Big Data : données personnelles

École Polytechnique- Palaiseau

Jeudi 15 octobre 2015

Coordination Scientifique

Patrick Legand (Xirius Informatique)

Pascal Alix (Virtualegis)

Thiên-Hiêp Lê (ONERA)

Editorial Board

Actes du séminaire du 23/09/2015

Big Data: données personnelles

Séminaire Aristote, 15/10/2015 à l'École Polytechnique

Coordination scientifique et organisation:

Patrick Legand (Xirius Informatique)

Pascal Alix (Virtualegis)

Thiên-Hiêp Lê (ONERA)

Sommaire

Compte-rendu des interventions

Introduction

Patrick Sitbon (Thales) débute la journée devant un parterre de 63 personnes, dont des étudiants. Il présente l'association Aristote dont il assure la présidence. Aristote est une association indépendante qui a pour vocation de favoriser l'innovation dans le numérique. Les adhérents viennent du monde académique, des services, de l'industrie. Elle réfléchit aux nouveaux usages qu'impliquent ces technologies. Elle organise des séminaires comme celui d'aujourd'hui, a des activités dans le domaine de la formation et réfléchit par des groupes de travail sur les thématiques qu'a choisi Aristote. Le séminaire d'aujourd'hui est atypique. Patrick Sitbon remercie Pascal Alix, Patrick Legrand et Thien-Hiêp Lê pour avoir organisé la journée et la souhaite fructueuse pour tous avant de donner la parole à Pascal Alix.

Thématique

Pascal Alix (Virtualegis) présente le thème de la journée et commence par remercier les futurs intervenants. De son point de vue, le séminaire tranche sur les autres de par son thème, car si le big data est magnifique, il est aussi inquiétant. Il interroge l'humain, l'éthique, la philosophie. La journée est donc un panorama le plus objectif possible sur le big data. Pascal Alix, qui est avocat, est naturellement sensible aux droits sur les données personnelles. Il rappelle l'article 1er de la loi 78-17 du 7 janvier 1978 : l'informatique doit être au service du citoyen; son développement doit s'opérer dans le cadre de la coopération internationale; elle ne doit porter pas atteinte à l'identité humaine ; les clefs de décision ne doivent pas être confiées au traitement informatisé et automatisé des données.

Il poursuit sur les lacs de données qui sont abondés par les gestes personnels (inscription sur des comptes en ligne, achats, paiements, publications). Mais d'autres données sont fournies par les objets qui nous entourent. Bientôt les voitures, les maisons, les réfrigérateurs vont être connectés et fournir des données.

Le big data, c'est du volume, de la vélocité, de la variété. Parfois on parle de véracité et de valeur ajoutée. Ces 3 ou 5 « V » se confrontent au droit et ses 3 « P » : prévisibilité, prudence, protection (intimité, vie privée, secret professionnel, non discrimination).

Les grands principes d'aujourd'hui sont la loyauté de la collecte des données, la limitation des finalités, la proportionnalité de la collecte, l'exactitude et la pertinence des données traitées, la limitation de la durée de la conservation, l'obligation de sécurité et le respect des droits. Mais demain, avec le RGPD (règlement général de protection des données) prévu pour 2018, cela va changer. Ce nouveau règlement va renforcer la protection des droits, mais aussi ouvrir à des innovations. Le principe nouveau est la nécessité d'une approche *privacy by design*. Il faut réfléchir à l'utilisation des données dès le début. Il faudra aussi effectuer une étude d'impact en cas de risques spécifiques. Le RGPD est un encouragement à la co-régulation.

L'affaire Volkswagen pose le problème des systèmes d'informations et de la fiabilité des algorithmes. Pascal Alix rappelle que selon F. Pasquale, de l'université de Yale, la neutralité des algorithmes n'existe pas et leur manipulation une réalité.

Il faut donc une approche pluridisciplinaire. Le big data n'est pas que de la data science. Dans une entreprise, le chef de projet big data ne devrait pas être un informaticien. Il doit comprendre des personnes de la direction de l'entreprise, le CTO, le CDO quand il y en a un, le RSSI, le responsable juridique et tous les métiers impactés par le projet.
Enfin, les projets big data fonctionneront s'ils sont acceptés par toutes les équipes de l'entreprise et par les usagers. Même les personnes « *digital native* » ont conscience que la confiance est primordiale. Il y a bien actuellement une prise de conscience des excès de la collecte de données.

1. Big data: du potentiel technologique aux écueils pour l'individu et la société

Vincent Corruble (Université Paris 6)

Vincent Corruble fait partie de l'équipe SMA du LIP6, le laboratoire d'informatique de l'université Paris 6. Il a fait le premier cours sur le *data mining* en Ile-de-France dans les années 2000. Les origines du big data remonte au début des années 1990 quand les scientifiques ont commencé à interagir. Puis les grandes entreprises comme IBM ont investi le domaine. Tous les ans, une grande conférence internationale (KDD) est organisée. À partir des années 2000, de plus en plus d'entreprises investissent le sujet. La collecte des données se généralise. Les organismes d'Etat (renseignement, santé) ont rejoint le mouvement.

L'aspect positif est la dimension économique. Il est apparu quand les scientifiques n'ont plus été seuls dans le *data mining*. Au début des années 1990, on a commencé à faire du *text mining* et du *web mining* pour comprendre, prédire et anticiper les comportements.
Vincent Corruble donne un premier exemple en Ecosse sur les données de patients en soins intensifs après des traumatismes crâniens sévères. Il fallait tirer partie des données médicales du patient et de ses données personnelles. L'étude a permis de mieux comprendre le phénomène qui suit le traumatisme crânien. Un deuxième exemple est celui de France télécom R&D. Au début des années 2000, on essayait de comprendre le comportement des internautes. Comment un site est visité, comment caractériser le visiteur d'un site. Cela s'est fait à partir des données de navigation et des données personnelles pour une réflexion globale.

D'un point de vue éthique, la perspective acceptée est un échange entre des données personnelles et un service. Cet échange est sensé être gagnant-gagnant. Mais vu le deuxième exemple, les données personnelles servent aux entreprises pour les appliquer à la société afin de construire un modèle à valeur générale (tous les patients, tous les administrés d'une ville). Les autres individus n'ont pas choisi d'être classés dans des catégories et n'obtiennent pas de service en échange.
Pour diminuer les risques des big data, Vincent Corruble propose quelques pistes de réflexion. Il faut intervenir sur quelques étapes du data mining (collecte des données, stockage, traitement, interprétation, exploitation). Or la Cnil s'est surtout intéressée au stockage. Mais la réflexion est sans doute plus facile au niveau de l'exploitation. On peut aussi intervenir au niveau culturel, sur le comportement des personnes, l'éducation ou la loi. Il faut une meilleure prise de conscience de son environnement informationnel et des conséquences sociétales des comportements individuels..

2. Aspects opérationnels, expérimentaux du Big Data en entreprise

Hugues Le Bars (NEOPOST)

On entre dans l'opérationnel avec l'intervention de Hugues Le Bars. Il est *Chief Data Officer* chez Neopost, créée en 1926. Neopost (CA de 1 Md d'euros) est en pleine transformation digitale avec une baisse des envois de lettres et de colis et une montée du numérique. Pour Hugues Le Bars, les big data ce sont les 5 V (volume, vélocité, variété (son, vidéo, texte), véracité afin de générer de la… valeur). La donnée a un comportement. Elle s'échangeait, elle devient une donnée d'interaction avec le web 2.0. L'information est descriptive et doit être en temps réel afin d'être au plus près de l'actualité. Elle doit être prédictive et même prescriptive afin de prévoir les comportements des clients. Les disciplines que le big data implique sont la visualisation, la science et l'ingénierie. Le big data peut expliquer le quand, le comment, le qui, mais pas le pourquoi. C'est toujours l'humain qui décide.

Le lac de données a commencé en 2004. C'est une nouvelle façon de stocker des données. Elles sont copiées une seule fois. En 2006, Yahoo lance Hadoop qui devient en open source en 2009. Mais le lac de données doit être suivi d'une raffinerie. Les entreprises vendent du support. Houtonworks data plateform trie les données et sort de l'information qui sert à décider. Le *chief data officer*, dont le rôle n'est pas encore bien défini, commence à imprégner les entreprises. Il cherche une organisation pluridisciplinaire, regarde tous les métiers et surtout les sciences cognitives (psychologie, philosophie, linguistique, anthropologie, neuroscience, informatique).Tous ces domaines interviennent dans les différentes étapes du *data mining*.

Le monde est volatil, incertain, complexe, ambigu. Il faut trouver la bonne innovation qui résous un besoin. Par exemple, comment trouver un taxi quand on n'en a pas (Uber), comment diffuser des films quand on a pas de salles (Netflix), comment monter des hôtels quand on en a pas (AirBnB). À valeur égale, ces sociétés utilisent 10 à 100 fois moins de personnels que les autres. Pour réussir, il faut sortir de la zone de confort, il faut de l'audace. Le chemin pour y arriver est déjà d'optimiser ce qu'on fait, trouver de nouveaux modèles économiques. Mais attention, il peut y avoir des ratés. Par exemple, le cybermarché Webvan : tout fonctionnait (1milliard de dollars, 1600 personnes) mais il

n’y avait pas de clients. Ils n’ont pas collecté les données pour voir s’il y avait un besoin. Pour connaître le succès, il faut aussi non seulement former les gens au big data, mais aussi un modèle économique *data driven* qui doit pouvoir contrôler les flux et garder une réciprocité avec le clients.
Le big data est une intelligence ambiante, mais Claude Levi-Strauss dit qu’il y a des coalitions de culture, chacune préservant son originalité.

Pour conclure, il faut être vigilant sur l’acquisition des données, vérifier les algorithmes et surtout, éduquer les jeunes générations.

3. Environnement juridique des projets Big Data : comment concilier principe de précaution et innovation ?

Célia Zolynski (Université Paris-Saclay)

Célia Zolynski est professeur de droit privé à l'Université de Versailles-Saint-Quentin. L'impact du big data est pour elle une préoccupation importante. On a décrit les transformations sociétales et économiques qu'il implique, le meilleure gestion des ressources, de la santé des publics… mais il y a aussi des risques important car la finesse d'analyse du big data permet le profilage fin. Avec des possibilités de censure et de discrimination. Il y a aussi un risque de dictature des données. Ces risques sont bien réels. L'enjeu est de taille surtout que le domaine est en pleine évolution. La définition des données personnelles est fondamentale car elle conditionne la loi. Ce n'est pas facile car les big data sont très variées et toute donnée a de la valeur. L'objectif n'est pas forcément d'identifier des personnes. Mais le caractère collectif du big data peut amener à l'analyse de personnes déterminées. Il y a plusieurs approches : tout traitement peut être considéré comme une donnée personnelle. Pour certains, au contraire, toute donnée personnelle peut être sujette à traitement, afin de ne pas brider l'évolution du big data. Il ne faut pas faire des principes mais réagir au cas par cas et voir si une donnée est personnelle ou si le traitement conduit à une donnée personnelle. Ce qui crée une incertitude sur le droit. Elle se ressent également sur les principes de la législation (consentement, finalité, proportionnalité). La Commission Nationale Informatique et Liberté (Cnil) réaffirme la robustesse de ces principes. Mais des ajustements sont à opérer, ce que tente de faire le Conseil d'État. Faut-il changer d'approche et se concentrer sur l'usage plutôt que sur la donnée ? Il faut dépasser l'approche statistique de la donnée que retient le droit français et européen. Dans son rapport, le Conseil d'État donne l'exemple de l'utilisation des capteurs dans les voitures ou des usages comme la publicité comportementale. On devrait considérer que les usages doivent faire l'objet de réglementation. Penser en terme de préjudice permettrait de libérer les données à des fins de recherche et encadrerait le profilage marketing.

Il convient de pouvoir gérer le risque informationnel. Il faut définir les obligations et la responsabilité des opérateurs et promouvoir une responsabilité pour risque. Les opérateurs doivent savoir évaluer les risques au cours des différentes étapes du big data et les minimiser. C'est une approche pragmatique au cas par cas que préconisent les autorités de régulation. L'obligation de vigilance va

être prolongée par un devoir de documentation, tel que les sociétés cotées doivent la faire pour les risques environnementaux.

Pour mettre en place ces obligations, il faut ajouter des préconisations sectorielles à la législation. Par exemple, une obligation de vigilance à la charge des opérateurs. Le principe de précaution peut être invoqué.

Il faut aussi protéger les utilisateurs. Le législateur doit veiller à protéger les personnes concernées et à ce qu'ils puissent être un contre-pouvoir à l'utilisation des données. On étudie aussi des outils d'*empowerment* comme la portabilité des données, ce qui permettra une meilleure répartition des bénéfices du big data.

La réflexion est en marche pour de nouvelles solutions comme Célia Zolynski le fait dans le cadre de D@nte à l'Inria. La réflexion doit être collective et impliquer des membres de la société civile. Il faut aussi savoir évaluer le préjudice actuel ou futur pour un individu ou un groupe d'individus comme cela se fait avec les risques environnementaux.

4. Accompagnement de projets Big Data (certification de processus industriels, établissement de pack conformité, par type d'usage)

Sophie Nerbonne (CNIL)

Sophie Narbonne, directrice de la conformité à la Commission Nationale Informatique et Liberté (CNIL), revient sur le discours des orateurs précédents en signalant que la vigilance à avoir sur les données porte sur les usages. La Cnil est l'autorité française en charge de la protection des données personnelles. Elle a été créée en 1978 pour veiller à l'application de la loi Informatique et liberté. Toute personne peut saisir la Cnil lorsque ses données ne sont pas traitées conformément aux règles applicables. Le régime juridique a été adopté au niveau européen pour assurer un haut niveau de protection. Dans le monde, si de nombreux pays ont des autorités de protection des données, le niveau de protection est souvent plus faible. Ce 15 octobre, à Bruxelles, les autorités de protection des données se réunissent afin de réfléchir de façon coordonnée. Le Big data semble toucher tous les domaines : marketing, e-commerce, assurance, lutte contre la fraude, tourisme, ressources humaines, sciences, santé, etc. La grille d'analyse de la Cnil s'appuie sur ses 5 règles d'or (finalité et proportionnalité, pertinence des données traitées, conservation limitée, sécurité et confidentialité, respect des droits des personnes). Pour les Cnil européennes (G29 de septembre 2014 et groupe de Berlin de mai 2014), le big data doit se développer en conformité avec les principes fondamentaux de la protection des données. Mais une approche ouverte semble nécessaire. Si le recours à des techniques d'anonymisations est utilisé pour sortir de l'application de la loi, il faut mettre en place des procédures robustes d'anonymisations.

Un exemple concret, le partenariat FIEEC (fédération des industries électriques, électroniques et de communication)-Cnil : un groupe de travail a été créé en 2012 pour aboutir à la publication d'un référentiel de bonnes pratiques en intégrant le *privacy by design* (intégrer la protection des données dans toutes les étapes). Cela a concerné les appareils connectés de domotique en aval des compteurs électriques, mais ce référentiel a été repris par les industriels automobiles. Le groupe a différencié 3 usages. : le In-In (si les données collectées dans le logement restent dans le logement (thermostat-chauffage)) ; le In-Out (quand elles sortent à l'extérieur (consommation, bilan énergétique)) ; le In-Out-In (quand les données collectives reviennent chez les utilisateurs). Les règles

à mettre en place sont différentes. Cela a fait un « pack de conformité énergie », qui peut être appliqué à d'autres domaines (assurance, banque). L'ensemble de ces recommandations va être porté au niveau européen, afin d'avoir des outils permettant d'atteindre un marché unique numérique en Europe.

5. Ethique des Big Data dans le domaine de la santé

Jérôme Béranger (Université de Marseille)

Jérôme Béranger est chercheur à l'Université de Marseille. Sa recherche est la médecine 3.0 et le big data. Le online a remplacé le silicium, tout passe dans l'infosphère. La médecine est devenue celle d'e-ppocr@te. La médecine 3.0 est personnalisée, prédictive, préventive et participative. Contrairement aux autres orateurs, Jérôme Béranger ajoute un 6ème V à la définition du big data : la visualisation. Il y a des enjeux éthiques à cette médecine 3.0. que sont la culture, la déontologie, le secret médical. Jérôme Béranger se pose des questions : la numérisation remet-elle en question certaines valeurs et principes humains ? Quel droit d'accès aux données ? Quel contrôle du citoyen, quelles garanties de confidentialité ? La technique thérapeutique n'était pas reléguée au second plan ? Comment intégrer les systèmes experts ? Le médecin deviendra-t-il un data manager ? Est-ce choquant ? Pas forcément, selon lui. Il faut savoir contrôler et encadrer. Mais des risques émergent : quelle valeur accordée à la pertinence scientifique ? Quelle place au jugement du praticien ? Quelle sécurité des données médicales ?

Pour répondre à ces questions, on peut avoir un cheminement éthique. Il s'agit d'évaluer les conduites humaines en les appuyant sur un système de valeurs, d'exigences, de principes. Il y a l'éthique descriptive et appliquée, fondée sur les finalités, l'éthique normative basée sur les normes et l'éthique réflexive centrée sur les principes et les valeurs morales. Cela peut être appliqué aux données, aux algorithmes et aux pratiques. Les valeurs universelles de l'éthique sont les principes d'autonomie, de bienfaisance, de non-malfaisance et de justice. Ces principes sont associés à des émotions (respect, compassion, crainte, indignation). À partir de ce travail, Jérôme Béranger a mis au point un modèle d'analyse éthique en cancérologie. Il a appliqué les principes à des paramètres environnementaux (structurel, stratégique, organisationnel, relationnel). Avec cela, il construit des indicateurs de qualité des big data, élabore une cible éthique des big data et peut suivre la valeur des données personnelles suivant différents axes. On peut associer des actions à chaque valeur afin de les améliorer.

Ainsi la dimension technique est compatible avec l'humain. Les big data sont les liens relationnels entre les acteurs. L'éthique est le garde fou de l'usage des big data contre la déshumanisation. Il ne faut pas avoir peur des big data du moment qu'elles sont encadrées éthiquement.

6. Big data et protection des données personnelles : enjeux et limites du règlement européen

Florence Bonnet (CIL Consulting)

Florence Bonnet, fondatrice de la société de conseil CIL Consulting, est aussi correspondante Informatique et Liberté. Elle a une approche critique du nouveau règlement européen. Elle s'y perd et imagine que les entreprises seront encore plus perdues. Si elle est critique, c'est qu'elle se demande si on peut réguler le big data. C'est un changement brutal et violent, une révolution. C'est un déluge de données occasionné par le nombre de capteurs bon marché, si ce n'est gratuit. Le big data, c'est aussi des algorithmes qui impactent l'individu. En Chine, chaque individu est observé afin d'améliorer la société. Le big data est une révolution, car il multiplie les systèmes sans intervention humaine. Enfin il est imprévisible. Avec le big data, les sources de risques sont plus nombreuses et plus graves. D'une part, tout devient connecté, d'autre part l'univers des big data est celui de start-up pour qui les questions éthiques viennent en second. De plus, on a davantage de données qui vont devenir plus sensibles. Enfin les événements deviennent plus graves. Les technologies de suivi et de profilage sont opaques.

Le cadre juridique doit donc évoluer. Même s'il y a un consensus mondial, personne ne fait pareil. Le nouveau règlement est indispensable mais c'est un frein à l'innovation et il y a un risque de détournement de la loi. Comment alors prévenir des menaces non identifiées ?
Pour Florence Bonnet, la règle sera applicable aux entreprises qui offrent des services aux citoyens européens ou qui utilisent des données des citoyens européens. Mais les États ont-ils les moyens de contrôler les géants du net, alors qu'ils en sont dépendants ?

Le nouveau texte intègre les données de localisation, de génétique. Il ne s'applique pas aux données anonymes mais le big data remet en cause la distinction entre données personnelles et non personnelles. La cible n'est pas l'individu, mais le groupe auquel appartient un individu. La publicité qui cible les gens en surpoids les prend quand elles sont les plus vulnérables. La frontière entre anonymisation et données personnelles est très légère. Il y a un risque de re-identification dans les données anonymes.

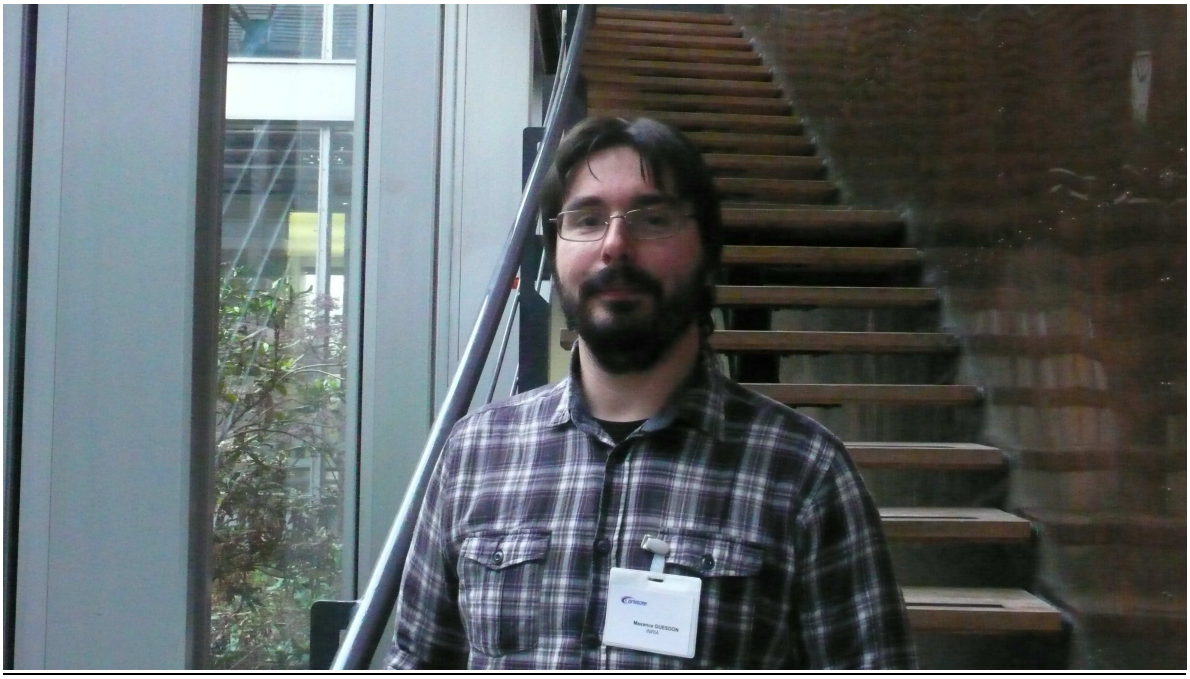
Le principe de finalité spécifique, qui est dans la loi actuelle, est aussi dans le nouveau règlement. Mais comment faire alors que le big data permet de découvrir des inférences qui n'étaient pas connues ? Toutes les données sont donc ainsi potentiellement pertinentes. Le règlement dit qu'il faut faire un test de compatibilité entre la finalité initiale et la nouvelle. Avec une nouvelle base légale, un traitement d'incompatibilité pourra être réalisé avec le consentement des personnes. Sauf si l'opérateur prouve qu'il a un intérêt légitime à devoir faire son nouveau traitement. Le profilage basé sur des données sensibles est interdit. Sauf que des données peuvent devenir sensibles. De plus, Florence Bonnet se demande si le consentement, quand il est obligatoire, est vraiment éclairé.
Un autre article du nouveau règlement est celui sur l'impact sur la vie privée. La Commission, le Parlement et le Conseil européen sont d'avis un peu différents. Par exemple, contrairement aux autres, le Conseil appuie sur la discrimination, les risques de fraudes et d'usurpation d'identité. Mais comment faire une étude d'impact sur chaque application alors qu'on prévoit 50 billons d'objets connectés en 2020 ?

Plusieurs questions se posent : a-t-on le choix de refuser son consentement ? Quid des personnes vulnérables ? Comment informer, alors que personne ne lit les politiques des différentes applications ? Il y a donc des limites pour protéger le citoyen.
Autre sujet, le droit à l'oubli et à l'effacement est une mission impossible. C'est pourtant un principe de base. Le droit à l'effacement sans délai est acté.
Pour conclure Florence Bonnet évoque quelques idées : il faut des incitations mais aussi des contrôles a posteriori et des sanctions dissuasives (les discussions sont entre 2 et 5% du CA de l'entreprise). Il faut aussi impliquer les professionnels responsables et éviter les lobbies. Il serait bien de délivrer des gages de confiance (certification, labels).

7. Appariement de données médico-sociales: techniques et organisations

Maxence Guesdon (INRIA)

Avec Maxence Guesdon, on parle technique big data. L'exemple choisi est le monde médical, dans lequel on a besoin d'apparier les données sociales et médicales en vue de recherches d'antécédents, d'études statistiques ou épidémiologiques. Pour conserver l'anonymisation, l'équipe de l'Inria auquel appartient Maxence Guesdon « pseudonymise » la donnée de façon aléatoire. Mais l'appariement nécessite que les données ne soient pas complètement anonymes.



La technique utilisée est le « hachage ». On calcule une empreinte de taille fixe à partir de données de n'importe quelle taille. Le risque de collision (même empreinte pour deux personnes différentes) est quasi nul. L'attaquer de ce hachage est possible mais on peut s'en prémunir avec une clé secrète qui perturbe l'empreinte. L'inconvénient d'une pseudonymisation pour l'appariement est que la moindre différence donne deux empreintes radicalement différentes. L'appariement (le chaînage) consiste à éviter les collisions (appariement de deux personnes différentes) et son contraire, à ce que deux enregistrements d'une même personne correspondent à deux personnes différentes. Pour éviter cela, on utilise la méthode statistique de Jaro avec calcul de poids, selon les noms, prénoms et date de naissance. Mais il faut faire attention aux seuils de calculs. Selon qu'ils soient trop hauts ou trop bas, on peut soit apparier tout le monde même les collisions, soit au contraire prendre trop de précautions et rater des appariements. Cette technique de chaînage est très artisanal et dépend des besoins, des données et de leur qualité, et des possibilités de vérifications.

Une autre technique est le chiffrement. Le message est codé et lisible uniquement par ceux qui ont la clef. L'inconvénient est qu'on peut revenir en arrière, tout en sachant qui a la clef. Pour protéger l'anonymat, il faut compartimenter les données, en séparant les données identifiantes des autres afin de maitriser les appariements. La technique est de confier le hachage des données à deux acteurs différents (avec génération d'une clef unique) alors qu'un troisième acteur apparie les données identifiantes. Il donne une clef à chacun des deux premiers producteurs. Ce qui fournit des tables avec numéro d'index. L'organisme utilisateur récupère les données chiffrées des deux producteurs mais ne peut pas remonter aux données de base.

Pour conclure Maxence Guesdon revient sur la loi sur le numérique. Selon lui, on pourra faciliter les études statistiques pour la recherche tout en protégeant la vie privée. Mais la confiance n'exclut pas le contrôle.

8. Quelle sécurité pour les données personnelles dans le contexte mondialisé des Big Data ?

Patrick Legand (Xirius)

Patrick Legand est Directeur de Xirius Informatique, entreprise spécialisée dans les systèmes de cyberdéfense et de sécurité. Nos gestes disséminent des traces en permanence. Les mouchards sont partout. Ce sont les ordinateurs ou les smartphones (adresse IP, clics, vidéos, messageries, déplacements, cookies, etc.) et les applications (Facebook, Google, plate-forme vidéos, jeux, pass Navigo). Il y a aussi des « ratés » (déclaration d'impôts numériques, bulletins de paie, relevé de compte bancaire, dossiers médicaux, photos). Cela fait beaucoup de données personnelles. Sans compter les objets connectés (sport, santé, domotique, automobile, loisirs). On peut tout avoir comme données. Une fois celles-ci déterminées, il faut savoir où elles se trouvent. Les choix sont multiples. Elles peuvent être chez l'opérateur, dans les data centers, chez le fabricant et son ou ses hébergeurs, chez es sous-traitants, chez les industriels du big data qui offrent des services à des entreprises. Chacun de ces acteurs peut être partout dans le monde. Sans compter que les start-up proposent à des développeurs d'utiliser leurs données pour créer d'autres applications pas forcément légales.

En France, la protection juridique est assurée par la Cnil ou bientôt par le règlement européen encore en projet. Mais sa complexité fait que la mise en conformité est difficile. Il y a aussi les organisations à but « non bienveillant » et les applications ouvertes à des développeurs d'applications tierces.

Un règlement s'applique aux entreprises hors UE, pour ce qui se passe dans l'UE, mais pas pour l'extérieur. Aux États-Unis, les citoyens non-américains sont sous écoute globale et continue (Patriot Act). Les données sont massivement exploitées. En Russie, le cloud s'arrête aux frontières. En Chine, les pratiques, encore trop récentes, sont fortement orientées vers l'intelligence économique. À part l'Europe, le Canada et l'Argentine, aucun pays n'a une politique de haute protection des données.
Il y a un bémol à mettre sur la protection des données personnelles en France : elles sont captées par les opérateurs depuis la loi sur le renseignement. C'est une boite noire qui intercepte tous les trafics et envoie ces informations aux services de renseignements. Ce qui pose typiquement un problème

aux journalistes quant à la protection de leurs sources. Il y a une tendance générale au « big Brother data ».

Les risques pour les données personnelles viennent de la sécurité limitée des plates-formes informatiques, des grands réseaux d'opérateurs, des fuites techniques des constructeurs… Mais il y a aussi des incertitudes induites par les techniques de big data, des erreurs dues aux algorithmes et une fiabilité aléatoire des modèles prédictifs. Or les algorithmes sont de plus en plus présents dans les processus de décision. Les attaques sont donc possibles.

9. Enjeux économiques du big data et des objets connectés

Patrick Waelbroeck (Télécom ParisTech)

L'angle économique est abordé par Patrick Waelbroeck, professeur d'économie industrielle et économétrie à Télécom Paris-Tech. Il travaille depuis avril 2013, au sein de la chaire « valeurs politiques et informations personnelles ». Chaque discipline de la chaire (philosophie, économie, etc.) donne son avis. La donnée diffère des autres produits car elle n'a de la valeur que si elle se transforme en information. En rassemblant des données, on peut faire payer des prix différents à différents clients. Les philosophes appellent cela « l'identité idem ». Il ne faut pas négliger ce sujet. Identité, identification, c'est le sujet de la protection des données. Cette protection coûte cher ou plus exactement les bénéfices générés sont plus importants sans protection qu'avec. Selon certains économistes, il faut laisser faire le marché, car poser des contraintes sur les informations freine le business. Ce n'est pas bon pour l'économie. Cela suppose que le marché ne commet pas d'erreur. Or c'est faux. Il y a des « externalités » qui ne passe pas par le marché, mais qui influe sur les clients. Certaines externalités sont négatives. Uber par exemple : vols de données, filles importunées, diffusion de localisation de personnes qu'on veut importuner, envoi passif de spams. Mais le marché n'en tient pas compte. La protection des données privées a donc un sens. Autre exemple : les cookies qui impliquent l'ouverture d'espace publicitaire qui vous concerne. Même si vous n'acceptez pas de fournir vos données, les croisements de cookies permettent de vous identifier.

D'autre part, le contraire de l'identité est l'anonymat. En économie, l'anonymat est l'ennemi des entreprises. Mais beaucoup d'entreprises savent où vous habitez et où vous êtes. Pourtant, le postulat de base en économie néoclassique est que les acteurs sont anonymes, ne se connaissent pas et ne se parlent pas. Il en sort qu'il existe un bien être maximum. On ne peut pas augmenter le bien être de quelqu'un sans baisser celui des autres. La plupart des résultats montre que l'identification est un frein à l'économie. C'est la liberté de choix en étant anonyme contre le choix unique. À force de vous cibler, vous ne pouvez acheter que la même chose. Le ciblage réduit les choix des citoyens.

La sécurité n'est pas qu'un problème informatique. La valeur économique d'un service augmente avec le nombre d'utilisateurs. Les externalités positives font un effet de boule de neige positif. C'est

ce qui se passe avec Facebook, Google et autres. Dans ce type d'économie, il faut grossir le plus vite possible, puis ensuite de gagner de l'argent. Il faut diminuer donc les questions de sécurité et ne s'en préoccuper qu'ensuite. D'autre part, les cookies font que les données ne sont plus captives et partent dans différents serveurs. Dans l'internet des objets, chacun travaille en liaison avec les autres. Ainsi la sécurité d'une étape bénéficie aux autres. Pourquoi une entreprise investirait pour le bien des autres ?

Enfin la concurrence est fondamentale en big data. Microsoft a sorti Microsoft internet gratuitement, alors que Netscape était leader mais payant. Apple a installé par défaut son logiciel de cartographie. Facebook tente de mettre par défaut une adresse mail dans son serveur pour capter davantage de clients.

10. Big Data : éclairage philosophique sur l'usage des données personnelles

Laura Lange (Consultante Philosophe)

Laura lange est consultante philosophe en entreprise et doctorante à Paris-Est. Elle commence par la métaphore du poisson volant. Il utilise ses ailes pour surplomber son environnement et revenir dans son milieu avec plus de connaissances. C'est ce que fait la philosophie, qui est l'aile humaine. Il faut faire attention aux mots, car chaque mot est une représentation et convoque une imaginaire différent. Laura Lange est adepte de la philosophie des poupées russes. Chaque domaine est inclus dans un plus grand. Le contexte est important. Qu'est-ce que le monde numérique ? Le dernier des mondes en date. Il est complexe car relié, dirait Edgar Morin. Les ordinateurs ont relié les textes, les gens, les données. Nous avons chacun une identité virtuelle, une empreinte numérique. Elle est impalpable, car même si on la génère, on ne la gère pas.

Le big data, ce sont les données dans les nuages. Dans ces nuages, il y a des techniques avec des promesses et des menaces. Ne jamais oublier que l'enthousiasme se termine pas des craintes. Internet a commencé dans les années 1960 avec la post-modernité, mouvement fort où la liberté était le mot d'ordre. Internet s'oppose ainsi au structuralisme. Le big data au contraire remet du structuralisme sous forme de frontières afin de nous proposer des possibles réduits à notre personne. Le premier gain des big data est la personnalisation. On va au delà des apparences. Cela révèle des profils. Ce qui permet des interactions avec l'environnement, avec autrui, avec vous-même. Et permet aussi une amélioration de la qualité de vie, dans le sens où il faut lutter contre ce qui va freiner votre liberté. Les objets connectés nous libèrent de taches qu'on peut ne pas faire. Ce qui libère du temps pour des taches choisies. Sauf que l'individualisme des Lumières n'est pas l'individualisme actuel. On cherche à lutter, à gagner sur les autres. C'est un individualisme concurrentiel. Dans cette culture-là, le big data sert à conquérir et non à aider.

Des masses de menaces. La culture actuelle est dite « hypermoderne ». C'est le toujours plus, la promotion de l'individualisme en masse, le retour identitaire, le communautarisme. On se croit hyper-libre, mais on est hyper-contraint. Il existe une menace d'une nouvelle catégorisation, d'un

nouveau système de méritocratie, de se voir enfermer dans des propositions non choisies. Quel est la place du désir personnel ?

Aujourd'hui nous sommes dans des relations en rhizome. Tout est emmêlé, mais on se demande où se trouve la racine. L'être humain perd racine. La philosophie, qui cherche les racines, peut aider à répondre à la question.

11. Les formes modernes du contrôle social : de la sorcellerie au village à la numérisation dans les sociétés urbaines

Dominique Desjeux (Université Paris Descartes)

Dernière intervention, celle de l'anthropologie. Dominique Désjeux, professeur émérite à l'université Paris-Descartes, a écouté les autres interventions. Il considère qu'il y a un paradoxe. La crainte du marketing est d'avoir des clients infidèles. Celle des clients est d'être capté. Or, sans être libre, nous avons quelques fois et par moment des marges de manœuvres. Il faut avoir des angoisses, mais pas trop. Sans angoisse ou quand on en a trop, on ne fait rien.

Les marges de manœuvre ne sont pas individuelles. L'approche économique est de bien voir le rapport coût-bénéfice. Elle nous rappelle que la sécurité coûte cher, plus cher que le curatif. L'éducation à la sécurité ne suffira pas si on ne change pas le coût.

Les big data c'est à la fois ancien et nouveau. L'anthropologue, mi-ethnologue, mi-sociologue, sait que les comportements humains sont constants dans leur diversité. Ce qui est nouveau est la technologie. La constante est la sorcellerie africaine. Dans les années 1970, en travaillant dans un village au Congo, Dominique Dejeux s'est rendu compte que la sorcellerie permet d'expliquer le bien et le mal comme la paranoïa qui dit qu'il y a toujours quelqu'un qui vous veut du mal et qui vous jette un sort. Lorsqu'il y a un malheur, on invente une intention et la sorcellerie met les deux ensemble alors qu'ils n'ont rien à voir. En big data, c'est pareil. Il n'y a pas de lien entre intention et effet. Le contrôle social le meilleur est le contrôle magico-religieux. Aujourd'hui on a une augmentation des sociétés libérales et en même temps des sociétés très traditionnelles.

On a eu les 30 Glorieuses. En 1973, on s'est rendu compte qu'il y avait des défauts. Il y a toujours de promesses et des menaces. La thèse de Dominique Dejeux est que la numérisation de la société est un retour à la normale, avec un contrôle social. L'urbanisation donne plus de liberté par rapport au contrôle des villages. Les big data est le contrôle social de la numérisation. Au niveau micro-social, les marges de manœuvre existent pour le contrôle des données. Mais les groupes de pression ont du mal à changer le monde. Ce n'est pas facile, mais possible. Le jeu des acteurs rend l'avenir optimiste.

Conclusion

En fin de journée, **Pascal Alix** remercie les intervenants et les auditeurs et donne rendez-vous à un prochain séminaire. À 18 h 20, la salle se vide.