CYCLE DE CONFERENCES SUR LE NUMERIQUE : LE BIG DATA

Conférence par Alain Dupuis, responsable informatique de la Comunnauté d'agglomération de l'Albigeois.

MEDIATHEQUE PIERRE-AMALRIC D'ALBI

Samedi 28 mai 2016 – 16h00

Thème de la conférence

Le BIG DATA, quézako?

Pour faire face à l'explosion du volume des données, un nouveau domaine technologique a vu le jour: le Big Data. Inventées par les géants du web, ces solutions sont dessinées pour offrir un accès en temps réel à des bases de données géantes. Mais de quoi parle-t-on réellement ? Comment ces données sont-elles gérées et par qui ? Quel est l'intérêt économique ? Et nos données personnelles, dans tout ça : que deviennent-elles ? N'y a-t-il pas un risque de dérive et comment nous en prémunir ?

Introduction

Le sujet est la «donnée», «l'information» , particulièrement celle qui est numérisée et qui nous concerne tous puisque, peu ou prou – et on pourra mesurer que c'est plutôt «prou» (du sens beaucoup, mais aussi de l'origine latine prode = profit) – elle nous caractérise, elle décrit nos comportements, elle a un impact fort sur les systèmes économiques, sociaux ou politiques dans lesquels nous nous inscrivons.

L'histoire de la données numérique est très récente - grosso modo moins de 50 ans – et est liée à l'évolution de l'électronique et de la science de la cybernétique.

Repères:

Cybernétique : Norbert Wiener -1947- tentative de vision unifiée des domaines de l'automatisme, de l'électronique et de la théorie (mathématique) de l'information.

Théorie de l'information : Claude Shannon -1948- cryptographie, schéma de Shannon.

Avec l'explosion des technologies de stockage (disques), de calcul (processeurs de plus en plus rapides et puissants), de transport (fibre optique, réseaux étendus), de télécommunications (communications de longues distance et en mobilité notamment), mais aussi de traitement (intelligence et puissance des logiciels), la donnée numérique est devenue rapidement un bien immatériel de plus en plus facile à collecter, stocker, analyser.

Repères:

Les premiers calculateurs électroniques (années 1950/1960)

L'informatique distribuée (années 1970)

L'essor des PC (années 1980)

L'arrivée de l'Internet (1995, en France ADSL en 2002)

Les premiers smartphones et les réseaux de type EDGE (2001)

La 3G (2005)

Les tablettes (2010)

La 4G, le début du FTTH (2012)

Le concept d'objets connectés, la réalité du Big Data (2015)

Pour qualifier ces volumes de données de plus en plus grands, de plus en plus immatériels et aux perspectives de traitement énormes (et en partie encore insoupçonnées), le terme de Big Data est apparu, presque à l'arrivée du XXIe siècle (1997, Association for computing machinery USA-NYC) mais on le voit surtout exploser dans de nombreux articles depuis deux ans, soit comme concept susceptible de révolutionner l'organisation de nos sociétés ou de nos économies.

Le terme est assurément à la mode, mais que cache-t-il?, c'est quoi (ou qui) ce Big Data?

Je vous propose une intervention qui va s'attacher à répondre à quelques questions en apparence basiques:

QU'EST-CE QUE LE BIG DATA, PEUT-ON EN DONNER UNE DEFINITION?

QUI «GERE» TOUTES CES DONNEES COLLECTEES, OU SONT-ELLES LOCALISEES?

QUEL INTERET IL Y A-T-IL A LA MISE EN OEUVRE DU BIG DATA?

LES DONNEES PEUVENT ETRE PERSONNELLES, QUELS SONT LES RISQUES DE DERIVES?

QU'EST-CE QUE LE BIG DATA, PEUT-ON EN DONNER UNE DEFINITION?

Questions préalables: Qui n'a pas un téléphone portable, un smartphone, une tablette? un PC ou portable à son domicile? Qui n'a pas créé une page sur un réseau social?

Commençons par un panorama d'applications liées à des données qui sont désormais systématiquement numérisées.

1 – LES RESEAUX SOCIAUX

Un réseau social est un outil numérique qui permet de constituer des groupes «d'amis» et de partager «entre amis» - ou plus si affinités - des opinions, commentaires, événements, photos, vidéos...

Approche fonctionnelle: vous souhaitez vous intégrer à un réseau social, vous créez alors un compte: même sans ami, vous avez déjà généré de la donnée (voir le parcours de création). Vous avez des amis, vous échangez, vous générez des données comportementales qui sont stockées et analysables.

Nous parlons ici d'un réseau de type Facebook, une des premières sources de trafic au monde et une des plate-formes numériques disposant du plus grand nombre de renseignements personnels sur ses utilisateurs. Pour l'anecdote, en décembre 2013, il a été démontré que ce réseau social peut également avoir connaissance de ce que nous écrivons mais effaçons avant publication.

Depuis Février 2015, les conditions d'utilisation de Facebook lui permettent également d'utiliser les données de messagerie instantanée (WhatsApp) et publication instantanée de photos (Instagram).

Ceci s'effectue en principe avec le consentement de l'utilisateur.

Ce qui n'est pas annoncé : même si vous n'êtes pas inscrit sur Facebook, le site dépose à votre insu un "cookie" sur votre machine dès que vous consultez une page Facebook publique (événement, page fan…). Ce discret petit fichier peut suivre à la trace votre navigation internet, et relever des informations propres à votre identité (localisation, langue, machine utilisée...).

2 – LE MOTEUR DE RECHERCHE, un moteur 10 puissance 100 (Gogol d'où GOOGLE) et les applications satellites

Lorsque vous utilisez un moteur de recherche, votre navigateur lui transmet automatiquement certaines informations (votre adresse Ip, l'adresse de la page que vous consultez). Dans le cas du moteur Google, le plus utilisé au monde (88,66% de part mondiale en 2015), il peut également déposer des cookies dans votre navigateur ou lire ceux qui sont déjà présents.

A ce moteur sont associés des services ou outils :

Le compte Google+ : informations de création du compte (et elles sont nombreuses), courriel envoyés ou reçus, contacts, événements de l'agenda, photos, vidéos, documents sur Google drive

La cartographie googleMaps : lorsque vous l'utilisez, elle recueille ce que vous recherchez, votre position géographique, les annonces que vous «cliquez ou touchez», l'adresse Ip et les cookies

Youtube: idem

Androïd : principalement utilisé sur les outils de mobilité, cet OS peut recueillir toutes les positions GPS du mobile, smartphone ou tablette.

Ceci s'effectue en principe avec le consentement de l'utilisateur.

3 – LES TELECOMMUNICATIONS, l'opérateur ORANGE et l'application FLUX VISION

Accroche commerciale : Je souhaite mieux connaître mes clients, décrypter leurs habitudes de consommation, leurs comportements d'achat pour améliorer les services offerts. Parce que la connaissance des clients est essentielle, Flux vision convertit chaque minute 4 millions de données mobiles en indicateurs statistiques pour mesurer la fréquentation d'une zone géographique et le déplacement des populations. (D'où viennent les clients, comment ils se déplacent sur un territoire, combien de temps ils y restent -> l'exemple de Chamonix Mont Blanc ainsi que Bouches-du-Rhône tourisme)

Flux vision est également en mesure de fournir la segmentation des individus par tranche d'âge et par genre.

Il s'agit ici d'informations collectées et analysées à partir d'un couplage mobile – antenne, en croisant avec l'information relative au client connue de l'opérateur.

Orange garanti l'anonymisation complète et irréversible, mais il s'agit bel et bien de vos données liées à l'usage du mobile.

Cette collecte s'effectue à l'insu de l'utilisateur.

4 – D'AUTRES SECTEURS

LE SECTEUR BANCAIRE : Combien de fois communiquons nous des informations plus ou moins personnelles (formulaires administratifs, cartes commerciales, informations bancaires, paiements par carte bancaire? C'est une quasi évidence que de dire que les banques savent généralement tout de leurs clients (salaire, épargne, propension à dépenser ou non, commerces favoris voire habitudes de consommations).

L'ENERGIE, ERDF et les compteurs Linky et Gazpar : Ces compteurs peuvent recevoir des ordres et envoyer des données sans intervention physique d'un technicien. «L'accès à nos données de consommation », selon ERDF, « permet de mieux la maîtriser et de proposer de nouvelles offres d'énergie et de nouveaux services (pilotage des appareils à la maison)» (Données transmises: relevé quotidien à distance de la consommation réelle et courbe de charge, puissance apparente, période tarifaire en cours).

Les outils de partage (le covoiturage, le partage de biens …). L'expérience de Séoul,

Sur ces secteurs, la collecte s'effectue en principe avec le consentement de l'utilisateur.

5 – LA SECURITE DES ETATS ET LE CRIME ORGANISE

Autre domaine de collecte massive : La sécurité des états (interception des communications numériques notamment...).

Ici la justification pour de la collecte d'information, souvent très privée, est la sécurité. Dans un passé récent, autant il était relativement complexe d'effectuer des écoutes téléphoniques, et surtout nécessitait d'importants moyens humains, autant avec l'avènement du numérique il devient « assez » simple de capter le trafic numérique puis de le confier à des systèmes d'analyse capable de repérer des éléments «clé».

Plusieurs loi récentes, en France, justifient (et «encadrent») le recours à cette collecte, nous nous y arrêtons deux secondes.

La loi 2013-1168 du 18/12/2013 relative à la programmation militaire 2014-2019 en France et ses articles 232 et 246

Art. L. 232-7. I. — Pour les besoins de la prévention et de la constatation des actes de terrorisme, des infractions mentionnées à l'article 695-23 du code de procédure pénale (par exemple et entre autres plus graves: trafic de vol de véhicules volés, sabotage, aide aux séjours irréguliers, contrefaçons et piratage de produits...) et des atteintes aux intérêts fondamentaux de la Nation, du rassemblement des preuves de ces infractions et de ces atteintes ainsi que de la recherche de leurs auteurs, le ministre de l'intérieur, le ministre de la défense, le ministre chargé des transports et le ministre chargé des douanes sont autorisés à mettre en œuvre un traitement automatisé de données.

Sont exclues de ce traitement automatisé de données les données à caractère personnel susceptibles de révéler l'origine raciale ou ethnique d'une personne, ses convictions religieuses ou philosophiques, ses opinions politiques, son appartenance à un syndicat, ou les données qui concernent la santé ou la vie sexuelle de l'intéressé.

Art. L. 232-7. II. — Pour la mise en œuvre du traitement mentionné au I, les transporteurs aériens recueillent et transmettent les données d'enregistrement relatives aux passagers des vols à destination et en provenance du territoire national, à l'exception des vols reliant deux points de la France métropolitaine. Les données concernées sont celles mentionnées au premier alinéa de l'article L. 232-4 du présent code. Les transporteurs aériens sont également tenus de communiquer les données relatives aux passagers enregistrées dans leurs systèmes de réservation.

Art. L. 246-1.-Pour les finalités énumérées à l'article L. 241-2, peut être autorisé le recueil, auprès des opérateurs de communications électroniques et des personnes mentionnées à l'article L. 34-1 du code des postes et des communications électroniques ainsi que des personnes mentionnées aux 1 et 2 du I de l'article 6 de la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique, des informations ou documents traités ou conservés par leurs réseaux ou services de communications électroniques, y compris les données techniques relatives à l'identification des numéros d'abonnement ou de connexion à des services de communications électroniques, au recensement de l'ensemble des numéros d'abonnement ou de connexion d'une personne désignée, à la localisation des équipements terminaux utilisés ainsi qu'aux communications d'un abonné portant sur la liste des numéros appelés et appelants, la durée et la date des communications.

La loi 2015-912 du 14/07/2015 relative au renseignement: polémique dite des «boîtes noires», (finalement validées par le conseil constitutionnel). Ce dispositif prévoit de pouvoir contraindre les fournisseurs d'accès à Internet (FAI) à *«détecter une menace terroriste sur la base d'un traitement automatisé»* en surveillant tout le trafic. En pratique, les services de renseignement pourraient installer chez les FAI ces «boîtes noires» chargées d'examiner les métadonnées de toutes les communications: origine ou destinataire d'un message, adresse IP d'un site visité, durée de la conversation ou de la connexion… Dans le but de détecter des activités «typiques» des terroristes. La loi autorise également la mise en place de fausses antennes permettant d'intercepter les communications de mobiles à proximité.

Pour mémoire, voir aussi la NSA, le programme PRISM et les révélations d'Edward Snowden en 2013.

Le cybercrime s'inscrit également et bien évidemment dans la collecte de données, soit pour des actions de cyberterrorisme, de cyberattaque ou, très prosaïquement, pour cibler et propager le commerce illicite (contrefaçons, armes, drogues...).

Ces cinq exemples illustrent l'existence de dispositifs de collecte de données numériques de plus en plus massives. Ces dispositifs , on le voit, concernent de multiples acteurs, institutionnels, économiques, criminels et sont autant d'occasions de nourrir des bases de données de plus en plus conséquentes...

Les exemples illustrent également le fait que nous sommes tous producteurs de données numériques, consciemment ou inconsciemment, et que le volume des données est considérable, voire humainement inconcevable.

Il s'agit donc d'un potentiel fort de numérisation des comportements humains (masse, variété, durée et possibilité de croisement).

A ce stade, proposons une définition du BigData :

Le ***big data (ont trouve également "mégadonnées")*** désigne des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information.

Il s'agit de données provenant de sources diverses et qui sont enregistrées en vue de permettre leur exploitation et leur analyse sans but prédéterminé et sans limite de temps.

Les données répondent en principe à 4 caractéristiques ; volume, vitesse, variété, valeur (4V) :

- **Volume :** les Big Data représentent d'énormes quantités de données.

- **Vitesse :** les données sont générées, capturées et partagées à une vitesse toujours plus importante Les délais d'actualisation et d'analyse des données sont toujours plus courts et elles sont le plus souvent traitées en temps réel ou quasi réel.

- **Variété** (ou hétérogénéité) : les données analysées ne sont pas forcément structurées. Elles peuvent provenir de sources différentes (et avoir un format différent comme du texte, des images, du contenu multimédia, des traces numériques, etc.) et être combinées entre elles. Des données enregistrées dans un fichier clients interne peuvent

être combinées avec des données externes provenant de réseaux sociaux, de moteurs de recherche, de feuilles d'avis officielles ou de portails de données ouvertes gérés par des autorités publiques.

- **Valeur (ou véracité) :** la dernière caractéristique est la plus-value que l'analyse des données représentent et les usages qu'il est possible d'en faire.

Repère technologique : La difficulté de traitement avec des outils classiques de gestion de BD ou de l'information.

Jusqu'à peu, on pratiquait plutôt un stockage d'information spécifiques, relativement homogènes, stockées en silo et en vue d'atteindre un but précis. On mesure cependant avec ces exemples qu'il peut s'agir d'une collecte «tout azimut», sans but précis, mais en vue de...

Cette nouvelle dimension implique l'émergence d'outils de traitement nouveaux mais aussi de métiers nouveaux autour de l'analyse et de la mise en modèles des données.

Quelques métiers:

- *Data miner (celui qui assure la collecte et le traitement)*
- *Data analyst (organise, synthétise, traduit l'information)*
- *Data scientist (celui qui fait parler la données, les valorise et en tire des indicateurs concrèts)*
- *Chief data officer (le patron des données)*

Quelques outils qui qualifient généralement les infrastructures du Big Data :

- *NoSQL (famille de SGBD qui s'écarte du classique SGBDR R pour relationnelles)*
- *BigTable (SG de base de données compressées, haute performance, développé par Google)*
- *Hadoop (structure logicielle libre ou framework libre facilitant la création d'applications pouvant traiter des pétaoctets - 10^{15} - de données)*
- *MapReduce (architecture de développement informatique conçue par google et permettant de manipuler de très grandes quantités de données en les distribuant dans un cluster de plusieurs machines pour être traitées)*

Pour conclure cette partie : De la difficulté de donner une définition formelle et universelle du Big Data...

Pour conclure cette première partie, mesurons que la définition précédente donnée pour le Big Data s'avère finalement un peu réductrice (ce qui constitue un paradoxe en matière de traitement de l'information) car on comprend à l'aide des exemples que le Big Data est un objet complexe et polymorphe et que sa définition peut varier selon les communautés qui s'y intéressent , soit en tant qu'usager, soit en tant que fournisseur de services, soit en tant qu'institution.

On peut également mesurer que seule une approche transdisciplinaire permet d'appréhender le comportement des différents acteurs du Big Data, qu'il s'agisse des concepteurs et fournisseurs d'outils (les informaticiens), des catégories d'utilisateurs (gestionnaires, entreprises , décideurs politiques, chercheurs), et bien évidemment des producteurs des données, c'est-à-dire, et majoritairement, nous tous.

QUI GERE LES DONNEES COLLECTEES, OU SONT-ELLES LOCALISEES ?

De nouveau, il va être un peu difficile de répondre précisément à ces deux questions du fait de l'immatérialité de la donnée, de l'étendue du réseau sur lequel les données peuvent voyager et de l'implantation géographique des acteurs qui collectent les données.

La confidentialité (secret des affaires, protection des entreprises, économie à l'échelle planétaire) est un autre aspect de cette difficulté à appréhender le sujet.

Enfin, chaque continent, chaque pays, a sa propre perception du sujet d'où une certaine hétérogénéité des réglementations.

En principe - et de bon sens - celui qui collecte la donnée en assure la gestion (stockage, sauvegarde, confidentialité s'il y a lieu) mais il peut acheter ou vendre de la données ou signer des accords permettant l'accès à des données d'autres collecteurs, racheter des sociétés ou outils collectant de la donnée.

On peut le voir dans le cas de l'ensemble des outils de la nébuleuse Google ou du réseau Facebook où les possibilités d'échange, croisements ou cessions de données collectées sont assez importantes et peu transparentes.

De manière générale, il est clair que pour tous ces acteurs de l'Internet, les données doivent être disponibles partout, pour un grand nombre de traitements et d'exploitations et avec une rapidité optimale.

Quelques éléments sur les technologies de stockage et transport:

La technique des fermes de serveurs, interconnectées par des réseaux de télécommunications très rapides permet une répartition géographique en fonction des impératifs de charge et de sécurité (redondance).

OVH, hébergeur français, dispose d'un Data Center sur les berges du saint-Laurent, depuis 2013 (capacité de 360 000 serveurs, servis par un réseau d'une capacité de 2,5 Tb/s). OVH dispose d'un réseau qui comporte 33 points d'interconnexion.

En 2014, on répertorie 3209 centres de données dans le monde dont près de 40 % aux USA (Rackspace, Verizon, IBM). En France 137 (c'est le 4e pays en 2014)

Quelques éléments sur la localisation de la donnée :

Malgré les restrictions liées à la sécurité et à la confidentialité industrielle, quelques éléments sont connus :

Google gère ses propres Datacenters, la firme est assez transparente sur le sujet et indique même sur carte où sont les centres de stockage (en Europe: Dubin[IRL], Eemshaven[PAYSBAS], Saint-Ghislain[BELGIQUE], Hamina[FIN], en Asie : Taïwan, Singapour, aux USA bien évidemment ainsi qu'au Chili).

Microsoft (Cloud OneDrive) gère ses propres Datacenters.

Apple choisit depuis 2016 de stocker certaines des données de ses clients chez Google. Jusqu'alors elles étaient principalement hébergées par AWS (Amazon web services : service de stockage de données proposé par Amazon à des sociétés tierces)

Dropbox : N'a pas de Datacenters, vos fichiers sont stockés chez Amazon (AWS). Les Datacenters d'AWS sont situés en Irlande et à Francfort pour l'Europe, Tokyo, Séoul, Pékin, Singapour pour l'Asie, Sydney pour l'Australie et bien sûr Amérique du Nord et du Sud.

Facebook a construit son premier centre de données à Prineville en Oregon et son premier centre européen en Suède, proche du cercle polaire (Luléa), notamment pour justifier d'économies d'énergie (refroidissement des fermes).

D'après une estimation de la banque d'investissement Jefferies, en 2015, les 10 plus gros acteurs mondiaux de l'Internet ont mobilisé près de 36 milliards de dollars dans des infrastructures de Cloud pour suivre la croissance d'activité sur le web et les mobiles.

Un tout dernier exemple associant stockage et gestion des données importantes et sensibles :

(source Reuters France– 5/06/2015)

Amazon et Google sont en concurrence sur le stockage sur le Cloud de données relatives à l'ADN humain, un marché de l'ordre de 100 à 300 millions de dollars actuellement mais qui pourrait atteindre le milliard de dollars d'ici 2018, estiment des analystes et des consultants. Microsoft et IBM sont également sur les rangs. La croissance de ce segment particulier du cloud est portée par l'évolution vers une médecine dite personnalisée s'appuyant sur l'ADN du patient, ce qui suppose des masses de données gigantesques pour déterminer comment tel ou tel profil génétique répond à tel ou tel traitement.

« Pour l'heure, ce sont les universités et les projets de recherche pharmaceutique qui sont les principaux clients de la génomique en Cloud mais les applications cliniques les supplanteront dans les dix années à venir », pense David Glazer, l'un des responsables de Google Genomics.

Les médecins accéderont donc régulièrement à un service cloud pour comprendre dans quelle mesure le profil génétique d'un patient l'expose à telle ou telle maladie ou évaluer de quelle manière il réagira à un traitement donné. "Nous en sommes à présent à ce stade de transition", souligne Glazer. Les experts de l'ADN et des données affirment tout simplement que sans accès au Cloud, la génomique moderne ferait du sur-place.

Pour conclure cette partie : De la difficulté à savoir...

Finalement, et contrairement à une classique armoire papier ou salle d'archives, nous ne pouvons pas vraiment localiser nos données à chaque instant, elles peuvent se déplacer au gré des nécessités d'hébergement des gestionnaires, mais aussi en fonction des impératifs économiques des groupes qui les collectent.

QUEL INTERET IL Y A-T-IL A LA MISE EN OEUVRE DU BIG DATA?

De toute évidence économique :

Repères :

- *FB a une capitalisation boursière de l'ordre de 300 milliards de dollars (novembre 2015),*
- *Au 31 mars 2016 (source Journal du Net) FB comptait 1,09 milliards d'utilisateurs actifs chaque jour et 989 millions d'utilisateurs actifs sur mobile, toujours chaque jour.*
- *Alphabet (maison mère de Google) a une capitalisation boursière de l'ordre de 540 milliards de dollars (février 2016).*
- *1,4 milliard d'utilisateurs Androïd dans le monde en septembre 2015.*
- *Youtube : plus d'un milliard d'utilisateurs dans le monde (source youtube.com)*

1 : De manière générale, toutes les informations collectées sont susceptibles d'être vendues à des entreprises intéressées → intérêt marchand direct

2 : Compte tenu du nombre important d'utilisateurs, les annonceurs souhaitent être présents sur le média → intérêt publicitaire

3 : Le croisement des habitudes de navigation, de déplacement des usagers (mobilité) et d'usage des outils permet de déduire une personnalité, des comportements donc de cibler des consommateurs → intérêt marketing.

Repère : Google annonce - de manière assez transparente sur son site - qu'il exploite les informations collectées pour « accroître l'efficacité des annonces et transmettre des rapports sur l'activité des annonces aux annonceurs et propriétaires de sites web ».

4 : Certaines sociétés se positionnent d'emblée en tant que data operator (ORANGE FLUX VISION par exemple), en proposant une prestation complète de collecte de données, de transformation des données en information, puis de transformation des informations en connaissance → analyse des comportements, connaissance des habitudes sur un territoire.

On perçoit, en filigrane, que l'analyse des comportements peut mener assez facilement à la modélisation de ceux-ci. On touche ainsi peut-être là le cœur des enjeux économiques car en croisant, analysant et déduisant, ce qui conduit à une modélisation mathématique (on parle d'algorithme prédictif), on peut susciter l'intérêt d'une variété importante d'acteurs économiques, comme les assurances, le marketing, la finance.

Imaginons quelques domaines (source IT-Expert magazine) :

Le marketing prédictif consiste en l'utilisation des algorithmes pour la détermination, par exemple, de la réceptivité des personnes à des textes ou offres promotionnels, du taux de performance de l'objet d'un e-mail publicitaire, ou encore du comportement des consommateurs, en vue notamment de l'amélioration de l'expérience utilisateur et d'une personnalisation des produits et services proposés.

La détermination de scores de risques : risques d'impayés ou de fraudes associés à une commande de produits ou services sur internet ou encore à une demande de crédit à la consommation par exemple, profils à risque parmi les passagers des compagnies aériennes, typologie de client risqué calculée par les compagnies d'assurance… Sur ce point, Facebook a d'ailleurs déposé un brevet qui permettrait aux banques et sociétés de crédits à la consommation d'évaluer le risque associé à un client en fonction de ses amis sur le réseau ou de son comportement d'achat. (brevet déposé le 04/08/2015 intitulé « autorisation et authentification basée sur le réseau social de l'individu »).

Repère : *Des preuves ?*

- *16/07/2015 – Source Les Echos.fr : BNP Paribas s'allie à Facebook, Twitter, Google et LinkedIn. Dans le détail, l'accord avec Google aura pour but de « renforcer l'accessibilité des services de la banque ", notamment en maîtrisant mieux les formats mobiles. Cet accord vise aussi à améliorer « la pertinence de ses offres en fonction des attentes des clients ". Le partenariat*

avec Facebook est sans doute le plus spectaculaire. Il a pour objectif de prolonger l'exemple de certaines filiales, telles que TEB en Turquie, qui propose l'ouverture de comptes sur Facebook. Avec Twitter, BNP Paribas compte mettre l'accent sur l'utilisation des données publiques du réseau social. Quant à LinkedIn, il s'agit d'améliorer son offre pour recruter.

- *29/005/2015 – Source Cbnews : BPCE accélère sa transformation digitale en signant un partenariat stratégique avec Facebook, partenariat qui sera utilisé et pour les clients et pour l'interne. Facebook fournira au Groupe BPCE un accès à ses outils, technologies et innovations ainsi qu'un support personnalisé avec la contribution active de ses équipes techniques, créatives et marketing pour développer plusieurs axes de collaboration.*

Mais aussi :

Le domaine culturel : prédiction de la qualité littéraire d'un roman, de la rentabilité d'un film cinématographique, du succès de séries à proposer par les services de VOD ou encore des nominations aux oscars.

La recherche de l'amélioration de la qualité de vie : anticipation de l'affluence dans les transports en communs ou dans tout autre lieu ouvert au public, localisation de places de parking à une heure et dans un lieu donnés, amélioration des déplacements, économies d'énergie (Linky), services d'assistant personnel prédictif, prédiction sur les prix des billets d'avion afin de déterminer si un prix va augmenter ou baisser ou encore à quel moment les prix seront les plus bas…

La médecine prédictive : à titre d'exemple, les algorithmes prédictifs sont utilisés pour la réalisation de diagnostics génétiques en vue de prédire quels seront les patients qui ont le plus de probabilité d'être atteints d'une pathologie donnée, ou encore pour déterminer les mutations génétiques impliquées dans le développement de certaines maladies. Pour une autre illustration, des chercheurs sont parvenus à créer un programme d'analyse linguistique ayant pour objectif de détecter chez un patient les risques de développer une psychose. Les services d'assistance et d'urgence se lancent également dans cette tendance, avec le déploiement d'algorithmes prédictifs permettant de prédire la gravité des blessures des occupants d'une voiture accidentée par exemple afin d'anticiper et d'organiser de manière adéquate l'arrivée des secours ou encore l'admission aux urgences d'un établissement hospitalier.

La police prédictive : les algorithmes prédictifs sont notamment utilisés par les forces de police et de gendarmerie en vue de déterminer les relations sociales entre des personnes dans le cadre d'enquêtes policières, et devraient à terme être utilisés pour prédire la perpétration de crimes ou de délits (des expériences sur ce point étant déjà menées à l'étranger).

Repère sur les modèles prédictifs : l'apprentissage automatique (deep learning : le programme alphaGo vs Lee Sedol) à comparer à la puissance de calcul d'il y a 20 ans (deep blue et Kasparov).

Repère : l'OpenData partie du BigData (open data = ouverture des données publiques numériques par les administrations)

- *Tranquilien : application pour smartphone qui permet de connaître à l'avance le taux d'occupation de ses trains en Île-de-France. Elle représente une première concrétisation de l'initiative de la SNCF dans l'open data. Les voyageurs équipés de smartphones sont également mis à contribution puisqu'ils peuvent informer le service de la fréquentation de leur train en temps réél. Pour fonctionner, Tranquilien s'appuie sur un algorithme qui utilise les données de la SNCF. Pour la gestion de son réseau, l'opérateur de transports utilise en effet des statistiques sur le trafic, collectées ces dernières années. Mais pour affiner les résultats, l'application propose également à l'utilisateur d'indiquer en temps réel la fréquentation de son train. Une interface collaborative permet ainsi aux voyageurs de corriger en direct des prévisions qui se seraient révélées fausses et fournir une meilleure information aux passagers des gares suivantes sur un trajet donné. L'application utilise aussi des données de localisation des utilisateurs mais assure garantir l'anonymat des contributeurs.*

- *HandiMap, la belle histoire de l'Open Data de Rennes : Rennes a mis à disposition les données géographiques de 80000 points de trottoirs surbaissés et deux développeurs ont décidé de créer une application proposant des itinéraires pour personnes à mobilité réduite. La valeur ajoutée de l'application est très claire pour les habitants de Rennes ou pour les gens de passage dans la ville. La valeur ajoutée pour la ville qui a libéré ses données est claire aussi : le développement de cette application en interne aurait été un chemin long et compliqué. Il aurait déjà fallu avoir l'idée, prendre la décision de la développer, allouer un budget, embaucher un développeur, etc. Quelques jours ont suffi à la mise à disposition de ce service.*

Les promesses du BigData sont donc multiples : commerce, santé, aménagement du territoire, environnement, transports, sécurité ...

Pour conclure cette partie, en matière de BigData, si le pacte semble à la hauteur, attendons de mesurer tous les effets et surtout, vérifions que le résultat est bien à la hauteur des enjeux...

LES DONNEES PEUVENT ETRE PERSONNELLES, QUELS SONT LES RISQUES DE DERIVES?

Et nous dans tout ça...

Le côté vertigineux du BigData, c'est qu'il est insaisissable, que ses contours sont flous, qu'il est international… et qu'il nous concerne tous.

Au delà des avancées réelles ou supposées que permet le BigData, il y a, me semble-t-il, à travers cette collecte massive de données numériques, une exigence de transparence toujours plus importante de l'activité humaine au travers d'une lisibilité accrue de l'activité de chaque individu. Cette évolution ne devrait cependant pas impliquer, toujours à mon sens, qu'on fouille sans scrupule la sphère privée.

Avant tout, celui qui procède à la collecte et l'analyse de données en mode « Big Data » devrait faire preuve de bonne foi et de transparence et prendre toutes les mesures utiles pour garantir autant que possible l'anonymat des données, leur sécurité et le respect de la sphère privée.

> *Repère : Est-bien garanti lorsque Vinton Cerf, (Chief Internet Evangelist chez Google et co-inventeur du protocole TCP/IP au début des années 1970) déclare - en novembre 2013 - que « La vie privée peut être considérée comme une anomalie ... » Il estime qu'« il sera de plus en plus difficile pour [nous] de garantir le respect de la vie privée ».*
>
> *Repère : le mouvement transhumaniste(Google en est un des principaux sponsors et emploie dans son équipe dirigeante Raymond Kurzweil, spécialiste de l'intelligence artificielle et théoricien du transhumanisme) prône le développement de l'IA pour accroître la longévité et le confort de l'espèce humaine. Et le développement de l'IA peut passer par les modèles prédictifs...*

Alors, des normes de protection des données sont-elles applicables ? Que dit le droit en la matière ?

En France comme dans de nombreux endroits dans le monde la protection des données nominatives d'une personne est sensée être assurée (France, voir CNIL loi du 6 janvier 1978 consolidée, droit à l'oubli, consentement de l'utilisateur requis ainsi que la loi LCEN de confiance en l'économie numérique) mais on mesure vite que le BigData pose un défi à cette protection puisque lorsque des données anonymes sont combinées à d'autres données, elles peuvent rapidement rendre les personnes parfaitement identifiables.

> *Repère : Les quasi-identificateurs - des combinaisons d'attributs, comme la date de naissance, le sexe et le code postal - doivent être traités avec précaution. Des scientifiques américains ont montré que les quatre cinquièmes de la population américaine pouvaient être identifiés a posteriori sur la seule base de ces trois caractéristiques.*

Sans compter que le consentement, lorsqu'il est requis, est obtenu via des clauses incompréhensibles qu'on accepte souvent sans lire et qui permettent au gestionnaire des applications d'utiliser finalement nos données comme il l'entend.

A noter aussi que toutes ces applications pour smartphones, certaines très pratiques, ne sont malheureusement accessibles que si l'on accepte un nombre invraisemblable d'accès aux données privées de nos smartphones et tablettes (contacts, sms, position GPS, bluetooth, photos, achats).

Et, bien évidemment, le Cloud et l'activité économique mondiale qui peuvent quelquefois si facilement s'affranchir des barrières et législations nationales.

> *Repère : Février 2016, la CNIL fixe un ultimatum à Facebook et se lance dans un « bras de fer » juridique (voir première partie sur Facebook et les cookies de visite de pages publiques. La belle affaire... l'amende la plus élevée est de 150 000 €.*

On peut donc être légitimement sceptique sur la possibilité d'établir un droit international associant BigData et respect de la vie privée.

Citons néanmoins quelques aspects importants associant données massives et protection juridique des personnes :

- Les possibilités techniques du BigData présentent un grand défi en ce qui concerne l'**exigence de transparence** (des dispositifs, pas de l'individu, cette fois). Chacun a le droit de savoir quelles sont les données le concernant qui sont traitées, par qui et dans quel but. Dans le cas des données massives, le traitement et la connexion de données provenant de différentes sources est très opaque et difficilement vérifiable par les personnes concernées. Les utilisateurs de données massives doivent donc être particulièrement vigilants quant à la transparence du traitement et à l'information des personnes concernées.

- Le traitement de données massives à caractère personnel requiert le **consentement** des personnes impliquées. À cet égard, le **but** des procédures impliquant des données massives doit pouvoir être connu clairement et sans ambiguïté par les personnes concernées, et ceci dès la collecte des données (Cette approche contredit toutefois un principe de collecte massives, celui qui implique la constitution de stocks de données qui serviront ultérieurement à un but non encore déterminé).

- L'exigence de l'**exactitude des données** constitue une difficulté supplémentaire : les algorithmes appliqués aux données massives analysent de grandes masses de données de manière autonome, automatisée, à la recherche notamment de corrélations. Ces procédures d'analyse créent de nouvelles informations liées à des personnes, sans qu'il soit possible de les qualifier d'exactes ou de fausses, puisqu'elles ne constituent que des probabilités ou des interprétations.

Et nos comportements ?

Nous avons encore une habitude des données numériques (au moins celles que nous fournissons consciemment) que nous pensons traitées dans l'unique but pour lequel elles ont été collectées et qui doivent en principe être détruites une fois ce but atteint. Les Big Data reposent au contraire sur l'exploitation de données dans d'autres buts, voire dans la conservation de données pour une utilisation ultérieure éventuelle et dans un but non encore déterminé.

Nos comportements doivent donc s'adapter.

Lorsque nous envoyons (partageons) une information (de localisation, ou personnelle, ou like, ou avis, ou commentaire...) sur l'Internet, lorsque nous confions des informations personnelles, nous devrions toujours savoir pourquoi nous le faisons et si cela revêt une quelconque utilité pour nous, en tant qu'individu isolé ou pour les autres en tant que groupe social.

Paradoxalement, l'Internet (et donc quelque part le Big Data) est là pour nous aider dans cette démarche car il constitue une source de connaissance inépuisable – à condition de croiser les informations et de séparer le bon grain de l'ivraie. De la même manière, l'Internet peut constituer un moyen de véhiculer massivement et salutairement toute information objective relative à des dérives d'acteurs du BigData.

Connaissance, décryptage et vigilance me semblent donc dans l'immédiat les éléments essentiels qui peuvent nous permettre de minimiser raisonnablement l'exposition au BigData, de comprendre comment il fonctionne et qui est susceptible de profiter de ce BigData... peu ou prou !

- FIN -