COMPTE-RENDU TABLE RONDE
BIG DATA ET CYBERSÉCURITÉ

Animateur

Camille SPOKOJNY, *Polyconseil*
Diplômée de l'Ecole de Guerre Économique, préside le Club Big Data *et Intelligence Économique de cette dernière. Elle a travaillé avec l'observatoire des* Big Data*, et est actuellement en poste chez Polyconseil.*

Intervenants

Marie-Thérèse ANDRÉ, Direction Générale de l'Armement (DGA)
Après une formation d'ingénieur en analyse système de l'ENSTA Paris, forte d'une première expérience de 8 ans dans le privé dans le développement de solutions logicielles, elle travaille dans le domaine de l'ingénierie des systèmes navals. Marie-Thérèse André a ensuite rejoint la DGA en 1998, successivement en tant que : chargée de la maîtrise d'ouvrage du développement de produits de sécurité gouvernementaux, architecte conseil pour la sécurité dans des systèmes d'information, elle devient expert technique et responsable du laboratoire d'analyse de la menace système, puis du laboratoire de techniques de détection. Depuis 2015, elle est chef du département défense et protection système.

Charles HUOT, Ecole de Guerre Économique (EGE)
Docteur en sciences de l'information et de la communication (Université de droit, d'économie et des sciences d'Aix Marseille), et auteur d'une thèse intitulée « Analyse Relationnelle pour la veille technologique : vers l'analyse automatique des bases de données » en 1992. Charles Huot multiplie les mandats puisque il est Président de l'association GFII (Les acteurs du marché de l'information et de la connaissance), Vice–Président du Pôle de compétitivité Cap Digital en charge des relations avec les experts, Président du comité éditorial de L'Alliance Big Data, Enseignant à l'École de Guerre Économique (EGE) sur l'introduction au Big Data, mais aussi Coordinateur du groupe de travail de l'AFNOR sur les Big Data et la Normalisation. Co-fondateur et Corporate Development Officer en charge du développement stratégique et de l'innovation, il représente la société Expert System auprès des industriels de son secteur et d'instances françaises et européennes.

Colonel Patrick PERROT, Service Central de Renseignement Criminel (SCRC), Gendarmerie Nationale
Ingénieur télécoms de formation, il est titulaire d'un doctorat en intelligence artificielle (Telecoms Paris Tech). Il est l'auteur de nombreuses publications dans le domaine de la biométrie et du renseignement. Également Officier de gendarmerie, le colonel Patrick Perrot a alterné les commandements opérationnels et les fonctions scientifiques notamment dans le domaine forensic et du renseignement. Depuis 2013, il est affecté au service central de renseignement criminel de la gendarmerie nationale à Pontoise.

Quang-Minh LESPECHEUX, Microsoft
Titulaire d'une maîtrise en droit des affaires et d'un DESS en droit des nouvelles technologies, il a commencé sa carrière en tant que responsable juridique et des affaires publiques de l'Union des Fabricants. Depuis 5 ans, il a rejoint Microsoft, en occupant le poste de juriste spécialisé en cybercriminalité et propriété intellectuelle en charge de la France et l'Europe du sud, puis celui de responsable des affaires publiques pour la France. Il est actuellement toujours en poste.

Charles HUOT
EGE

Il s'agit ici de faire une petite introduction aux Big Data, et présenter le projet en cours. La cybersécurité a un sens particulier. On s'intéresse aux Big Data pour la sécurité pour être capable de traiter la variété et le volume de données. La variété d'informations est beaucoup plus complexe que celle traitée dans les métiers de l'industrie.

L'expression « *Big Data* » a émergé en 2011, même si le *data mining* est apparu dans les années 90. Nous pouvons remonter plus loin chercher les racines de la pratique, l'analyse des données remontant aux années 70. Un grand nombre de méthodes utilisées dans le traitement des données émerge de ce champ d'application.

« La complémentarité des savoirs est indispensable, il faut faire travailler ensemble des savoirs de natures différentes comme ceux de l'EHESS et celui des ingénieurs. Nous rencontrons alors un enjeu propre au système d'information, celui de l'ouverture des silos d'informations. »

La question est de savoir ce qui est « Big » ? Google, Facebook, Microsoft ont accès à un grand nombre de données, mais est-ce vraiment là qu'est le Big Data ?
Le Big Data est construit par trois grands piliers :

1. Les données de type industrielles, l'*open data* : ce sont des données issues d'Internet, ou émergeant de métiers spécifiques. Il faut parler alors d'espace des données. Le numérique s'est attaché à traiter les chiffres. Il y a une difficulté à traiter les images ou les textes, à comprendre le contenu d'une vidéo. Une nouvelle vague est née, initiée par la capacité des systèmes à analyser le texte ou l'image, comme en témoigne le patron de Facebook à Paris, Laurent Solly, qui a relancé le *machine learning*.
2. Les techniques d'information : avec le traitement automatique des images et des textes, des techniques d'interprétation, de visualisation, de stockage, ou encore d'interprétation de la connaissance sont rendues accessibles.
3. Le traitement de l'information : les organisations ont compris leur intérêt dans le traitement de la donnée. La visibilité donnée par des données maîtrisées permet l'amélioration de leur processus voire de produire de nouveaux produits ou services.

Pour les non-initiés, nous pouvons citer la série *Person of Interest*, qui s'illustre par la présence d'une Intelligence artificielle (IA) capable, par le croisement et la transformation des données, de produire des informations stratégiques.

Avec *Expert System*, nous travaillons sur le portage des algorithmes sur de nouvelles structures. L'enjeu est d'être capable de faire évoluer ses structures algorithmiques sur de nouveaux supports. Le projet IMM (projet d'Information Multilingue Multimédia) réunit des intégrateurs de contenu et des PME technologiques pour alimenter en donnée les PME.

Traduction automatique, texte, communication... Un grand problème de l'information non structurée est le traitement de l'information venant de plusieurs sources. Combien de temps pour que la machine comprenne et traduise une nouvelle langue ? Poursuite par IMA → intégrer plus audiovisuel et image, le CEA, le CNRS, capteurs d'informations pour aider à la décision ou à l'identification des sujets nouveaux. Par exemple, avec la radicalisation, il s'agit de déterminer les prédicats par l'analyse des sites web, des réseaux sociaux couplés avec les sources de propagande de DAECH. Dans cette optique, la complémentarité des savoirs est indispensable, il faut faire travailler ensemble des savoirs de natures différentes comme ceux de l'EHESS et celui des ingénieurs. Nous rencontrons alors un enjeu propre au système d'information, celui de l'ouverture des silos d'informations.

Quang-Minh LEPECHEUX
Microsoft

Il abordera le sujet des Big Data de manière concrète en mettant en valeur leur potentiel en matière de sécurité, en s'appuyant sur des exemples, puis il appréhendera les enjeux juridiques, économiques et sociétaux de ces technologies.

Le Big Data est là pour faire le lien entre le présent et l'avenir. Il s'agit de fournir des produits adaptés aux clients. D'ici à 2020, le marché du Big Data représentera 5 milliards de dollars, alors que les attaques informatiques coûtent 400 milliards de dollars. Microsoft n'est pas nouveau dans le secteur du Big Data, le correcteur orthographique de Word était déjà alimenté par plus de 10.000 mots, il y a 10 ans.
Quels sont les implications des Big Data du *machine learning* ? Intelligence Artificielle et droits juridiques.
La 4ème révolution industrielle est une révolution numérique. Les *Big data* emmènent les entreprises. Il y alors une nécessité d'avoir un cadre juridique pour pouvoir évoluer. Les clients de Microsoft sont attaqués en permanence, avec des attaques de plus en plus complexes : la question n'est plus de savoir « quand » on sera attaqué, mais comment.

Les *Advanced Persistent Threats* (APT) sont des attaques de longue durée mais d'intensité différentes, mises au point pour récupérer des logs sur un temps long. Afin de les détecter, l'analyse du système s'appuie sur l'étude comportementale des utilisateurs afin d'identifier les connexions anormales au réseau, comme la vitesse de frappe, où des mouvements de souris qui seraient d'une rapidité inhabituelle.

« La question n'est plus de savoir « quand » on sera attaqué, mais comment. »

La *Digital Crime Unit* de Microsoft dispose d'un serveur rempli par les infections par botnets, ainsi que des chemins d'accès demandés par les machines infectées par ces mêmes botnets. Ce serveur est rempli par l'intermédiaire d'actions en justice. Ces dernières permettent de stopper l'atteinte au système, en demandant une coupure entre la machine infectée et le serveur. On peut très bien demander de rediriger le pc vers un serveur spécifique. Le ping demandé au serveur permet de nourrir une base de données d'adresse IP infectées. Cette technique permet de visualiser les données. Les attaques peuvent être calibrées au cas par cas pour une infection ciblée. Avec le botnet *Citadelle*, une carte de visualisation des machines infectées fut produite. Étonnamment avec la *Datavizualisation*, il apparaît que le botnet respecte la frontière ukrainienne et russe. Après une investigation poussée du botnet, il est apparu qu'il ne s'activait qu'uniquement lorsque l'alphabet de l'OS choisi était en caractère latin.

Microsoft cherche à développer les services aux individus en protégeant les données sensibles de ces derniers, avec les nombreuses informations captées par les systèmes d'information et la personnalisation des services proposés. Le droit d'auteur pose aussi des problèmes aux scientifiques. Des chercheurs pourraient traiter des données anonymisées pour les scientifiques. La limitation de communication des données de géolocalisation apparaît un non-sens pour le développement.
La concentration des données augmente leur risque de vulnérabilité. Microsoft a produit un livre blanc : *Du Big Data au Big Business.*

« Microsoft cherche à développer les services aux individus en protégeant les données sensibles de ces derniers, avec les nombreuses informations captées par les systèmes d'information et la personnalisation des services proposés. »

Question de Guillaume Vincent, Air liquide : avec les GPR (*General Purpose Registers*), on ne demande pas de localisation. Il aimerait connaître leur position sur le conflit qui se pose entre les Big Data/computing et la réglementation.
Certaines entreprises comme les archives nationales ont besoin d'être géolocalisées en France, mais en pratique cela serait très difficile à appliquer à tout le monde. Géolocaliser les données au niveau européen est possible, mais assurer le fonctionnement des *datacenters* 24/24 serait difficile.
Avec les *free flow data*, nous lançons des *datacenters* en France, déjà présents en Irlande. La France est un canal de circulation des données et cela permet d'ouvrir de nouveaux marchés. Nous développons aussi une solution de *back up* en hollande.

Marie-Thérèse ANDRE
DGA

Elle aborde aujourd'hui le sujet sous l'angle suivant : « *Le Big Data et la détection des attaques ou des intrusions visant les systèmes acquis par la DGA au profit des forces armées* ».

Les attaques informatiques peuvent porter atteinte à l'intégrité opérationnelle. Il y a besoin de systèmes d'information déployés sur les terrains d'opération. Les rafales, les navires, sont des systèmes d'information. Les failles numériques sont omniprésentes et souvent là où on ne les soupçonne pas. Comme avec les industriels.
Avec la protection contre les intrusions, il faut développer un posture plus dynamique pour déceler les tentatives d'intrusion. Il s'agit alors d'observer l'activité du système d'information et les connexions entrantes (potentiellement toutes les activités sont intéressantes, qu'il s'agisse des flux légitimes, ou des flux illégitimes qui sont potentiellement malveillants). En croisant les données de plusieurs systèmes, cela devient intéressant. A l'échelle du Ministère de la Défense, nous rentrons dans les *Big Data*.

Les attaques de longue durée travaillent sur une échelle de temps long, produisant un volume de données important. Une analyse des données est menée conjointement avec l'ANSSI pour déceler de nouveaux modes opératoires. Utilisée avec ces méthodes, l'analyse rétroactive s'opère sur une durée de plusieurs années.

« Avec la protection contre les intrusions, il faut développer un posture plus dynamique pour déceler les tentatives d'intrusion. Il s'agit alors d'observer l'activité du système d'information et les connexions entrantes (potentiellement toutes les activités sont intéressantes, qu'il s'agisse des flux légitimes, ou des flux illégitimes qui sont potentiellement malveillants). »

Le constat est que les solutions de cybersécurité, même privées, ne permettent pas de croiser les informations sur une durée de temps long. Il faut passer à une autre technologie permettant une recherche rapide par rapport à quelque chose de connu et capable de détecter des anomalies sans connaître l'anomalie recherchée. Un comportement normal pourrait être qualifié d'anormal par un système, du fait de l'évolution des usages dans le temps.
Le *Big Data* n'est encore pas mature, il faut des architectures pour traiter de façons rapides les données produites, sur des données stockées sur plusieurs postes, un modèle décentralisé, une solution *NoSQL*. Le besoin de solution prône une nouvelle approche. L'intelligence artificielle et les *data analytics* du domaine de la santé se présentent comme un marché transposable vers la Défense du fait de la modélisation que ces activités demandent toutes deux.
La visualisation de la donnée permet de visualiser les données échangées. Cependant l'humain reste essentiel, il faut fournir des outils adaptés. Savoir quels autres outils peuvent les donner. Un pré-calcul doit être fait pour accéder à d'autre données. Les sets de données doivent être évalués avant de les mettre dans les centres de supervision.
Il faut évaluer la pertinence et l'efficacité de ces données. Une fois passées en production, difficile d'évaluer les attaques connues. Des données sur masse avec des traces d'attaques connues. Pour être sûr que le logiciel marche.
En interne, nous avons des chercheurs. Nous recrutons des statisticiens pour apporter d'autres regards. Nous élaborons un dispositif pour soutenir les PME, un dispositif d'aide à la décision duale. Nous invitons les PME à proposer des sujets sur la cybersécurité. (RAPID)
Sur la visualisation, beaucoup d'autres travaux sont prévus. Nous travaillons sur un processus d'anonymisation. Un manque de solution industrielles.

Colonel Patrick PERROT
SCRC

Son propos portera sur l'aide à la prise de décision par le *Big Data*, l'analyse prédictive et donc du *machine learning*.

Les redondances de nos exposés nous montrent au moins que nous ne sommes pas hors sol. « *Le* Big Data *et la cybersécurité sont un iceberg qui n'est pas soumis à la fonte* ».
Tout n'est pas mature, et nous devons faire part d'une prudence apparente. L'image négative des *Big Data* et de la sécurité sont toujours assimilées à *Minority Report* ou à *Big Brother*, ce qui est un vrai frein pour son développement, et pourtant la démarche est urgente.

« Le besoin de solution prône une nouvelle approche. L'intelligence artificielle et les data analytics du domaine de la santé se présentent comme un marché transposable vers la Défense du fait de la modélisation que ces activités demandent toutes deux. »

Il y a un avant et un après : avant, la donnée était rattachée à la notion de propriété, au moins de manière sous-jacente ; après, une donnée déstructurée est partagée avec intérêt. Le cadre juridique évolue en permanence pour cadrer les évolutions de son usage. Nous assistons à l'apparition de données en streaming, un flux continu de données plutôt qu'un stockage statique.
La gendarmerie n'est pas une start-up, nous ne travaillons pas sur le dernier concept à la mode mais plutôt sur les bases. Notre attention se porte donc sur les mathématiques. L'intelligence artificielle existe depuis longtemps. Les réseaux neuronaux existent aussi depuis des années, il faut donc continuer à s'appuyer sur les mathématiques pour le monde d'après.

« On me fait souvent la réflexion que l'analyse prédictive ne marche pas, sinon nous pourrions gagner au loto. (...) La prédiction criminelle a vocation à ne pas se réaliser par la réalisation de contre-mesures. »

La démarche de la gendarmerie se divise en deux pôles : le C3E, et un centre analytique, c'est-à-dire un espace cyber et un espace scientifique. Nous travaillons sur la détection de failles, de comportements anormaux.
Pourquoi faisons-nous des *Big Data* ? La donnée doit servir à la décision, pour décider nous utilisons nos connaissances, notre intuition.
Trois types d'analyses sont d'usage et appuyées par l'utilisation des Big Data :

- L'analyse prédictive : à partir d'un héritage du passé (les données), nous pouvons déterminer ce qui arrivera de la manière la plus probable possible.

- On me fait souvent la réflexion que l'analyse prédictive ne marche pas, sinon nous pourrions gagner au loto. Cette réflexion fut faite de la part d'un criminologue et fut reprise par les médias. La prédiction criminelle a vocation à ne pas se réaliser par la réalisation de contre-mesures. Il faut trouver la variable qui a modifié la situation. Le modèle reste valide par rapport au données antérieures. Il faut intégrer des données extérieures pour rendre l'analyse variée. Par exemple, à New York, les inspecteurs des incendies utilisent des données relatives aux constructions, comme la date, les matériaux... en tout, une cinquantaine de variables pour déterminer quels immeubles sont a visiter en premier. Avec l'idée d'une police prédictive, il faut toutefois laisser des patrouilles là où il ne se passe rien.

  « La construction d'un modèle de pensée repose sur un ensemble de comportements individuels analysés. Nous détruisons ensuite les données personnelles. »

- L'analyse prospective : elle consiste en l'élaboration de *scenari*, même les plus improbables ; mais il n'est pas possible d'évaluer la probabilité de cet événement. Le futur n'est pas entièrement prédictif.

- L'analyse situationnelle : il s'agit de coupler les données dont nous disposons avec d'autres données de l'environnement pour prévoir, non pas un modèle, mais plutôt un schème. L'homme est toujours en présence car ce dernier dispose de la

faculté d'intégrer des faits nouveaux. Nous revenons sur un caractère intuitif.

L'idée que la machine prend la décision ne peut pas être vrai en tout temps. Mais il faut reconnaître son aide.
Avec les données, il faut se demander laquelle faut-il prendre et quelle est sa valeur ? Celle de nos données internes ? Celle que nous diffusons de manière déstructurée ? Nous voyons aussi une augmentation des communications dans le *Darkweb* qui augmente de 80%, mais ce ne sont pas obligatoirement des criminels.

« Avec l'idée d'une police prédictive, il faut toutefois laisser des patrouilles là où il ne se passe rien. »

Les méthodes d'analyse :

- La méthode de régression : l'idée est d'agir le plus en amont possible afin d'impliquer le plus d'acteurs dans la chaîne pour avoir des relais.
- La méthode factorielle : chaque jour des résumés des infractions constatées sont produits. Cela permet d'identifier les secteurs sur lesquels agir. Soit, cela enfonce des portes ouvertes ; d'un côté heureusement que l'analyse valide des faits observés. Oui, les vols de carburants ont lieu là où il y a des véhicules. Mais aussi, il est apparu que le vol d'accessoires ne se pratiquait qu'uniquement sur certaines marques.

Le *quantum learning* est difficile a appréhender cognitivement, l'ordinateur quantique serait capable d'utiliser la superposition d'états.
Le *machine learning* est-il intrusif ? La construction d'un modèle de pensée repose sur un ensemble de comportements individuels analysés. Nous détruisons ensuite les données personnelles. L'arnaques au président est courant, il serait possible de produire un modèle d'après les attaques passées pour connaître les cibles potentielles.

« La prédiction est difficile surtout lorsqu'elle concerne l'avenir. »

La théorie des graphes : avec la notion de centralité, il s'agit de connaître les degrés, les liens qui composent une organisation, afin de pouvoir identifier qu'il faut toucher pour toucher de manière durable la structure. Celui qui est en position de centralité n'est pas toujours le leader. C'est le rôle fonctionnel qui est mis en évidence, le rôle de l'individu. Bien sûr, du moment qu'il n'y a pas tout le réseau, le modèle restera imparfait.

Je suis personnellement contre une uberisation de la sécurité. En Angleterre, quand on achète un logement, on peut avoir accès aux informations relatives à la criminalité du quartier sans passer par les services d'État. Il faut s'attacher à la notion d'évaluation, il y a un besoin d'être rigoureux. Il faut évaluer les faux positifs et ses propres moyens, c'est-à-dire la manière dont le personnel s'approprie ces outils. Cela pose les difficultés d'exploitation. Je tiens à rajouter que les solutions d'analyse prédictive ne sont laissées qu'aux chefs et ne sont pas accessible aux patrouilles.
Quelle transparence des algorithmes ? Il suffit d'aller sur Github pour avoir accès à la plupart des algorithmes utilisés.
En complément des méthodes d'investigation classique, nous utilisons l'analyse *a priori* au SCRC, et l'analyse *a posteriori* au SIP. Rappelons que « *la prédiction est difficile surtout lorsqu'elle concerne l'avenir* » (Pierre Dac).

<u>Points thématiques :</u>

- *Data* et *open data* : on nous a déjà demandé de mettre à disposition les faits de cambriolage qui sont répertoriés par adresse. Même la gendarmerie n'a pas le droit de l'utiliser. Nous sommes bloqués par le CNIL. Nous souffrons d'un manque de données.

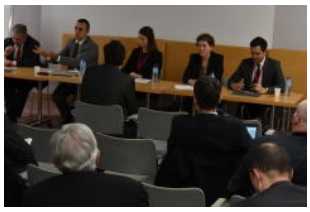
  « Avec les objets connectés, les enjeux vont être énormes. »

- Les boîtiers connectés : développés par AXA, pour capter les bons conducteurs en premier. AXA a formulé cela comme une demande d'utilisateurs. Ce fut une co-création, la communauté a ensuite porté l'offre. Cependant, il faut noter que certains automobilistes, bon conducteur avec un bon score, ont voulu arrêter car étaient trop stressés à chaque fois qu'ils prenaient la voiture.

- Avec les objets connectés, les enjeux vont être énormes. Du point de vue de la cybersécurité, après l'attaque du botnet MIRAI, il faut en prendre conscience. Les IoT sont de plus en plus nombreux et de plus en plus petits. Ce sont des petits capteurs qui transmettent une information structurée, il s'agit alors de l'exploiter.

- Cela est plus une question de normalisation. Nous pouvons nous demander si la dictature des 10 000 pas / jour a un sens du point de vue médical. Les données extraites sont sorties de l'aspect médical.

- Problème de l'effet tunnel : il faut penser à une diversification des compétences, des savoirs. Pour chaque nouveau projet lancé à la DGA, un comité d'éthique est mobilisé pour chaque allocation de recherche. Nous nous interrogeons sur comment utiliser les données des réseaux sociaux, par exemple, de plaintes d'effets secondaires de médicaments qui serait observables. Nous pouvons voir que celui qui détient la connaissance ne l'utilise pas forcément.

CE QU'IL FAUT RETENIR

⧫ D'ici à 2020, le marché du Big Data représentera 5 milliards de dollars. Les attaques informatiques coûtent déjà 400 milliards de dollars.

⧫ La complémentarité des savoirs est indispensable : il faut faire travailler ensemble des savoirs de natures différentes, des chercheurs issus des Sciences Humaines et Sociales (SHS) et des chercheurs issus des sciences dures (mathématiciens, ingénieurs…). Il faut penser à une diversification des compétences, des savoirs.

⧫ Le besoin de solution prône une nouvelle approche. L'Intelligence artificielle (IA) et les *data analytics* du domaine de la Santé se présentent comme un marché transposable vers la Défense.

⧫ Il faut soutenir les efforts pour développer une Intelligence artificielle (IA) capable, par le croisement et la transformation des données, de produire des informations stratégiques, d'aide à la prise de décision par le *Big Data*, ou encore développer l'analyse prédictive.

⧫ Le *Quantum learning* est difficile à appréhender cognitivement, mais est porteur de nouvelles révolutions à venir. L'informatique quantique doit être un champ d'études et d'application soutenu et protégé par l'État.