias commerciaux, criminelle et sociale, **Intrado Beware** algorithme attribue ensuite un score et une note de menace (vert, jaune ou rouge) à une personne, qui sont envoyés automatiquement à celui qui a fait une demande. Le logiciel a été d'abord piloté par la Ville de Thornton Département de la police dans le Colorado et est actuellement mis à l'essai à Fresno *(Intrado 2012, Hoggard 2015).*
Cependant, il y a également eu des services de police, comme Bellingham, qui ont décidé de ne pas acheter le logiciel après que les citoyens ont exprimé leurs préoccupations au sujet des répercussions sur les coûts et la vie privée.

II.5.2. Efficacité de la police prédictive

En général, il est trop tôt pour tirer des conclusions convaincantes sur l'efficacité des applications de police prédictive décrites ci-dessus. Le nouveau logiciel n'a pas été mis en place assez longtemps pour obtenir une image claire, et seulement quelques vastes évaluations indépendantes ont été réalisées (voir, par exemple, Hunt, Saunders et Hollywood 2014).

Les études menées par les policiers eux-mêmes sur des projets pilotes comme ***Predpol*** montrent qu'il existe une significative probabilité concernant des crimes contre les biens : les cambrioleurs ont un comportement territorial, quand ils trouvent une zone où il y a beaucoup de choses à voler ils y reviennent régulièrement.

D'autres résultats positifs ont été trouvés par Mastrobuoni (2014), qui a procédé à une évaluation des **Keycrime**.

Cet auteur a montré que Keycrime, qui est utilisé par la police à Milan, peut aider à réduire le taux de criminalité et peut aider les réduire les coûts de la police.

En revanche, des résultats peu concluants ont été obtenu durant un projet pilote dans le Kent Royaume-Uni où *Predpol* a été mis en place (Kent Police 2013). Une évaluation plus approfondie des logiciels de police prédictive, qui a été utilisé pour prédire des crimes contre les biens aux États-Unis, n'a pas pu trouver ici des effets statistiquement significatifs *(Hunt, Saunders et Hollywood 2014).*

En conclusion, ces technologies sont encore très nouvelles, en particulier les technologies prédictives d'identification, les évaluations futures devront faire la lumière sur leur efficacité pour la prévention du crime. À l'heure actuelle, aucune évaluation prédictive des applications d'identification n'a été trouvé. Cependant, tant que la méthodologie sous-jacente de ces nouvelles technologies ne sont pas remis en question critique, et tant qu'il n'y a pas de normes de qualité obligatoires d'évaluation *(Farrington 2003)* pour vous assurer que les évaluations sont menées correctement, ces types d'évaluations doivent être interprétées avec soin.

II.6. Outils d'analyse

II.6.1. *Anacrim*

Conçu au milieu des années 1990 pour la Gendarmerie nationale, *Anacrim* permet de traiter un grand nombre de données différentes et de faire des recoupements : téléphonie, comptes bancaires, procès-verbaux, analyses ADN, documents divers, etc.

Le logiciel se compose de quatre modules :

- ATRT pour l'exploitation automatisée de relevés bancaires et des données téléphoniques,
- ANB pour l'analyse et la représentation visuelle des données,
- IVC pour l'identification des victimes de catastrophe et
- Mercure pour l'analyse des données téléphoniques obtenues sur réquisition.

Le module ANB a été développé à l'origine par la société anglaise i2 Limited qui appartient maintenant à IBM. D'après IBM, le module ANB est un outil d'analyse visuelle qui transforme les données en renseignement. La solution offre des fonctions innovantes : visualisations des réseaux connectés, analyse de réseaux sociaux, vues géo spatiales ou temporelles qui mettent en évidence les connexions et les tendances cachées dans les données. Ces connaissances vous permettent d'identifier et stopper plus facilement les activités criminelles, les cyber menaces et les fraudes.

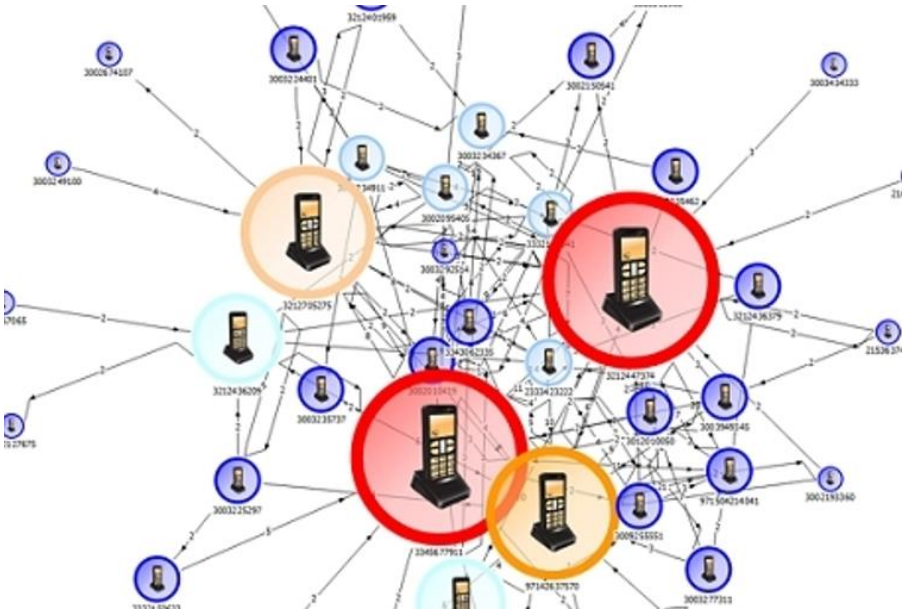


Tableau 7 : mise en relation

Ce logiciel apporte *« une vision globale de la procédure et permet de distinguer la logique qui se dessine au travers de la commission d'un fait criminel ou délictuel »* (*Didier Berger, chef du Bureau des affaires criminelles (BAC) de la gendarmerie.*)

Près de 400 analystes sont formés à ce logiciel dont l'objectif est de mettre en évidence des incohérences d'emploi du temps d'un témoin ou d'une personne mise en cause mais aussi des contradictions entre certains témoignages et des constatations effectuées par les enquêteurs.

L'analyse criminelle repose en grande partie sur la représentation visuelle par graphique, permettant d'identifier plus facilement les liens entre différents acteurs d'un réseau... ou les incohérences dans leurs déclarations pendant l'enquête. Pour chaque dossier, "on constitue une base de données avec tout ce que tout le monde dit et fait, ce volume global est ensuite transposé sur un gros schéma sur lequel on zoome en fonction de ses recherches", résumait un ancien Analyste criminel.

Dans le cas pratique, ce logiciel fait le "brakethroo" dans l'affaire nommé "Gregory".
Les forces de l'ordre ont accumulé 2000 courriers anonymes et 400 prélèvements ADN. Sans compter les nombreuses dépositions. En repassant ces données dans le logiciels *Anacrim*, il *« a relancé l'enquête sur la mort du petit Grégory en offrant un regard neuf sur la procédure car sur le fond il n'y a pas d'éléments nouveaux »*, (AFP). *« Comme les investigations récentes de police technique et scientifique sur des dizaines de scellés n'ont pu aboutir, les gendarmes ont décidé une remise à plat du dossier et Anacrim, sous sa dernière version, a permis de démontrer de nouvelles incohérences qui avaient jusque-là échappé aux enquêteurs ».*

II.6.2. COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*)

Le logiciel COMPAS de société Northpointe est un algorithme prédictif qui peut déterminer la tendance de récidive d'un criminel. Ce logiciel est utilisé par de nombreux tribunaux américains.

Aux Etats-Unis, de nombreuses juridictions locales utilisent des logiciels prédictifs pour tenter d'évaluer les risques de récidive des prévenus. Conçus comme des programmes « d'aide à la décision » pour les juges, lorsqu'ils doivent décider d'une mise en liberté sous caution ou d'une condamnation, ces programmes notent le plus souvent les prévenus sur une échelle d'un à dix, dix représentant un risque exceptionnellement fort de récidive.

Une enquête de ProPublica montre cependant que cet algorithmee est en réalité extrêmement peu efficace. En comparant les scores de risque de récidive attribués par le programme et les cas de récidive

réels – en excluant les personnes incarcérées – sur l'ensemble d'un comté pendant deux ans, les enquêteurs ont établi une série de statistiques sur l'efficacité du logiciel.

« Le score reflète de manière incroyablement erronée le risque de commission d'un crime violent : seules 20 % des personnes dont le programme estimait qu'elles commettraient un crime violent l'ont fait. Lorsqu'on prend en compte l'ensemble des crimes et délits, comme la conduite sans permis, le logiciel s'est avéré légèrement plus efficace qu'un pile ou face. Pour les personnes dont on pensait que leur récidive était probable, 61 % des personnes ont été arrêtées dans les deux années à suivre. »

En plus, le logiciel surpondère systématiquement le risque de récidive pour les Afro-Américains, qui se voient deux fois plus souvent que les Blancs attribuer un risque de récidive moyen ou important.

Plus précisément, le logiciel se base sur les réponses à une série de 137 questions, qui vont du contenu du casier judiciaire à l'adresse et aux revenus du prévenu – aucune des questions ne porte sur l'origine ethnique. Parmi ces questions figurent aussi celles sur la « moralité », le questionnaire demandant, par exemple, si le prévenu considère qu'il est normal qu'une personne affamée vole pour se nourrir.

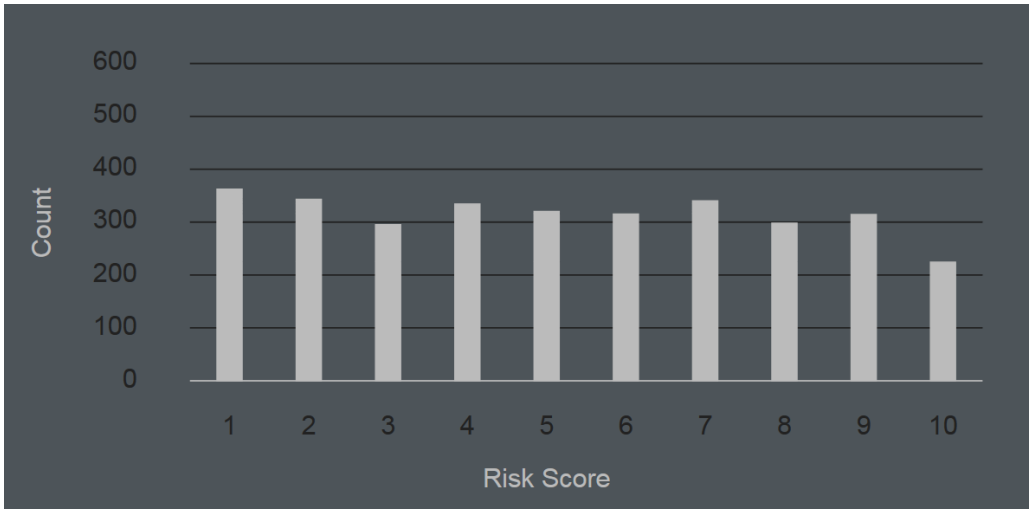


Tableau 8 : Scoring des personnes de couleur de peau blanc

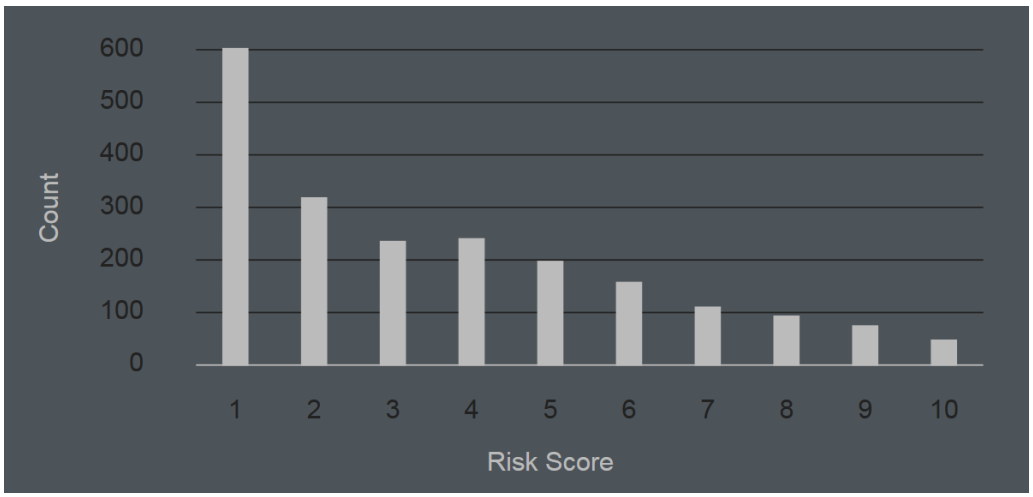


Tableau 9 : Scoring des personnes de couleur de peau noir

L'exactitude globale était fondée sur le taux auquel un défendeur avait correctement prévu de récidiver ou non. La recherche a également fait état de faux positifs, lorsqu'un prévenu est censé récidiver, mais ne le fait pas, et de faux négatifs, lorsqu'un prévenu est prédit de ne pas récidiver, mais qu'il le fait. Avec beaucoup moins d'informations que COMPAS, 7 caractéristiques comparées à 137, lorsque les résultats ont été regroupés pour déterminer la sagesse de la foule, ***les humains sans expérience présumée de justice pénale étaient exacts dans 67 %*** des cas présentés, soit la même signifiance statistique que ***la précision de 65,2 % de COMPAS***.

II.6.3. Faception

Faception, une start-up israélienne, associe aux traits d'un visage des caractéristiques comportementales, et affirme pouvoir ainsi repérer les pédophiles et terroristes. Sa méthode est toutefois contestée.

Joueurs de poker professionnels, individus au QI élevé, mais aussi pédophiles et terroristes… Autant de catégories de personnes déterminées arbitrairement par Faception et auxquelles la start-up a respectivement associé des traits caractéristiques.

Tableau 10 : capture d'écran Faception

Grâce à une technologie fondée sur la reconnaissance faciale, il lui serait possible de reconnaître une quinzaine de catégories d'individus dans 80% des cas, à partir de l'analyse de flux vidéo ou de simples photographies.

Faception se flatte également d'avoir réussi à identifier les meilleurs joueurs d'un tournoi de poker, organisé par l'un de ses partenaires : sur les cinquante participants de la compétition, deux des quatre joueurs désignés se sont effectivement rendus en finale. Sur la base de ses résultats, elle aurait d'ores et déjà conclu un contrat avec un organisme de sécurité pour identifier de potentiels terroristes.

En matière de traitement automatique des informations par les algorithmes, les biais restent encore nombreux. Ils résultent de la nécessaire intervention humaine dans leur élaboration. Une analyse menée par ProPublica, comme dans le cas de COMPAS, révélait ainsi qu'un algorithme destiné à prédire les récidives, et très utilisé dans l'univers carcéral américain, avait tendance à léser les Noirs.

En plein essor, le secteur de la reconnaissance faciale voit les innovations se succéder. Il y a près d'un an, DeepFace (Facebook) annonçait un taux de reconnaissance correct des visages de 87,25%, presque équivalent à celui des humains et depuis dépassé par FaceNet (Google), avec ses 99,63%.

II.6.4. BEWARE

Le logiciel **Beware** développé par la société **Intrado** s'est inspiré du modèle du renseignement militaire. Utilisé par la plupart des services de police américains, ce type de logiciel d'évaluation classe les personnes selon leur degré de dangerosité supposée. Un simple commentaire sur les réseaux sociaux peut conduire n'importe quel citoyen à se retrouver dans le collimateur des agences de sécurité.

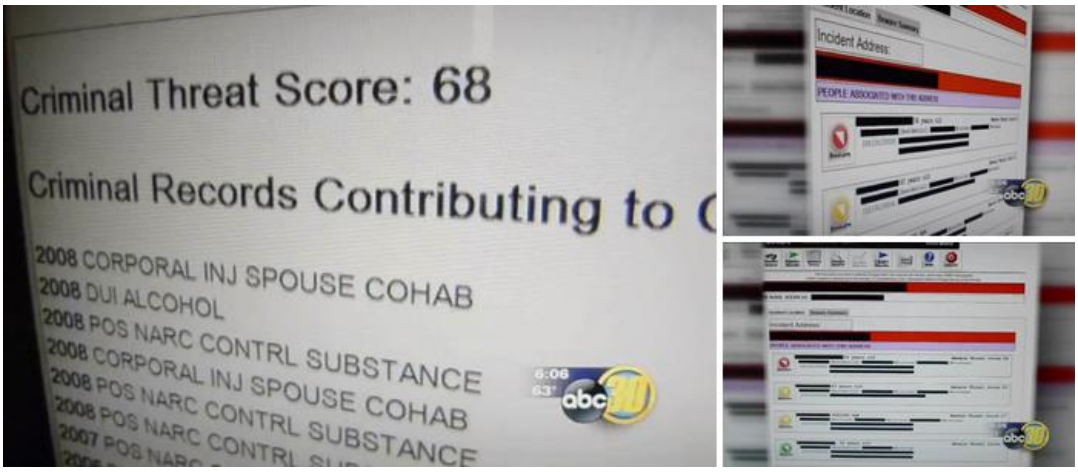


Tableau 11 : capture d'écran Beware

Beware rassemble un très grand nombre d'informations et de renseignements sur les individus, censés renseigner les forces de l'ordre sur la dangerosité potentielle de l'intervention. De moins en moins chère, ce type de technologie est de plus en plus utilisée aux Etats-Unis. Il y a deux ans, on estimait à 90% la proportion des 14.000 services de police américains ayant recours à ce type de logiciel d'évaluation. Mais elle n'est pas exempte d'erreurs et risque d'être utilisée à des fins de surveillance des citoyens innocents, (petit clin d'œil à *Minority report*) voire de « profiler » tout un chacun en vue d'évaluer la probabilité d'un comportement criminel.
Pour John Dyer, chef de la police de Fresno, l'important est de fournir autant de renseignements que possible à ses agents sur le terrain afin qu'ils puissent intervenir dans les meilleures conditions de sécurité.

L'argument sécuritaire de Beware est le premier cité par les partisans de ce type d'outils qui traitent des milliards de données tels que les rapports d'arrestations, les titres de propriété, les données cachées du Web et les achats par cartes bancaires, mais aussi les propos tenus sur les réseaux sociaux, pour associer un niveau d'alerte à trois couleurs : vert, jaune et rouge à chaque individu évalué.
La « note » peut baisser en fonction des messages mis en ligne sur les réseaux sociaux : selon le président d'Intrado, même « tout commentaire pouvant être considéré comme offensant influe sur le résultat ».

II.6.5. Deep Science : Coban

La police de l'État américain du Delaware, a commencé à déployer des caméras "intelligentes" dans ses véhicules afin d'aider les autorités à détecter un véhicule transportant un fugitif, un enfant disparu ou une personne âgée désorientée.

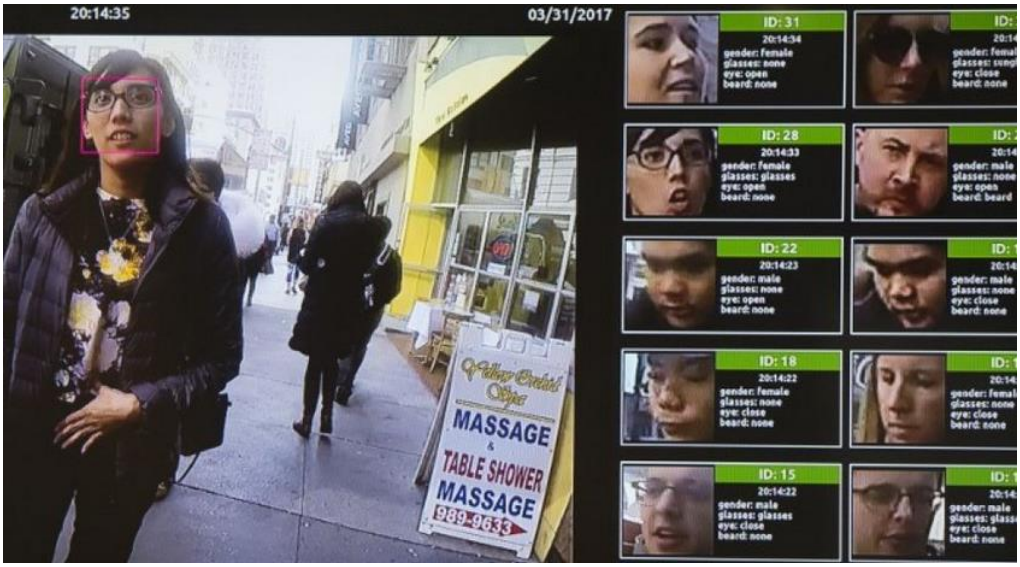


Tableau 12 : capture d'écran Coban

Le programme de reconnaissance faciale s'inscrit dans une utilisation croissante de l'alliance entre l'intelligence artificielle et la vidéo pour lutter contre la criminalité. Les vidéos seront analysées par des logiciels d'intelligence artificielle capables d'identifier les véhicules grâce à leur plaque d'immatriculation ou d'autres caractéristiques, et donner ainsi *"des yeux supplémentaires"* aux agents en patrouille, selon David Hinojosa de *Coban Technologies*, la société qui fournit ce matériel.

La start-up américaine Deep Science utilise la même technologie pour aider les commerçants à détecter en temps réel un vol à main armée, en identifiant des armes ou des agresseurs masqués, ce qui permet de déclencher des alarmes automatiques.

II.6.6. *Predpol*

Predpol est un logiciel de police prédictive fait aux Etats-Unis par la société *Predpol*. Son algorithme secret, toujours comparé aux précogs de Minority Report, c'est-à-dire à une forme de perception extra-sensorielle, semble tenir plus de la magie que de la science, comme si les mathématiques étaient miraculeuses. La société, elle, affiche partout des résultats là où sa technologie est déployée. A savoir si l'on en croit le site de *Predpol*, une petite dizaine de villes américaines pour l'instant, Los Angeles et Atlanta étant les plus importantes : une baisse de la criminalité de 10 à 30 % selon le type de crimes.

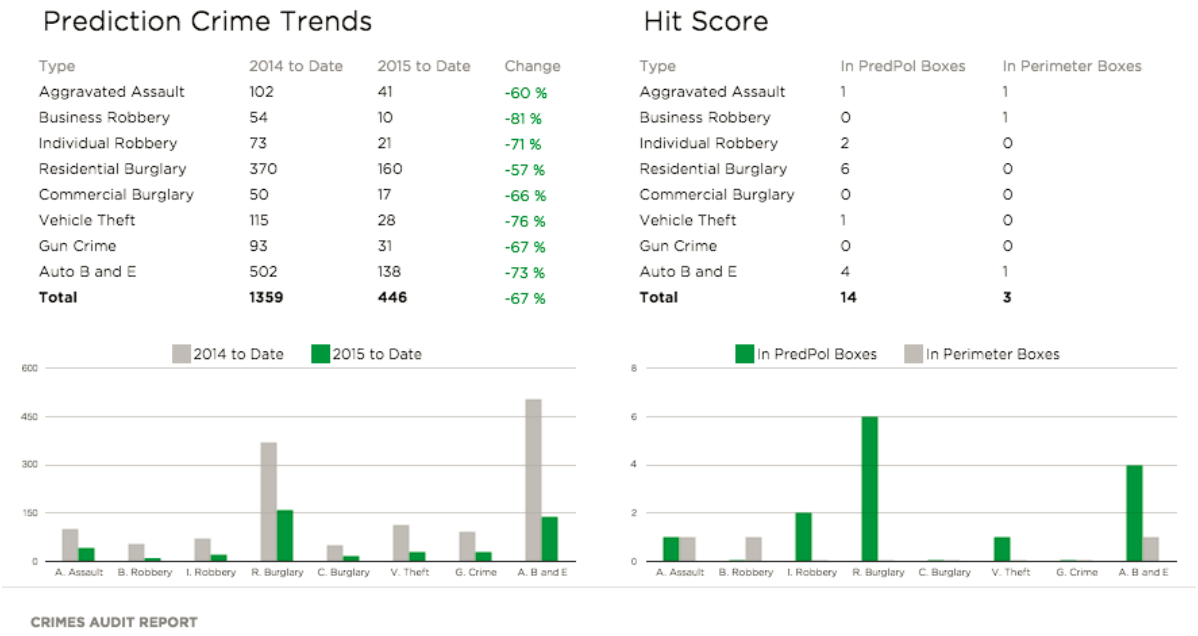
Tableau 13 : Capture d'écran de l'interface cartographique du logiciel de Predpol et sa prédiction de zones à risques

Pour mettre au point leur outil prédictif, les chercheurs se sont inspirés d'algorithmes utilisés dans la prévision des tremblements de terre. Si les sismologues peinent à prédire les secousses primaires, ils peuvent en revanche prévoir leurs répliques en théorie proches en temps et en lieu du séisme initial. Il en irait de même pour les délits : un délit a de fortes chances de se répéter dans un même quartier et se diffuse de proche en proche.

Ce phénomène de la diffusion, que les spécialistes appellent near-repeat, colle assez bien avec les résultats des enquêtes qualitatives conduites auprès des cambrioleurs. Ces derniers expliquent aux enquêteurs qu'il leur arrive régulièrement de revenir cambrioler un même logement lorsque l'effraction n'est pas compliquée et qu'ils n'ont pas pu tout emporter lors de leur premier passage. Les cambrioleurs opèrent par secteur et ils obtiennent parfois auprès de leurs réseaux des informations sur la vulnérabilité des cibles détectées lors des phases de repérage.

L'algorithme de ***Predpol*** est une version améliorée de celui mis au début des années 2000 par des chercheurs du JDICS, le Prospective Crime Mapping dit ***PROMAP***. Il consiste à modéliser les changements spatio-temporels de la victimisation à répétition sur un territoire donné afin de développer un outil opérationnel de prédiction du crime. Pour anticiper les répliques, les chercheurs s'inspirent des méthodes de lissage ordinairement utilisées en analyse spatiale pour trouver les points nodaux sur une carte. On

trouve déjà l'idée maîtresse qui fera le succès de *Predpol* : alors que les cartes des hotspots policing se contentent de répertorier les zones à risque à partir de la répartition spatiale des délits déjà commis, l'algorithme de *PROMAP* intègre, dans les éléments de paramétrage des formules, les théories criminologiques sur la contagion, notamment les résultats-clés des recherches sur la victimisation à répétition. Des recherches ont établi que le risque de victimisation se diffuse sur un rayon de 400 mètres, avec un risque plus élevé pour les maisons du même côté de la rue et sur une période de deux mois.



Prediction Crime Trends

Type	2014 to Date	2015 to Date	Change
Aggravated Assault	102	41	-60 %
Business Robbery	54	10	-81 %
Individual Robbery	73	21	-71 %
Residential Burglary	370	160	-57 %
Commercial Burglary	50	17	-66 %
Vehicle Theft	115	28	-76 %
Gun Crime	93	31	-67 %
Auto B and E	502	138	-73 %
Total	**1359**	**446**	-67 %

Hit Score

Type	In PredPol Boxes	In Perimeter Boxes
Aggravated Assault	1	1
Business Robbery	0	1
Individual Robbery	2	0
Residential Burglary	6	0
Commercial Burglary	0	0
Vehicle Theft	1	0
Gun Crime	0	0
Auto B and E	4	1
Total	**14**	**3**

Tableau 14 : capture d'écran Predpol Score

II.7. Prédiction et discrimination

II.7.1. Préjugés, la discrimination algorithmique et la stigmatisation

Les règles ou catégories utilisées dans le logiciel prédictif doivent être traduites en code informatique et la traduction elle-même peut avoir plusieurs conséquences sachant que les intentions initiales peuvent être perdues dans la traduction.

La discrimination algorithmique peut être une conséquence potentielle qui peut être le résultat de trois types de déformation de polarisation.

- Le premier type de déformation s'insinue involontairement dans l'étiquetage des exemples ou des règles qui sont codées dans l'algorithme.
- Le deuxième type peut être déformé en raison d'hypothèses déformées par la façon dont des données ont été recueillis.
- Troisièmement, sans que cela soit explicitement écrit dans le logiciel, fluage de polarisation peut se produire en raison de défauts techniques, des défauts et des bugs dans le système, ce qui peut conduire à plus de faux positifs qui répondent à certains critères.

Bien que l'exploration de données puisse hériter des préjugés des avant-décideurs ou les préjugés répandus qui persistent dans la société en général, la discrimination qui en résulte est presque toujours une propriété émergente non intentionnelle de l'utilisation de l'algorithme plutôt qu'un choix conscient par ses programmeurs, de sorte qu'il peut être particulièrement difficile d'identifier la source du problème ou de l'expliquer au tribunal.

Toutefois, cette observation ne repose pas sur des données empiriques, et on peut faire valoir qu’en particulier dans le contexte de la lutte contre le terrorisme et les politiques de lutte contre la radicalisation, cette observation n'est pas vrai, et la discrimination est un choix conscient par les programmeurs comme sous forme de profilage ethnique.

Cela met en évidence la question de la responsabilité morale des programmeurs.

I.1.1. Faux positifs/négatifs

On prétend souvent que le Big Data est fiable et que la causalité saisit à la matière parce que la taille de l'échantillon est si grand *(Mayer-Schönberger et Cukier 2013).* Cependant, l'erreur est un problème crucial dans l'application du Big Data quand on met l'accent sur la prise de décisions sur les individus et les groupes, dans le but d'identification prédictive. La fiabilité dépend aussi de la façon dont la technologie est mise en œuvre dans une organisation.

Par conséquent, ces systèmes de détections peuvent évaluer certaines personnes à être faussement identifiés (faux positifs) et les criminels et les terroristes « vrais » pour rester non identifiés (faux négatifs). Un taux élevé de faux positifs augmente le risque que certains individus et les groupes sont systématiquement ciblés de manière disproportionnée comme des criminels potentiels.

PARTIE III

III. ∞ PROPOSITION ∞

Après avoir vu différentes solutions de profilage mises en œuvre avant l'application de RGPD, en faisant une rétrospective, je vais exposer une solution possible à appliquer en France dans le cadre du profilage criminologique.

Le défi s'amplifie avec l'entrée en vigueur de RGPD. Dans un premier temps je vais parler des freins posés sur le profilage avec le nouveau cadre légal et la protection de la vie privée et des données personnelles.

Ensuite, je vais brièvement parler des problèmes des différentes sources de données et du temps de traitement des données de masses.

La solution que je propose se constitue de plusieurs briques mise en place de telle façon que le profilage puisse continuer à se faire dans le cadre de RGPD et avec un temps de traitement considérablement réduit, presque en temps réel. En fait, on peut appeler ce traitement le pré-profilage.

III.1. LE PROFILAGE AVANT LE RGPD : ENJEU DU BIG DATA

Comme le profilage est une technique de traitement de données personnelles, avec l'entrée en vigueur du Règlement européen général sur la protection des données (RGPD) le 25 mai 2018, on assiste à un changement profond des règles de traitement des données personnelles. Désormais, le profilage fait l'objet d'un encadrement strict par rapport à d'autres traitements de données.

Le profilage ne touche que les personnes physiques.

Comme vu dans la partie précédente, la notion de profilage est strictement encadrée par le RGPD.

III.1.1. Obstacle RGPD (extrait d'article 4)

"Profilage, toute forme de traitement automatisé de données à caractère personnel consistant à utiliser ces données à caractère personnel pour évaluer certains aspects personnels relatifs à une personne physique, notamment pour analyser ou prédire des éléments concernant le rendement au travail, la

situation économique, la santé, les préférences personnelles, les intérêts, la fiabilité, le comportement, la localisation ou les déplacements de cette personne physique ".

C'est une définition très large de la notion de profilage. Le profilage regroupe donc un grand nombre d'activités de domaines très différents comme par exemple le marketing comportemental ou la détection de fraudes.

Comme vu auparavant, le profilage est une technique de traitement de données personnelles spéciale et automatisée qui nous permet d'évaluer certains aspects propres à la personne concernée par le traitement des données.

En revanche, les traitements de données personnelles non automatisés ne font pas partie de la notion de profilage étant donné qu'ils ne sont pas mentionnés à l'article 4 du RGPD.

Le profilage aujourd'hui est autorisé seulement sous certaines conditions car il peut porter atteinte aux droits fondamentaux et aux libertés fondamentales des personnes concernées.

III.1.2. G29

Le G29 est un groupe de travail qui réunit les autorités de contrôle des Etats Membres de l'Union Européenne. La CNIL (Commission nationale de l'Informatique et des Libertés) en fait donc partie.

Le G29 divise la notion de profilage en deux types de profilages :

- Profilage associé à une prise de décision automatisée
- Profilage non associé à une prise de décision automatisée

Le profilage non associé à une prise de décision automatisé est autorisé. Il en est de même pour les processus de décisions qui ne sont pas automatisés mais basés sur un profilage avec une intervention humaine.

Dans ces deux cas, l'entreprise ou l'organisme souhaitant effectuer le profilage doit seulement se plier aux règles du RGPD et être conforme à ses exigences. Aucune condition supplémentaire n'est requise. Le consentement de la personne concernée n'est pas requis.

Les exigences à respecter dans le cas du profilage non associé à une prise de décision automatisé sont par exemple les suivantes : un traitement de données licite et suivant un intérêt légitime et l'information de la personne concernée.

En revanche, le texte n'est pas assez clair sur certains points et il mérite plus de précisions. Je vais m'intéresser à deux termes : légitime et raisonnable…

L'intérêt légitime doit s'évaluer attentivement. Comme pour le traitement des données, la personne concernée par le profilage doit pouvoir raisonnablement s'attendre à être la cible du profilage. Elle doit avoir pu anticiper le moment et le cadre de la collecte de ses données personnelles et elle doit pouvoir comprendre la finalité du profilage.

Le profilage associé à une prise de décision automatisé s'effectue sans intervention humaine et est susceptible d'engendrer des effets juridiques vis-à-vis des droits et liberté de la personne concernée par le profilage. Sa mise en œuvre fait donc l'objet de conditions strictes, qu'importe que les effets juridiques soient négatifs ou non.

Un profilage associé à une prise de décision automatisé avec intervention humaine est rattaché à la notion de profilage associé pleinement à une prise de décision automatisé si l'intervention humaine est insignifiante ou n'a pas d'impact sur le résultat final du profilage.

Le profilage doit absolument être accompagné de garanties permettant de maintenir le respect des droits de la personne concernée.

En principe, le profilage associé strictement à une prise de décision automatisé est interdit. Néanmoins, le RGPD prévoit des exceptions à ce principe dans son article 22.

III.1.3. Droit d'opposition au profilage basé sur une prise de décision automatisée

III.1.3.1. Article 22 du RGPD

" La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire ".

La personne concernée peut s'opposer au profilage basé sur une prise de décision automatisée.

III.1.4. Les exceptions au droit d'opposition de la personne concernée au profilage automatisé – le temps perdu

L'article 22 du RGPD prévoit trois exceptions au droit d'opposition de la personne concernée.

III.1.4.1. Extrait d'article 22 a) du RGPD

1. " [Le profilage] est nécessaire à la conclusion ou à l'exécution d'un contrat entre la personne concernée et un responsable du traitement ".

En réalité, on doit mettre en place des moyens permettant de vérifier et de justifier du caractère nécessaire de l'opération de profilage pour l'exécution ou la conclusion d'un contrat entre une personne et l'organisme ou l'entreprise. Le meilleur moyen de vérifier la nécessité d'un traitement de données est d'effectuer une analyse préalable d'impact du traitement par rapport aux finalités recherchées par l'entreprise.

III.1.5. Un droit national contraire en vigueur :

III.1.5.1. Extrait d'article 22 b) du RGPD

2. " [Le profilage] est autorisé par le droit de l'Union ou le droit de l'État membre auquel le responsable du traitement est soumis et qui prévoit également des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ".

Les Etat membres ont la liberté d'ériger une disposition légale permettant de s'opposer au droit d'opposition des personnes concernées au profilage automatisé.

En effet, les Etats membres peuvent être amenés à autoriser le profilage pour prévenir de risques de fraude ou d'évasion fiscale.

III.1.6. Le consentement explicite de la personne concernée

III.1.6.1. Extrait d'article 22 c) du RGPD

3. " [Le profilage] est fondé sur le consentement explicite de la personne concernée. ".

Si la personne concernée donne son consentement explicite au profilage basé sur une prise de décision automatisée, le profilage est autorisé.

Le consentement explicite doit avoir été donné par écrit pour éviter toute confusion. Il doit être clair et précis.

Le consentement explicite de la personne concernée ne dispense pas le responsable de traitement des données de devoir démontrer que la personne concernée a été bien informée et conseillée sur les conséquences de son consentement et du profilage automatisé. Il faut que la personne concernée puisse prendre un choix éclairé.

Le responsable doit de plus dans tous les cas prévoir à l'avance les conséquences que pourrait avoir un traitement de données sur la personne concernée et ses droits et libertés.

Enfin, il est strictement interdit d'effectuer du profilage sur des enfants.

Jusque-là, le profilage concerne beaucoup de travail effectué préalablement, des permissions, de prétraitement et autorisation. En plus, les données de masse nous n'aidons pas dans le sens où nous devons traiter tous les données, les analyser et prétraiter. Certainement, ça prend beaucoup de temps.

Pour éviter ce problème, je me pose la question de sources des données utilisées pour les analyses et le procédé avec des outils disponibles dans les services de renseignement vu auparavant.

III.2. LE PROBLEME DES SOURCES DES DONNEES

Tous les logiciels de profilage reposent sur la solution BigData, notamment les 3V. Le Big Data permet ainsi de faire passer de l'analyse reporting à l'analyse prescriptive.

Pour traiter ces données, des processus sont automatisés, ce qui nous permet de gagner en temps de traitement. Or, tous les processus automatisés sont régulés par le RGPD, comme vu auparavant.

L'information produite s'organise dans des bases de données, elles-mêmes agrégées dans des entrepôts de données (datawarehouses ou datamarts).

Ces données sont ensuite traitées sous forme de cubes décisionnels pour permettre de visualiser des indicateurs sous différentes dimensions (temporelle, géographique, catégories, segmentation, …).

Si l'utilisation des données personnelles est encadrée avec RGPD, je vais m'intéresser aux autres sources disponibles…

Comme vu dans la partie Big Data, on s'appuie sur quatre sources de données :

- Les " logs " des sites web
- Les " insights " des médias sociaux
- Les " third party data "
- L'Open data

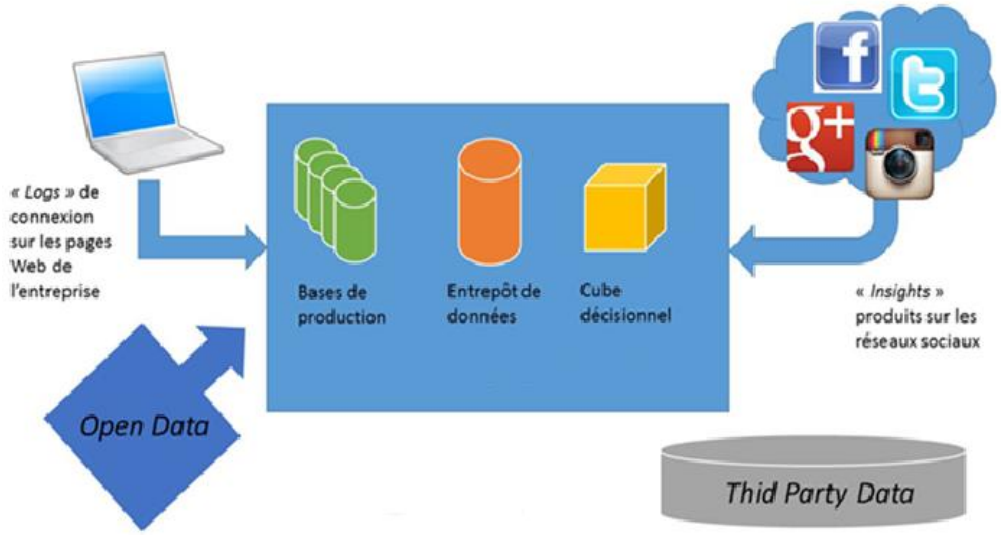


Tableau 15 : sources des données

III.2.1. Les " logs " (journaux de connexion) issus du trafic sur le site officiel de l'entreprise

Des sites internet génèrent du trafic qui est également analysé. Pour une approche plus fine, et donc plus riche en informations, on disposera des *trackers* sur les différentes pages afin de mesurer les chemins de navigation, ou encore les temps passés sur chaque page… Voire les déplacements de la souris sur l'écran.

D'autres questions intéressantes, et donc d'autres sources de données, sont les chemins pris par les visiteurs pour parvenir sur le site : moteurs de recherche, annuaires, rebonds depuis d'autres sites.

Parmi les solutions d'analyses les plus connues on peut citer : Google Analytics, Adobe Omniture, Coremetics.

III.2.2. Le contenu et les mesures de réputation issus des médias sociaux

Une approche complémentaire, mêlant méthodes quantitatives et qualitatives, consiste à recueillir les commentaires aux publications et à y appliquer des algorithmes d'analyse de sentiment.

III.2.3. La " third party" data

Les données sur les internautes (third party data) sont récoltées via des formulaires ou des cookies. Au-delà des classiques informations d'identité (sexe, âge, CSP…), il est maintenant beaucoup plus efficace de mesurer les comportements (navigation, configuration matérielle, temps passé sur les pages…).

III.2.4. L'open data

Le terme Open Data désigne des données auxquelles n'importe qui peut accéder, que tout le monde peut utiliser ou partager. Les critères essentiels de l'Open Data sont la disponibilité, la réutilisation et la distribution, et la participation universelle. Il s'agit là de la définition donnée par l'Open Knowledge Foundation en 2005.

III.2.5. Les données ouvertes en France : Open data Gouv

En France, le site data.gouv.fr de l'open data gouv permet à tout un chacun d'accéder librement aux données publiques pour les partager, les améliorer et les réutiliser. Cette plateforme officielle permet de répartir les données ouvertes dans plusieurs catégories.

Ces catégories sont l'agriculture et l'alimentation, la culture, l'économie et l'emploi, l'éducation et la recherche, l'international et l'Europe, le logement, le développement durable et l'énergie, la santé et le social, la société, et les transports, tourisme et territoires. Grâce à cette catégorisation, les utilisateurs peuvent facilement sélectionner le domaine qui les intéresse.

III.2.6. Grey data

Les données des sources grises représentent la source utilisée dans le profilage. La plupart des informations d'une organisation - plus de 80% - sont constituées de données non structurées, dont la majorité est directement contrôlée par des employés individuels sur leurs postes de travail individuels et leurs parts de fichiers. Cela signifie que ces données sont pour la plupart non gérées.

En fait, un sondage 2012 du Conseil de conformité, gouvernance et surveillance (CGOC) a révélé que dans le magasin de données d'entreprise moyen, 1% des données étaient sujettes à des litiges, 5% étaient assujetties à la réglementation et 25% avaient une valeur commerciale. A retenir - laissant 69% potentiellement sans valeur.

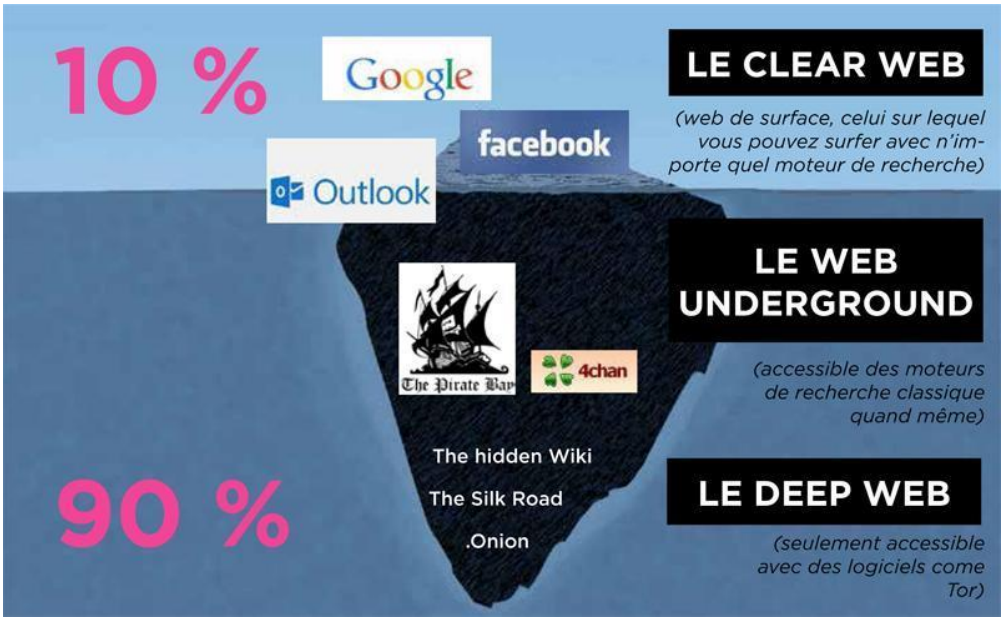


Tableau 16 : Data Iceberg

III.3. PROPOSITION DE SOLUTION D'ANALYSE DES DONNEES

III.3.1. Répondre aux RGPD et profilage

Comment utiliser des données laissées par des utilisateurs ?

Selon RGPD, une donnée à caractère personnel est toute information relative à une personne physique susceptible d'être identifiée, directement ou indirectement.

Peu importe que ces informations soient confidentielles ou publiques.

Pour que ces données ne soient plus considérées comme personnelles, elles doivent être rendues anonymes de manière à rendre impossible toute identification de la personne concernée : noms masqués, visages floutés, etc.

S'il est possible par recoupement de plusieurs informations (âge, sexe, ville, diplôme, etc.) ou par l'utilisation de moyens techniques divers, d'identifier une personne, les données sont toujours considérées comme personnelles.

Le RGPD prévoit que la police puisse utiliser des données à caractère personnel en cas de prévention et détection d'infractions pénales :

"La protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, y compris la protection contre les menaces pour la sécurité publique et la prévention de telles menaces et la libre circulation de ces données, fait l'objet d'un acte juridique spécifique de l'Union."

Bien qu'on ait la possibilité d'utilisation des données personnelles dans le profilage, comment peut-on savoir sur quelle personne orienter notre investigation sans pour autant faire un abus de pouvoir et pratiquer le traitement de profilage de masse ?

On a vu auparavant que le temps de traitement des données de masse (Big Data) dans le profilage prend beaucoup de temps.

De mon point de vue, la réponse est dans l'anonymisation des données (les définitions ont été rassemblé par *Charlotte Galichet, Avocat au Barreau de Paris,* et publié dans le magasin ARTICLE EXPERT) et la surveillance du changement de pattern avec la surveillance des connections plutôt que sur le contenu des informations.

Si l'on oriente l'investigation vers des liens qui se créent entre des personnes et non vers des données et leurs contenu, ce problème disparait.

III.3.2. Anonymisation

L'anonymisation offre une double garantie : celle de la sécurisation de l'exploitation des données personnelles et celle du respect des droits fondamentaux des personnes dont les données personnelles sont traitées.

L'appréciation du caractère irréversible de l'anonymisation, lequel offre la possibilité ou non d'identifier une personne, dépend « des moyens susceptibles d'être raisonnablement mis en œuvre, soit par le RT, soit par une autre personne ».

Au regard de la Loi Informatique et Libertés, le modèle d'anonymisation à utiliser peut-être l'anonymisation à bref délai.

Dans ce cas, le processus d'anonymisation suit immédiatement la collecte des données (quelques minutes). Toutefois, du fait de l'existence d'un temps, bien que bref, entre la collecte des données et leur anonymisation réelle, la CNIL reste compétente pour autoriser la mise en place du procédé d'anonymisation.

Dans la mesure où l'anonymisation à bref délai permet à de contourner l'application des certaines règles de la loi de 1978 (notamment en matière d'information préalable des personnes) et RGPD, la CNIL appréciera l'efficacité du procédé envisagé afin de garantir la sécurité des personnes dont les données personnelles sont traitées.

III.3.2.1. Les techniques d'anonymisation

Deux grandes familles de techniques visent à altérer le lien entre les données personnelles collectées et l'individu auxquelles elles se rapportent : la randomisation et la généralisation.

Dans la mesure où nous souhaitons anonymiser des données et si besoin retrouver la vraie identité, le modèle conventionnel n'est pas adapté.

III.3.2.2. Les techniques de pseudonymisation

L'article 4 du RGPD définit la pseudonymisation de la manière suivante : " (…) on entend par pseudonymisation : le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable. "

La pseudonymisation permet toujours d'identifier un individu grâce à ses données personnelles car elle consiste simplement à remplacer un attribut par un autre au sein d'un enregistrement. En effet, le considérant 26 du RGPD rappelle que « Les données à caractère personnel qui ont fait l'objet d'une pseudonymisation et qui pourraient être attribuées à une personne physique par le recours à des informations supplémentaires devraient être considérées comme des informations concernant une personne physique identifiable ».

Le Règlement communautaire évoque beaucoup la notion de pseudonymisation comme technique de respect du principe du Privacy By Design et de la minimisation des données (article 25 du RGPD notamment).

Les techniques de pseudonymisation sont nombreuses et offrent des garanties de sécurité variées, d'autant plus que les erreurs dans leur mise en œuvre sont courantes.

III.3.2.3. Système cryptographique à clé secrète :

Dans le cas d'un système cryptographique à clé secrète, le détenteur de la clé peut aisément ré-identifier chaque personne concernée en décryptant l'ensemble des données, puisque les données à caractère personnel y figurent toujours, quoique sous une forme cryptée. En supposant qu'un système cryptographique conforme à l'état de la technique a été appliqué, le décryptage ne serait possible qu'à condition de connaître la clé.

III.3.2.3.1. Fonction de hachage :

La fonction de hachage renvoie un résultat de taille fixe, quelle que soit la taille de l'entrée encodée (l'entrée peut être un attribut unique ou un ensemble d'attributs). Evidemment, le risque consiste en la découverte de la fourchette dans laquelle se situent les valeurs. Afin de réduire ce risque, la fonction de hachage avec salage (où une valeur aléatoire, appelée « sel », est ajoutée à l'attribut qui fait l'objet du hachage) permet de réduire la probabilité de reconstituer la valeur d'entrée.

III.3.2.3.2. Fonction de hachage par clé, avec clé enregistrée :

Il s'agit d'une fonction de hachage particulière qui utilise une clé secrète comme entrée supplémentaire (à la différence d'une fonction de hachage avec salage, où le « sel » n'est généralement pas secret). Un responsable de traitement des données peut ré exécuter la fonction sur l'attribut en se servant de la clé secrète, mais il est beaucoup plus difficile pour un attaquant de ré exécuter la fonction sans connaître la clé car le nombre de possibilités à tester est suffisamment grand pour rendre la tâche impraticable.

III.3.2.3.3. Chiffrement déterministe

Chiffrement déterministe ou fonction de hachage par clé avec suppression de la clé :

Cette technique équivaut à sélectionner un nombre aléatoire comme pseudonyme pour chaque attribut de la base de données et à supprimer ensuite la table de correspondances. En supposant qu'un algorithme conforme à l'état de la technique soit appliqué, il sera difficile pour un attaquant, en termes de puissance de calcul requise, de décrypter ou de ré exécuter la fonction, car cela supposerait d'essayer chaque clé possible, puisque la clé n'est pas disponible.

III.3.2.3.4. Chiffrement a trois niveaux

On part du principe que des données sont pseudonymies a l'entré et cryptées avec les trois clés différentes. Les informations récupérées sont librement utilisables car l'identité de la personne n'est pas mise en évidence à l'aide de l'anonymisation.

Pour s'assurer le non-abus et de respect de la vie privé, les trois clés nécessaires pour retrouver l'identité de la personne sont donné aux autorités suivantes :

Une clé est délivrée à l'autorité juridique, une au parlement et une a la police.

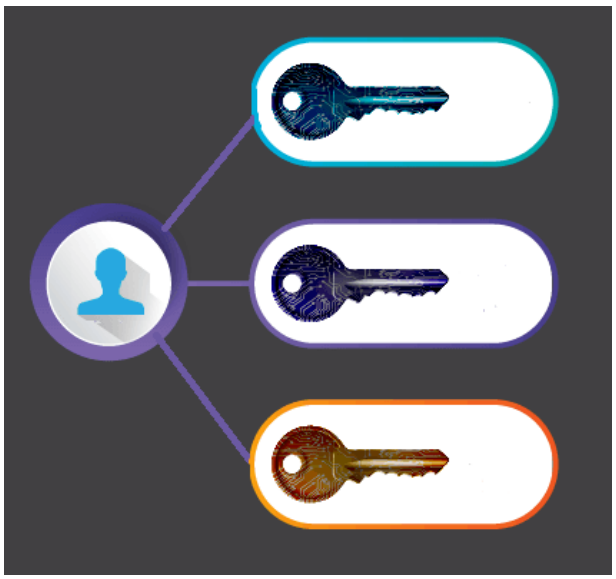


Tableau 17 : chiffrement a trois niveau

Pour lancer une opération de renseignement ciblée, on a besoin de connaitre l'identité d'une personne. Pour y parvenir, les trois clés sont nécessaires. Automatiquement, des trois instances doivent être d'accord entre elles pour nous délivrer les clés nous permettant de faire la corrélation.

Si notre traitement révèle qu'un individu est potentiellement dangereux, à l'aide des trois clés et avec l'accord des trois parties on peut retrouver l'identité de la personne visée.

Ce système nous permettra de traiter des données anonymes donc de contourner le problème de RGPD et de gagner un temps considérable concernant les autorisations nécessaires pour le profilage.

III.4. PROPOSITION DE REDUCTION DU TEMPS DE TRAITEMENT

Le temps de traitement des données structurés comme non structurés est considérable. Une des solutions à privilégier sera de se baser plutôt sur des métadonnées que des contenus de données. En plus, en utilisant des métadonnées, le processus d'anonymisation des données est fait en amont.

III.4.1. Métadonnées

Les métadonnées servent à décrire ou à définir une autre donnée (photo texte son ou vidéo). Elles sont définies dans le cadre du modèle ressource description Framework (RDF).

La première métadonnée contenue dans un fichier numérique est son nom de fichier. Mais il est possible de compléter ces informations avec le titre, les mots-clés, les informations de droits d'auteur, le nom de l'auteur etc.

Ci-dessous, à titre d'exemple, je vais m'intéresser essentiellement aux métadonnées qui concernent les photos.

En photo on distingue deux types de métadonnées : les EXIF et les IPTC.

III.4.1.1. Où sont stockées les métadonnées ?

Pour des ressources non digitales (livre, disque, ou objet de musée) les métadonnées sont évidemment externes à ses ressources. Cette externalisation des métadonnées est notamment utilisée des systèmes documentaires (bibliothèque, médiathèque).

Il faut faire attention à toujours garder associer l'objet et ses métadonnées.

Lorsqu'il s'agit de ressources digitales (photo, MP3, documents MS Office ou OpenOffice etc.) les métadonnées peuvent être internes à ces ressources.

Cela simplifie le classement, mais aussi la recherche de ces documents.

Le balisage d'une ressource informatique consiste à enregistrer les métadonnées utiles à la description du document utile notamment pour les images (EXIF et IPTC) pour les fichiers MP3 (chmps ID3) les autres fichiers multimédias.

Une photo (ou autres types de documents) ainsi balisée transporte avec elle ses propres métadonnées il n'y a donc plus risque de perte parce que celle-ci est téléchargée, copiée ou répliquée.

Le balisage des documents numériques présente des limitations importantes :

- Il ne peut se substituer à la description à l'aide de métadonnées stockées dans une base de données.
- Tous les logiciels ne sont pas capables de lire ou de préserver les métadonnées incluses dans un fichier numérique. Ainsi il arrive avec certains logiciels (généralement gratuit) que les métadonnées disparaissent lorsqu'on modifie le fichier.

III.4.1.2. Les métadonnées EXIF :

Exif est une abréviation de ***EX****changeable* ***I****mage* ***F****ile*. ces métadonnées sont automatiquement enregistrées dans les documents elles définissent les informations d'ordre technique.

Ce format de métadonnées n'est pas standardisé, il est toutefois utilisé par de nombreux constructeurs d'appareils photographiques numériques.

Un éditeur de métadonnées EXIF est un non-sens, dans la mesure où ces métadonnées sont automatiquement enregistrées par le boîtier qui réalise les photos.

III.4.1.3. Les métadonnées IPTC :

L' IPTC (International Press and Telecommunications Council) est une organisation internationale créée en 1965 pour développer et promouvoir des standards d'échange de données à destination de la presse.

Le sous-ensemble de ce modèle appelé *Application record N° 2* a servi de base en 1994 à la société *Adobe* pour définir dans son logiciel *Photoshop* les informations associées à une image. C'est ce sous-ensemble qui est communément appelé *métadonnées* (ou *champs* ou informations ou *en-têtes* ou *headers*) IPTC.

Ces métadonnées peuvent être très variées selon le type de document.

Sur des photos c'est encore pire, on peut connaitre des réglages très précis comme le modèle d'appareil photo, la sensibilité ISO, des coordonnées GPS, la date, si le flash a été utilisé, éventuellement le nom du logiciel de retouche utilisé, etc.

Il faut s'en douter, ces informations peuvent être très utiles dans le domaine de l'analyse du comportement.

Exemple avec une simple photo, l'argument **-lang fr** est optionnel, c'est simplement pour avoir la sortie en français :

```
root@bt/# exiftool -lang fr voiture.jpg
Version ExifTool                : 8.60
Nom de fichier                  : voiture.jpg
Dossier                         : /tmp
Taille du fichier               : 176 kB
Date/heure de modification du fichier: 2018:12:22 03:54:08+02:00
File Permissions                : rw-r--r--
Type de fichier                 : JPEG
Type MIME                       : image/jpeg
Indicateur d'ordre des octets Exif: Little-endian (Intel, II)
Description d'image             :
Fabricant                       : SONY
Modèle d'appareil photo         : DSC-H3
Orientation de l'image          : 0° (haut/gauche)
Résolution d'image horizontale  : 72
Résolution d'image verticale    : 72
Unité de résolution en X et Y   : Pouce
Date de modification de fichier : 2018:12:24 12:65:49
Positionnement Y et C           : Côte à côte
Temps de pose                   : 1/4
Nombre F                        : 3.5
Programme d'exposition          : Programme normal
Sensibilité ISO                 : 640
Version Exif                    : 0221
Date de la création des données originales: 2018:12:24 12:65:49
Date de la création des données numériques: 2018:12:24 12:65:49
Signification de chaque composante: Y, Cb, Cr, -
Mode de compression d'image     : 4
Décalage d'exposition           : 0
Ouverture maximale de l'objectif: 3.5
Mode de mesure                  : Multizone
Source de lumière               : Inconnue
Flash                           : Flash non déclenché, mode flash forcé
Focale de l'objectif            : 6.3 mm
Color Reproduction              : Vivid
Macro                           : Activé
Mode mise au point              : AF-S (prise de vue unique)
```

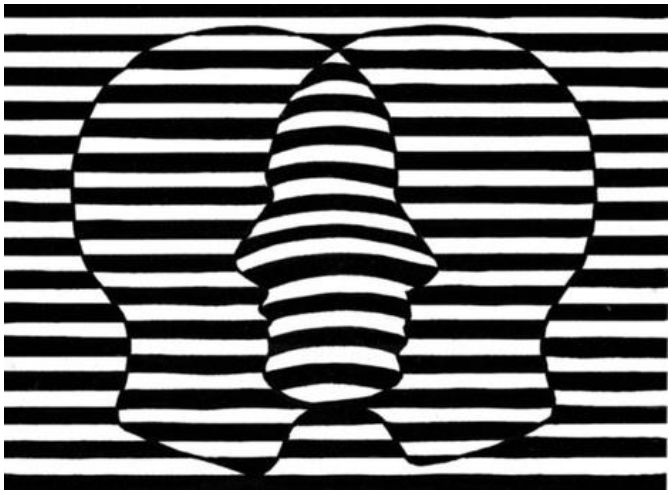
```
Mode AF                          : Multi AF
AF Illuminator                   : Auto
Qualité                          : Normale
Flash Level                      : Unknown (2)
Release Mode                     : Burst
Numéro de Séquence               : 1
Anti-Blur                        : On (Shooting)
Réduct. bruit longue expo.       : Arrêt
Version Flashpix supportée       : 0100
Espace colorimétrique            : sRVB
Largeur d'image                  : 640
Hauteur d'image                  : 480
Identification d'interopérabilité: R98: fichier de base DCF (sRGB)
Version d'interopérabilité       : 0100
Source du fichier                : Appareil photo numérique
Type de scène                    : Image photographiée directement
Traitement d'image personnalisé  : Traitement normal
Mode d'exposition                : Exposition automatique
Balance des blancs               : Equilibrage des blancs automatique
Type de capture de scène         : Standard
Contraste                        : Normale
Saturation                       : Normale
Accentuation                     : Dure
PrintIM Version                  : 0300
Schéma de compression            : JPEG (ancien style)
Thumbnail Offset                 : 9384
Thumbnail Length                 : 3004
Largeur d'image                  : 640
Hauteur d'image                  : 480
Procédé de codage                : Baseline DCT, codage Huffman
Nombre de bits par échantillon   : 8
Composants colorimétriques       : 3
Rapport de sous-échantillonnage Y à C: YCbCr4:2:0 (2 2)
Ouverture                        : 3.5
Taille de l'Image                : 640x480
Temps de pose                    : 1/4
Vignette                         : (Binary data 3004 bytes, use  -b option to
extract)
Focale de l'objectif             : 6.3 mm
Luminosite                       : 2.9
```

III.5. PROPOSITION DE METHODE D'ANALYSE

III.5.1. Pattern et profil

Pour faire l'analyse des données des masses automatisées et sans que les données utilisées nous amènent directement vers une personne identifiable, nous utilisons des données anonymes et notre recherche peut se baser sur le pattern commun.

Le terme de pattern est généralement utilisé pour décrire un ensemble de cas. Le terme de profil quant à lui est principalement exploité pour décrire un auteur ou un groupe d'auteurs. Ces nuances terminologiques apparaissent comme le résultat de la décomposition du domaine en deux communautés : les analystes du crime et les analystes du criminel. De manière plus générale, ces deux termes décrivent toutefois des concepts différents.



Le terme de pattern a notamment un double sens qu'il s'agira de ne pas confondre. Un pattern décrit un schéma récurrent ou une structure particulière. Il sera toutefois également utilisé pour décrire une solution adaptée pour un problème récurrent de conception.

En criminologie, la théorie des patterns *(Brantingham & Brantingham, 1993)* décrit les régularités des activités criminelles : les crimes sont regroupés en pattern, les décisions prises par les criminels suivent des patterns, et des patterns décrivent des processus mis en œuvre pour commettre les crimes. La théorie des patterns fait partie d'un ensemble de théories regroupées dans le domaine de la criminologie environnementale. La théorie des choix rationnels et la théorie des activités routinières en font partie *(Felson & Clarke, 1998).*

Si les crimes et les criminels suivent des schémas récurrents, il doit être possible de les détecter, formaliser et représenter. Ces patterns pouvant alors servir de base pour le développement d'évaluations opérationnelles ou stratégiques.

En matière d'analyse du crime, un pattern décrit un ensemble de cas partageant des caractéristiques communes, telles que le type de délit, le mode opératoire, le type de cible, une même zone géographique, etc.

III.5.2. Méthode inverse : Surveillance de changement de pattern et des relations

Si on connait un pattern pour un type de crime, ça sera plus facile de suivre le changement où sortie d'un pattern dit "normal" vers un pattern a risque.

Contrairement à l'analyse du comportement où on suit et analyse le comportement d'un individu, l'idée d'inclure et de se baser sur ses relations peut sembler plus pertinente.

En même temps, un pattern est basé sur le comportement et le changement de comportement des individus. Pour déterminer un pattern, il est important de s'intéresser aux comportements prévisibles qui permettent d'établir un schéma type.

Suivre les relations des personnes et des cercles d'intérêts semble être une méthode qui palie le RGPD et réduit le temps de traitement et notamment de décision.

Le temps de traitement est réduit dans la mesure où l'on surveille le pattern commun d'un groupe de personnes et non plusieurs personnes appartenant à un groupe. Si on ajoute la notion de métadonnées plutôt que des données à traiter, on peut constater que le temps de traitement ne sera pas le même.

Partons de principe que des données sont traité à la source et qu'on traite des relations trouvées dans des métadonnées. Ça nous permet d'avoir le résultat presque en temps réel, car la quantité des données a traité a considérablement réduit. On obtient le résultat dit "just in time".

III.5.2.1. Pattern

Pour surveiller le changement de pattern, on peut se référer aux différents types d'analyse selon le type des données.

Globalement, deux dimensions sont décrites de manière récurrente et semblent dominantes :

- La composante temporelle
- La composante spatiale

Ces deux dimensions ne sont toutefois pas suffisantes pour décrire l'ensemble des questionnements impliqués lors de l'analyse de données de criminalité.

Une troisième dimension principale est définie :

- La composante relationnelle

De manière générale, ces trois dimensions d'analyse semblent être pertinentes et fondamentales pour traiter et interpréter les phénomènes de criminalité *(Oatley et al., 2005) (Ribaux, 2008).*

Bien qu'une relation puisse être de nature temporelle ou spatiale, il est nécessaire de définir une dimension au sein de laquelle d'autres formes de relations se distinguent. Adrienko et Adrienko définissent cette troisième dimension sous le terme de population : un groupe d'entités pouvant être de diverses natures *(Andrienko & Andrienko, 2005).*

La notion d'entité est récurrente dans l'ensemble des définitions de cette dimension. Cette décomposition en trois dimensions est d'ailleurs également utilisée pour décrire la manière dont l'humain mémorise les connaissances selon trois structures :

- Savoir-quoi,
- Savoir-quand,
- Savoir-où

Etudier des dimensions d'analyse nous permet d'orienter notre analyse là où des autres outils d'analyse parviennent à la fin de leur analyse : de déterminer des relations entre des entités.

Les questions d'analyse semblent pouvoir se décomposer selon quatre dimensions principales. C'est-à-dire quatre composantes dans lesquelles la variabilité est observée et qui sont récurrentes dans les questions d'analyse.

III.5.2.1.1. La dimension temporelle

Elle couvre les questions liées à l'analyse d'une distribution temporelle. Le temps est la composante principale de l'analyse. Les réponses aux questions posées doivent se chercher dans la variabilité temporelle : quand, sur quelles périodes, à quelle fréquence, selon quelle régularité temporelle, etc

III.5.2.1.2. La dimension spatiale

Elle couvre les questions liées à l'analyse d'une distribution géographique. L'espace est la composante principale de l'analyse. Les réponses aux questions posées doivent se chercher dans la variabilité spatiale : où, dans quelle région, selon quel parcours, quelle étendue géographique, etc.

III.5.2.1.3. La dimension relationnelle

Elle couvre les questions liées à l'analyse des relations entre les entités pertinentes, telles que les événements, les personnes, les objets et les traces. La relation est la composante principale de l'analyse. Les réponses aux questions posées doivent se chercher dans la variabilité relationnelle : qui, quoi, entre qui, entre quoi, avec quoi, etc.

III.5.2.1.4. La dimension quantitative

Elle couvre les questions liées au dénombrement. Le nombre est la composante principale de l'analyse. Les réponses aux questions posées doivent se chercher dans la variabilité quantitative : combien de, selon quelle proportion, etc.

Pour chaque dimension, des représentations spécifiques sont exploitées. Cette décomposition selon quatre formes de questionnement principales permet de classifier et étudier les diverses formes de représentation exploitables en analyse criminelle.

Contrairement à l'ANACRIM où la dimension relationnelle a été établie et qui a permis de retrouver le coupable, cette méthode est peau adaptée pour la détection de radicalisation et passage à l'acte.

III.5.2.2. La dimension relationnelle

L'analyse relationnelle consiste en détection, compilation et en l'interprétation des informations afin d'identifier la présence de relations entre des entités d'intérêts (p. ex. des personnes, des événements, des véhicules, des sociétés, etc.) Elles permettent de rechercher et de poser des propositions telles que, quelle est la source de cette trace, qui a téléphoné avec qui, etc. Lorsque beaucoup de relations sont en jeu dans une affaire, il devient évidemment utile de les visualiser.

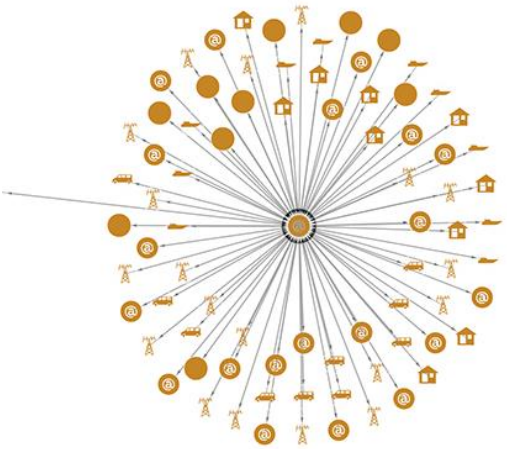


Tableau 18 : relation

Le terme d'analyse relationnelle est généralement exploité en analyse criminelle pour décrire l'exploitation de représentations relationnelles à des fins d'analyse. Le terme de visualisation relationnelle semble plus adapté dans ce sens.

Partons de principe que le pattern est connu pour une personne dit "standard où normale" et une personne dite "radicalisée où prête à agir".

A partir de ce point, sans s'intéresser aux identités des personnes, on peut surveiller le changement de pattern dans un, où l'autre sens.

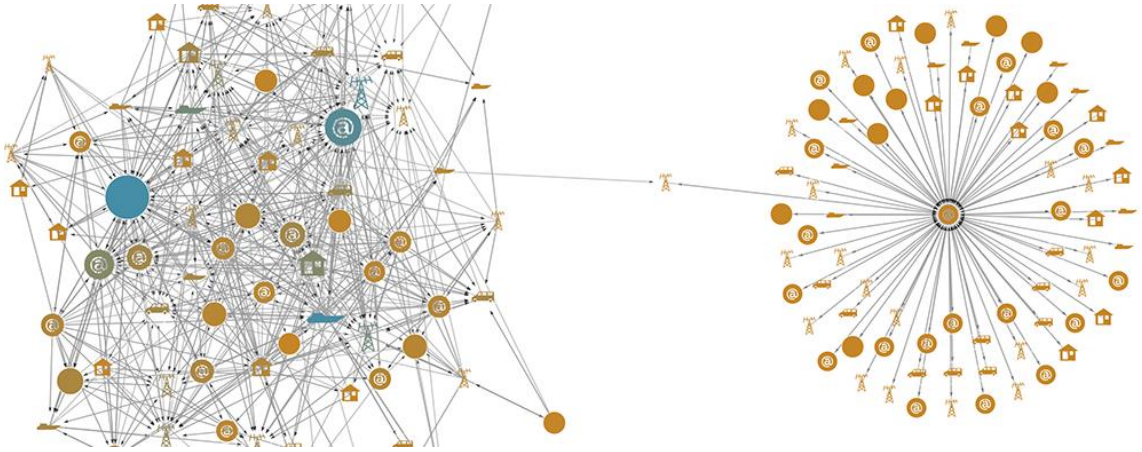


Tableau 19 : relation - corrélation

Une personne peut avoir plusieurs relations différentes au sains d'un groupe, mais ces relations restent dans le groupe. Dès que des relations sortent d'un groupe, là on peut parler de changement de pattern.

III.5.3. Comment peut-on accélérer et améliorer le processus ?

III.5.3.1. Relations et pattern

Selon l'analyse de journal "Perspectives on Terrorism", dans chaque attentat, quatre patterns communs ressort. Tous les criminels avaient communiqué avec leurs téléphones portables, ils ont utilisé leur ordinateur, ils étaient connus de la police pour un délit quelconque, et ils ont suivi la voix de radicalisation.

Tableau 20 : pattern commun

Comme expliqué auparavant, la police avait déjà ces données, mais il est impossible de faire le rapprochement entre ces données à temps. Vu que la quantité des données est assez conséquente, le temps d'analyse des informations est long et peut aller jusqu'à plusieurs mois. Les analystes se trouvent saturés avec des choix multiples et le débordement cognitif.

Même si on est en possession des toutes les informations, le rapprochement entre les données reste le point noir.

Pour accélérer ce processus, comme vu dans le paragraphe précédent, on va s'intéresser aux liaisons entre des individus et non aux contenus de leurs échanges.

En utilisant des données ouvertes provenant de Facebook, Twitter, blogues, des sites internet, dark web, des moyens de communication comme Télégramme et autres, on peut tirer de nombreuses conclusions.

Traitement et analyse des métadonnées nous permettent de ne pas rentrer dans l'analyse profonde mais de créer des "réseau sociaux" de chaque personne et de visualiser leurs relations avec d'autres réseaux.

Partons de ce principe, l'analyse peut se faire presque en temps réel car la quantité des données à analyser a considérablement diminué.

L'importance est mise sur les relations : qui communique avec qui, quand et comment, où et a quelle fréquence.

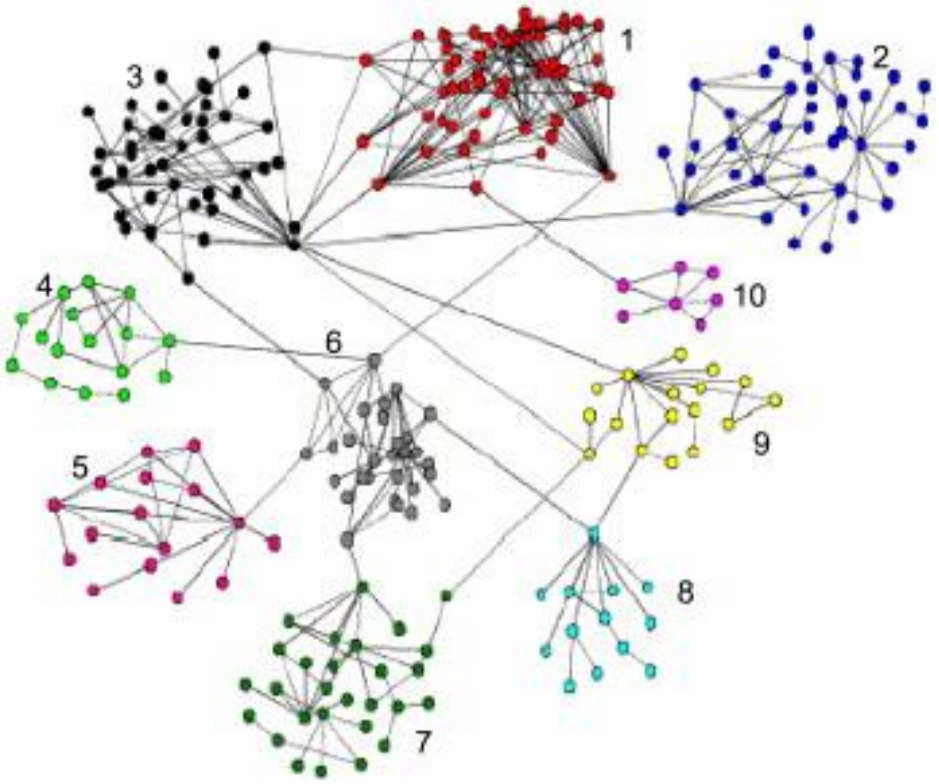


Tableau 21 : mise en relation des cercles individuels

A partir de ce moment-là, on peut arriver à cartographier des relations de réseaux. Une fois fait, on peut on peut visualiser le maillage fait et en appliquant des filtres on arrivera à isoler les cas à traiter en priorité en fonction des leurs relations, fréquence, et changement de pattern habituel.

Comme les informations récoltées des sources différentes sont pseudonymies, le traitement automatisé peut s'appliquer sans craindre des conséquences instaurées par RGPD.

Avec cette approche, la quantité des données stockées donc traitées est considérablement réduite.

En appliquant des algorithmes de Machine Learning, théorie des graphs, algorithmes génétiques et probabilité on peut arriver à visualiser rapidement des relations et tendances entre des relations.

Egalement, avec la méthode de maillage, on pourrait remonter jusqu'à ceux qui commandent, endoctrinent où embrigadent plus facilement.

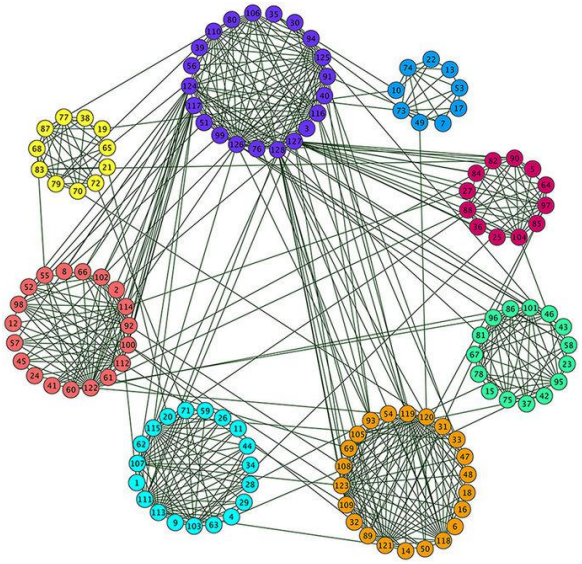


Tableau 22 : maillage relationnel

Avec l'aide d'une présentation graphique on pourrait facilement identifier ces individus. En présentation graphique, des acteurs se présentent sous forme de cercles avec des connexions. Plus le réseau compte des relations, plus la personne est dense.

On peut s'intéresser au degré de proximité avec des chefs de réseaux et aux relations avec d'autres membres.

Pour faire la présentation visuelle de réseau, on va s'intéresser aux méthodes de visualisation.

III.5.4. Visualiser des données dans l'analyse criminelle

Les méthodes de visualisation sont exploitées dans de nombreux domaines et concentrent diverses communautés de recherche. Ici la visualisation est brièvement présentée afin de compléter et préciser les rôles de la visualisation dans le processus de l'analyse criminelle.

La visualisation est l'un des supports externes qui a pour but d'augmenter des capacités cognitives. Son utilisation automatique dans le processus de profilage est nécessaire. Si on est obligé de se passer des décisions automatisées, cette manière de nous proposer des solutions restreintes est le bon choix.

Globalement, la visualisation peut donc être exploitée pour quatre objectifs :

- Mémoriser :

Afin d'effectuer des comparaisons, la visualisation permet de faciliter la mémorisation en regroupant un grand nombre de données dans l'espace visuel. Un objectif de synthétisation des informations peut également être atteint à l'aide de représentations.

- Explorer :

L'analyse exploratoire regroupe l'ensemble des tâches de recherche et d'identification de régularités ou d'anomalies particulières dans un ensemble de données. L'objectif est de découvrir de nouvelles connaissances.

- Confirmer :

L'analyse confirmatoire vise à tester les hypothèses définies. Elle sert de base à la prise de décision. L'objectif est de détecter les éléments confirmant ou non les hypothèses élaborées.

- Communiquer :

Afin de transmettre une connaissance pour faciliter la compréhension d'une problématique ou communiquer un message, convaincre.

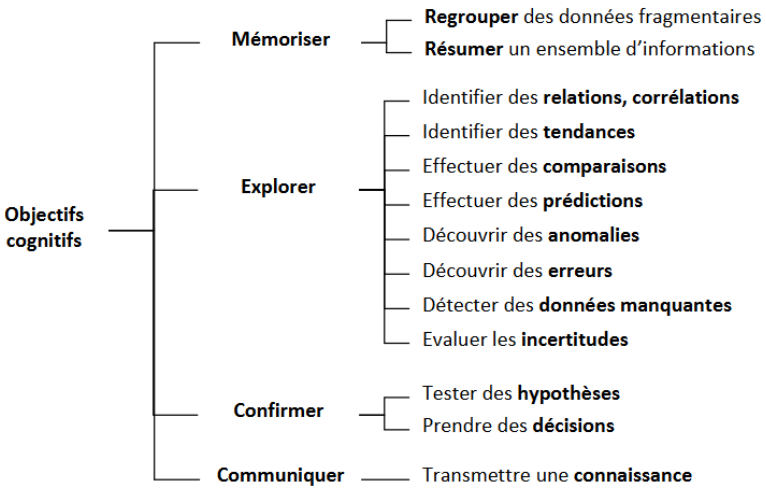


Tableau 23 : objectifs cognitifs

III.6. MISE EN PRATIQUE DES SOLUTIONS PROPOSEES

Comme expliqué auparavant, on va se baser sur des solutions proposées et les appliquer dans un cas pratique.

On suppose que le pattern a été établi. On commence par le traitement des métadonnées des sources ouvertes. Comme sources ouvertes on a pris des données de Facebook, twitter, des blogs, etc.

Un exemple des métadonnées Twitter peut être comme suivant :

```
<title>Dictionnaire du web - Les mots du webmarketing</title>
<meta name="description" content="Le dictionnaire du web vous explique Intern
<meta name="robots" content="noodp"/>
<link rel="canonical" href="http://www.dictionnaireduweb.com/" />
<link rel="publisher" href="http://plus.google.com/115231113632770601503"/>
<meta property="og:locale" content="fr_FR" />
<meta property="og:type" content="website" />
<meta property="og:title" content="Dictionnaire du web - Les mots du webmarke
<meta property="og:description" content="Le dictionnaire du web vous explique
<meta property="og:url" content="http://www.dictionnaireduweb.com/" />
<meta property="og:site_name" content="Dictionnaire du Web" />
<meta property="og:image" content="http://www.dictionnaireduweb.com/wp-conten
<meta name="twitter:card" content="summary_large_image"/>
<meta name="twitter:description" content="Le dictionnaire du web vous expliqu
<meta name="twitter:title" content="Dictionnaire du web - Les mots du webmark
<meta name="twitter:site" content="@dictionnaireweb"/>
<meta name="twitter:image" content="http://www.dictionnaireduweb.com/wp-conte
<meta name="twitter:creator" content="@dictionnaireweb"/>
```

A partir de ce moment-là, des données sont anonymisées et le traitement des relations et corrélations peut démarrer.

Toutes les métadonnées récupérées sont traitées. Suivant des filtres préétablis, si on ne détecte pas une déviation de pattern, on ne s'intéresse pas aux contenus des données.

Les relations d'une personne ressemblent à un réseau social personnel :

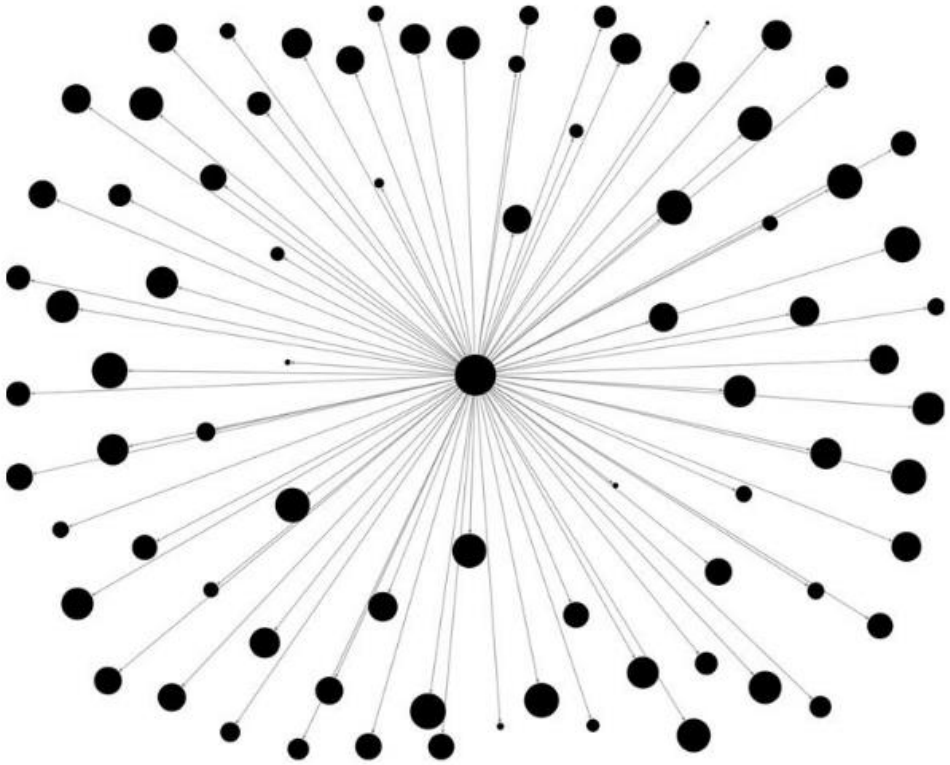


Tableau 24 : réseau social personnel

A partir de ce moment-là, on peut visualiser des relations et les suivre en temps réel. La surveillance des relations plutôt que des contenus de messages échangés nous permet de réduire considérablement du temps nécessaire pour détecter une anomalie dont la déviation de pattern.

Les relations sont toujours présentées sous forme de réseaux de connaissances individuels mis en corrélation entre eux.

Si, par exemple, on détecte qu'une personne change son comportement social, à titre d'exemple change son habitude sur internet, essaie de contacter des personnes connues pour des faits criminels, ou bien publie des informations dans des blogs où "like" des publications à caractère criminel selon des filtres prédéfinis, on découvre que cette personne représente une cible potentielle et on peut s'intéresser plus en détail a son activité.

Si pas la suite des investigations automatisés on arrive à détecter le niveau élevé de danger potentiel, lever le secret de son identité reste tout de même entre les mains des trois détenteurs de clé de cryptage.

Le modèle de détection de comportement :

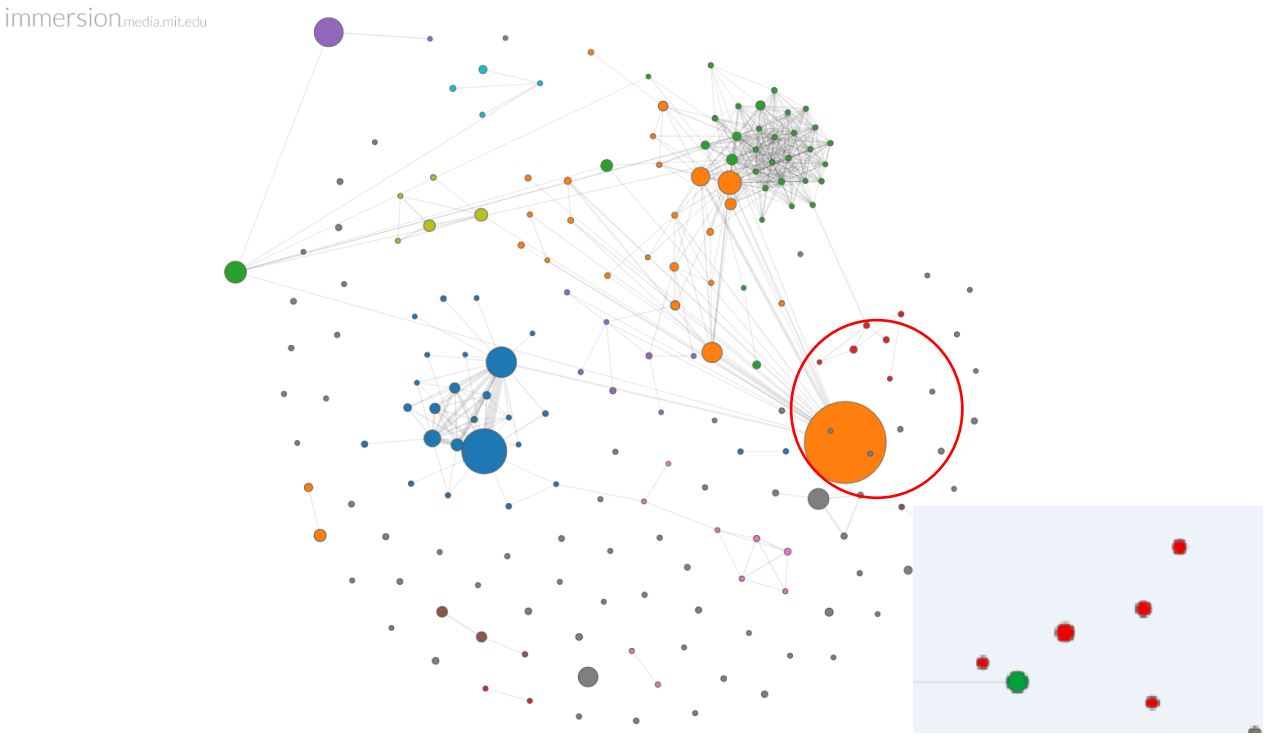


Tableau 25 : modèle de détection de changement de comportement

Le rouge représente le cercle de connexion radicalisé.

Le cercle violet représente la personne qui dévie de pattern. Elle contacte via le réseau vert la personne qui est en contact avec le réseau plus dense, vert clair, et qui est en contact avec des réseaux radicalisés, représentés par la couleur rouge.

A partir de ce moment-là, on peut vérifier si le contact a été fait au hasard ou bien, si le contact a été fait avec une intention (il se multiplie).

La quantité des données à traiter ont été minimalisés, le traitement automatisé, la présomption d'innocence également garantie car on ne porte pas de jugement sur des relations. Le traitement est basé sur des règles de type d'arbre de décision.

IV. CONCLUSION

Après plusieurs mois de recherche et réflexion au sujet de profilage criminel, j'ai réussi à m'empreigne du modèle actuel d'investigation criminelle et des techniques utilisées dans ce domaine. Pour permettre aux lecteurs de comprendre le fonctionnement et mécanisme de profilage, j'ai structuré ce mémoire de telle façon qu'avant de commencer de s'intéresser aux logiciels et au modèle de profilage, on se pose la question : pourquoi avons-nous besoin de tout ça et comment sommes-nous arrivés à faire naitre le profilage ?

Dans la première partie j'ai voulu montrer que la clé d'un bon profilage est dans la compréhension de la nature humaine et le mode de fonctionnement de notre comportement. Comme de nombreuses études l'ont démontré, plusieurs niveaux de comportement, notamment le comportement conscient et inconscient influence l'attitude des personnes observées. Avec des techniques de persuasion appliquée, le changement de comportement et d'attitude se montre autant bénéfique que dangereux.

En fonction du besoin d'une personne où un groupe, le changement de comportement donc l'influence sur un sujet est facilement faisable, ce qu'on a vu dans des étapes du changement de comportement.

Une fois le mode de fonctionnement de comportement et d'attitude démontré, nous avons la réponse à la question "comment faire pour changer le comportement d'un individu et par quels moyens ?"

Ensuite, la base de profilage aujourd'hui mise sur le point d'anticipation de changement de comportement dont un possible passage en acte. Pour y prévenir, l'étude de pattern de comportement et le changement de ce pattern se révèle le mode opératoire préconisé dans les services de renseignement.

Avec l'arrive du Big Data et du traitement des données de masse, les nouvelles techniques de renseignement se profilent ainsi que des nouveaux défis. En partant de la législation dont la protection de la vie privée jusqu'au temps de traitement considérablement long. Ces nouvelles techniques appliquées en France comme dans d'autres pays se révèlent moins adaptées aux quantités de données produites chaque jour, des données structurées comme non structurés.

Pour comprendre le mode de fonctionnement de ces logiciels, une rétrospective des algorithmes de traitement nous explique les bases opératoires.

Comme une suite logique, je m'intéresse aux types de police prédictive et leur efficacité. En parcourant des logiciels utilisés aujourd'hui, je mets en évidence leurs point commun : les données.

Si des données sont au cœur du profilage aujourd'hui, je soulève le problème lié à la protection de la vie privée et la difficulté supplémentaire dans le domaine de profilage.

Vu que les sources des données sont diverses, leurs collecte et exploitation pose le problème dans le sens de l'exploitation comme temps de traitement, la proposition de changement de modèle de profilage naît.

L'idée de s'intéresser aux relations plutôt qu'au contenu me parait une solution possible. Pour répondre aux exigences de RGPD, le pseudonymisation des données avant le traitement se montre un modèle idéal. Avec des niveaux de chiffrement différents, la protection de la vie privée me semble résolu.

Pour réduire le temps de traitement qui peut atteindre plusieurs mois avant de donner des résultats, je me suis intéressé aux prétraitements dont des métadonnées. En les exploitant, j'ai trouvé que le temps réduit considérablement ce qui nous permet d'avoir des résultats dit "just in time".

Quant à avoir un meilleur aperçu, la visualisation des relations et corrélations s'impose comme la solution choisie.

Aujourd'hui, il est évident que la technologie progresse de jour en jour et si on ne suit pas ce progrès dans le domaine du profilage, les outils utilisés seront vite obsolètes. Le changement de modèle d'analyse se propose comme une solution qui nous permet d'aller avec le temps et d'assurer la pertinence des résultats. Bien que cette approche soit utilisée aux Etas-Unis depuis quelques temps sur certains procédés, ce qui nous a été révélé récemment par Snowden, l'Europe dont la France doit considérer rapidement le changement de cap et transformation de modèle existant pour se mettre en adéquation avec des demandes croissantes.

En travaillant sur ce mémoire j'ai appris des techniques, des défauts et j'ai vu la possibilité de construire un modèle nouveau applicable non seulement en criminologie, mais dans toutes les sphères de la société. Avec cette approche, le ciblage économique devient plus performant et notre intimité reste tout de même protégée.

Les connaissances que j'ai acquises durant cette période de professionnalisation en entreprise comme pendant mes cours m'ont permises de mieux comprendre le besoin et elles m'ont guidé vers une réflexion sur le plan technique autant que sur le plan conceptuel.

Aujourd'hui, après avoir fait cette recherche, j'ai changé mon point de vue concernant les réseaux sociaux et les données que l'on sème partout où l'on passe. Du côté technique, ce travail m'a permis de réfléchir et chercher des solutions en analysant le problème dans son fond, et pas seulement en pensant à résoudre temporairement.

Le métier de profileur avec des nouvelles technologies et des approches peut bien ressembler aux films de science-fiction fait auparavant, où des solutions fictives correspondant à notre époque actuelle semblent réel. Les perspectives de profilage sont très larges. En partant des entreprises privées et publiques, vers des utilisations personnelles.

Les fondations qui découlent de ce concept multidisciplinaire ont la flexibilité d'absorber n'importe quel type et n'importe quel volume de données, y compris des données existantes et héritées, et des données structurées et non structurées.

Avec l'application de la technologie de représentation graphique, on peut se permettre à la suite d'analyse d'offrir des performances optimales pour le monde connecté.

La plate-forme proposée peut automatiser la collecte de données connectées, en économisant du temps et des ressources, en apportant des résultats en temps réel pour accélérer la recherche et la prise de décision.

On se basant sur la solution graphique qui nous permet d'afficher des données, la prise des décisions en fonction de la représentation des données et des relations deviendra plus facile.

Avec un logiciel intelligent, l'avertissement automatique des changements de comportement dans les réseaux en fonction des paramètres prédéfinis peut être utilisés dans plusieurs domaines d'activités. Le concept d'acquisition de données automatiquement de n'importe quelle source et le traitement en temps réel avec la détection de nouvelles connexions et corrélations de données peut non seulement nous aider à être plus performant dans la prédiction, mais aussi nous augmenter la productivité générale dans tous les domaines d'activité où elle s'applique.

J'aimerai finir cet exposé par la citation de mon compatriote, *Nikola Tesla*, qui est l'une des sources de mon inspiration…

"Les combats entre les individus, aussi bien qu'entre les gouvernements et les nations, sont l'invariable résultat de l'incompréhension dans l'interprétation la plus large de ces termes. Les malentendus sont toujours causés par l'incapacité à apprécier le point de vue d'autrui. Et cela est dû à l'ignorance de ceux qui sont concernés, pas seulement dans leur champ propre, mais dans leurs domaines communs. Le danger d'un choc est aggravé par un sens plus ou moins prédominant de combativité, représenté par chaque être humain. Pour résister à cette tendance inhérente à la lutte, la meilleure manière est de dissiper l'ignorance des actes des autres par la propagation systématique de la connaissance générale. Avec cet objectif en vue, le plus important est de faciliter l'échange des pensées et des relations. "

Bibliographie

Abadi, D. (2009). Data management in the cloud: limitations and opportunities. In *Data management in the cloud: limitations and opportunities*.

Allan W. Wicker . Attitudes versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research.

ANDRIENKO, ANDRIENKO., & G.(2005. (s. d.). *Exploratory analysis of spatial and temporaldata: a systematic approach. 1er éd*. Berlin Heidelberg: Springer.

Bargh, J. A. (2004). Being here now: Is counsciousness necessary for human freedom ? In J. Greenberg, S. L. Koole, & T. Pyszczynski (Éd.), *Handbook of Experimental Existential Psychology* (p. 385–397). New York: Guilford Press.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). *Once Upon a Crime: Towards Crime Prediction from Demographics and*. Mobile Data.

Brakel, V., & Rosamunde. (2016). Pre-Emptive Big Data Surveillance and its (Dis)Empowering Consequences: The Case of Predictive Policing. *SSRN Electronic Journal*.

Brantingham, P. L., & Brantingham, P. J. (1993). *Environment, Routine, and Situation: Toward a Pattern Theory*. of Crime.

Breiman, L. (2001). Random forests. In *Machine Learning*.

Brinol, P., & Petty, R. E. (2003). Overt head movements and persuasion: A self-validation analysis. *Journal of Personality and Social Psychology*, *84*, 1123–1139.

Bughin, J., Byers, A. H., & Chui, M. (2011). *How business uses social technologies*.

Cacioppo, J. T., & Petty, R. E. (1981). *Communication and Persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Cortes, C., & Vapnik, V. (1995). Machine Learning. In *Machine Learning* (Vol. 20, p. 273).

Davenport, T. H. (2014). *Stratégie Big Data*.

Farrington, D. P. (2003). *Key results from the first forty years of the*.

Feblowitz, J. (2013). Using Information Intelligence to Improve Projects in the Energy Sector. In *IDC Energy Insights # EI225900R2. Using Information Intelligence to Improve Projects in the Energy Sector*.

Felson, M., & Clarke, R. V. (1998). *Opportunity makes the thief: practical theory for crime prevention*. Police: Research Series.

Festinger, L., & Maccobby, N. (1964). On resistance to persuasive communications. *Journal of Abnormal and Social Psychology*, *68*, 359–366.

Fointiat, V., & Barbier, L. (s. d.). Persuasion et Influence : changer les attitudes, changer les comporte-ments. Regards de la psychologie sociale. *Journal d'Interaction Personne-Système (JIPS)*, *2015*(4), 1–18. https://doi.org/http://jips.episciences.org>. <hal-01207402

Fointiat, V., Girandola, F., & Gosling, P. (2013). *La Dissonance cognitive. Quand les actes changent les idées*. Paris: Armand Colin.

Greenwald, G., & A. (1968). On Defining Attitude and Attitude Theory. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Éd.), *Psychological Foundation of Attitude*. New York: Academic Press.

Groves. (2013). *Discussion on advancing the methods for quality improvement research*. Licensee BioMed Central Ltd.

Guéguen, N., & Pascual, A. (2000). Evocation of freedom and compliance: The « but you are free of » technique. *Current Research in Social Psychology*, *5*, 264–270.

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate data analysisUppersaddle River* (6e éd.). N.J: Pearson Prentice Hall.

Halevy, A., Norvig, P., & and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intell. Syst*, *24*, 8–12.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments on mass media communication*. Princeton, NJ: Princeton University Press.

https://www.cnil.fr/. (s. d.). Consulté à l'adresse https://www.cnil.fr/

http://www.lecerclemarketingclient.com/parole/y-a-t-il-un-consentement-au-profilage-les-precisions-des-guidelines-du-g29/. (s. d.). Consulté à l'adresse http://www.lecerclemarketingclient.com/parole/y-a-t-il-un-consentement-au-profilage-les-precisions-des-guidelines-du-g29/

Hunt, P., Saunders, J., & Hollywood, J. S. (2014). Evaluation of the Shreveport Predictive Policing Experiment. In *Evaluation of the Shreveport Predictive Policing Experiment*.

Jean-Paul Brodeur, Directeur CICC. Un article de la revue Criminologie, Volume 38, Numéro 2, Automne, 2005

Joule, R.-V., & Beauvois, J.-L. (1998). *Petit traité de manipulation à l'usage des honnêtes gens*. Grenoble: Presses Universitaires de Grenoble.

Joule, R.-V., Girandola, F., & Bernard, F. (2007). How can people be induced to willingly change their behavior ? The path from persuasive communication to binding communication. *Social and Personality Compass*, *1*, 493–505.

Joule, R.-V., Gouilloux, F., & Weber, F. (1991). The lure: A new compliance procedure. *Journal of Social Psychology*, *129*, 741–749.

Kaloxylos, A. (2012). *Computers and Electronics in Agriculture 89*.

Kiesler, C. A. (1971). *psychology of Commitment: Experiments linking behaviour to belief*. Academic Press.

La recherche et la gestion des liens dans l'investigation criminelle: le cas particulier du cambriolage. (s. d.) (Thèse de doctorat). Lausanne, Switzerland.

Lin, J., & Schatz, M. (s. d.). *2101. Design Patterns for Efficient Graph Algorithms*.

Loshin, D. (2010). The Practitioner's Guide to Data Quality Improvement. In *The Practitioner's Guide to Data Quality Improvement*.

Mastrobuoni, G. (s. d.-a). *Optimal criminal behavior and the disutility of jail: Theory and evidence*.

Mastrobuoni, G. (s. d.-b). *Police and Clearance Rates: Evidence from Recurrent Redeploy-ment Within a*. City.

Mastrobuoni, G., & Owens, E. (2014). *Criminal careers and criminal firms*.

Moore, G. (s. d.). *Cramming More Components onto Integrated Circuits.*

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. In *Bitcoin: A Peer-to-Peer Electronic Cash System*.

Oatley, G., Zeleznikow, J., Leary, Rewart, & B.(2005. (s. d.). From links to meaning: aburglary data case study. In *Dans International Conference on Knowledge-Based Intelligent Information and Engineering Systems* (p. 813–822). Melbourne, Australia: Springer.

Patil, T. H. D. (2012). *Data Scientist: The Sexiest Job of the 21st*. Century.

Pay, A. M. (s. d.). *reference points, and police performance*.

Petty, R. E. (1994). Two routes of persuasion: State of the art. *International Perspectives on Psychological Science*, 229–247.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental*

Social Psychology, *19*, 123–205.

Prochaska, J. O., DiClemente, C. C. e. N., & J.C. (1992). In search of how people change: Applications to addictive behaviors. *American Psychologist*, *47*(9), 1102–1114.

Richard H. Thaler, Cass R. Sunstein (2012). Nudge

Rosenberg, M. J. (s. d.). An analysis of affective-cognitive consistency. In M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson, & J. W. Brehm (Éd.), *Attitude organization and change*. Yale University Press.

Snijders, T. (2011). Multilevel Analysis. International Encyclopedia of Statistical Science. In *Multilevel Analysis. International Encyclopedia of Statistical Science* (p. 879–882).

Edward Joseph Snowden, ancien employé de la Central Intelligence Agency

The Quarterly Journal of Economics. (2006). *The Quarterly Journal of Economics*, *3*(783–821).

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension*. Cambridge, MA: MIT Press.

UE. (2016). RGPD. In *Parlement européen et du Conseil du 27 avril*.

Verplanken, B. (2006). Beyond frequency: Habits as mental construct. *British Journal of Social Psychology*, *45*, 639–656.

Verplanken, B., & Aarts, H. (1999). Habit, attitude and planned behavior: Is habit an empty construct or an interesting case of goal-directed automaticity? *European Review of Social Psychology*, *10*, 101–134.

Viktor Mayer-Schönberger, Kenneth Cukier. A Revolution That Will Transform How We Live, Work, and Think

www.statista.com. (s. d.). Consulté à l'adresse www.statista.com

Zanna, M. P. K., C, A., & Pilkonis, P. A. (1970). Positive and negative attitudinal affect established by classical conditioning. *Journal of Personality and Social Psychology*, *14*, 321–328.

Zikopoulos, P., & Eaton, C. (s. d.). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming*. Data.