QUI CONTROLE LE BIG DATA ?

Saint-Herblain, Maison des Arts

4 mai 2017

Conférences « Place Publique, Démocratie Participative »

Introduction de la soirée-débat par le maire de Saint-Herblain qui se réjouit de la forte participation 'intergénérationnelle' à cette manifestation.

L'animateur, qui relancera les quatre intervenants par ses questions tout au long de la soirée, présente chacun d'entre eux :

Nicolas GUY : Il a fondé la société 'SoyHuCe', une startup normande qui a développé une application utilisant des données pour repenser et améliorer le lien entre les citoyens et les collectivités.

Fabrice BENAUT : Il dirige 'iDeaTrans' mais est également membre de plusieurs associations du Digital Data Marketing.

Valentina FERREOL : Elle est présidente de l'Institut G9, spécialiste de la transformation des entreprises et elle étudie les impacts technologiques et humains du numérique.

Dominique CARDON : Il est sociologue à l'Université Paris Est Marne-la-Vallée et il est l'auteur du livre « *A quoi rêvent les algorithmes ; nos vies à l'heure du Big Data* ».

L'animateur va rappeler ce qu'on entend par 'Big Data', soit 'Données de masse ' et va d'entrée de jeu répondre à la question de la conférence en mettant la responsabilité sur les entreprises multinationales. Il cite le « biopouvoir », un type de pouvoir qui s'exerce, selon Michel Foucault, sur la vie des corps et celle de la population. Ces données sont caractérisées par les « 3V » : Variété, Vélocité et Volume. Sur la notion du 'volume' il donne comme exemple le 'Large Hadron Collider' (LHC), un accélérateur de particules mis en fonction en 2008 et situé entre la périphérie de Genève (Suisse) et le pays de Gex (France). Cet appareil produit l'équivalent de 3 millions de DVD de données par an. On pense aujourd'hui que Facebook produit encore plus de données.

D'une façon générale, dans cette nouvelle représentation du monde (nouveau paradigme) il faut produire toujours plus de données afin d'établir les algorithmes (séries d'opérations) qui vont les collecter, puis les algorithmes qui vont les traiter. Ces algorithmes sont comparés à des recettes de cuisine. Quelle est la recette du moteur de recherche Google ? Il rappelle néanmoins que s'il y a beaucoup de promesses le traitement de ces données a ses limites et que, par exemple, les données de masse ne peuvent pas 'tout' prédire. S'il y a beaucoup d'opportunités, il y a, par ailleurs, beaucoup de dangers. Quel est le monde des Big Data ?

Quel est le marché de la donnée ? Quels sont les usages d'Internet ? Vers quelle vision du monde allons-nous ? Autant de questions qui vont alimenter le débat entre les quatre intervenants.

Nicolas GUY

Historiquement on peut situer la naissance du Big Data en 1941 avec Alan Turing et l'ère des algorithmes. C'est en effet la 'Machine de Turing' qui a permis de déchiffrer les messages codés allemands pendant la seconde guerre mondiale. A partir de là on s'est aperçu que les algorithmes, les mathématiques, allaient permettre de résoudre énormément de problèmes pratiques au quotidien. L'autre étape sera 1976 avec le livre d'Emmanuel Todd *« La chute finale* » qui a permis de prédire la chute de l'empire soviétique en manipulant des concepts.

Un exemple d'application moderne est Google Map, autrement dit l'utilisation d'anciens algorithmes pour trouver le trajet le plus efficace afin d'aider les représentants de commerce. Un autre exemple est sans doute la voiture autonome qui est bardée de capteurs pour se comporter 'intelligemment'. Des millions de données sont générées et traitées chaque minute. Le Big Data est bien derrière l'ensemble des démarches de stockage et de traitement de ces données.

L'algorithme est donc une recette appliquée à des problèmes concrets. Il va permettre de sélectionner la solution que l'on estime adéquate pour résoudre un problème puis de la coder dans un langage interprétable par la machine.

Dans ma startup je veux arriver à créer la 'ville intelligente' (smart city). Jusqu'ici les opérateurs privés ont pris la main et ont 'siloté' les données concernant la vie en ville. Je propose d'utiliser la démocratie participative pour que les citoyens retrouvent une gouvernance : je me base sur les citoyens et les données qu'ils fournissent pour créer des algorithmes permettant d'améliorer les services publics. Il y a au départ un consentement du citoyen qui va fournir des données environnementales et non personnelles.

Qui contrôle le Big Data ? Avant tout celles et ceux qui maîtrisent les algorithmes. Aujourd'hui, en utilisant les données des caméras de surveillance, on peut, avec de nouveaux algorithmes, détecter qui est entré avec une valise puis est ressorti sans cette valise, et cela <u>sans regarder les images</u>.

Fabrice BENAUT

A partir du moment où on collecte de la donnée on va créer une chaîne de création de valeurs qui va aboutir à la monétisation de la donnée. L'économie de la 'Data' est une économie collaborative : on corrèle les données entre elles et ce faisant on crée donc de la valeur. Le modèle le plus simple est la vente des fichiers mais si on enrichie la donnée on va créer d'autres valeurs pour d'autres usages. On démultiplie ainsi la valeur initiale de la donnée.

Qu'est-ce qu'une donnée 'granulaire', la 'granularité' des données ? Plus les données sont sélectionnées, plus elles sont 'personnelles' et plus leur granularité sera 'fine'. Que va-t-on faire de ces données ? On touche là le domaine de la CNIL.

Pour créer de la valeur on peut partir de la donnée mais on peut aussi partir du 'besoin'. Mais pour libérer l'espace de création de valeurs il faut obtenir toutes les garanties nécessaires et donc 'anonymiser' les données par différents moyens (brouillage – codage – agrégation). Le marché français aujourd'hui représente 55 milliards d'euros avec une croissance annuelle de 35 %. Environ 50 % de ce budget est constitué par 3 secteurs d'activité : la Finance (utilisation de robots), la Santé et le Secteur Public. Il est étonnant de voir que 90 % de la donnée disponible ont été créés ces 2 dernières années. Il faut savoir que toutes les applications sur nos smartphones créent des données que l'on appelle 'traces'. Paradoxalement, même lorsqu'on ne fait rien, on produit de la donnée.

Qu'en est-il de la puissance du « GAFAM » (= Google + Apple + Facebook + Amazon) ? Les transformations digitales permettent d'être partout en temps réel. Toutes ces données sont homogènes et connectées au réseau IP. Ce qui va être important c'est le comportement de l'utilisateur aux 'points de contact'. Par exemple le smartphone qui va permettre de suivre les parcours du client. Aujourd'hui la tendance est à réduire de plus en plus ces points de contacts.

Pour réserver une chambre d'hôtel on va passer par Google qui est propriétaire de Booking.com mais on n'aura pas accès au téléphone de l'hôtelier : celui-ci est pris en otage par Booking.com.

Dès qu'une startup a une 'bonne idée' qui peut intéresser Google ou un autre groupe international ceux-ci rachètent cette société. Le marché est ainsi 'très concentré' sur le GAFAM. Lorsqu'Amazon a commencé il n'a pas gagné d'argent mais a surtout essayé de racheter des startups et d'héberger tous les concurrents potentiels. Tous les contenus des messages sur gmail sont stockés sur des plateformes conçues à cet effet pour être transmises au gouvernement américain (c'est prévu par la loi aux USA).

Valentina FERREOL

Les données des grands groupes sont 'stockées quelque-part' mais on ne sait pas trop comment. Cependant, les données publiques qui nous concernent sont sécurisées. Par exemple nos déclarations d'impôts, ou nos données 'santé' qui sont détenus par nos médecins. Cependant, si les données 'santé' sont traitées uniquement par le corps médical il y a une différence entre, par exemple, les données transmises au médecin par le 'boitier d'alerte' qui suit à distance une personne âgée (données conservées uniquement par le médecin) et les données de mesure à distance d'un paramètre (ex : tension) qui iront sur le 'cloud' et seront non protégées. Heureusement 'l'effet volume' tasse tout.

A côté des données médicales un autre secteur intéressant de la 'transformation digitale' concerne les transports. Depuis 2015 la SNCF a ouvert des interfaces d'application. Ceux-ci traitent des données 'neutres' (les horaires des trains) mais aussi des données personnelles 'voyageurs' qui ne seront jamais communiquées à l'extérieur mais vont être traitées par la SNCF. Mais il y a plus de 500 formats différents de données uniquement sur le sujet des transports ! D'où la difficulté pour les grandes et moyennes agglomérations. Une expérimentation a été lancée par la ville de Belfort pour traiter les données concernant le

réseau de bus et les données sur la consommation énergétique. Ces données sont 'en silo'. Elles sont exploitées séparément, chacune dans leur domaine.

La loi 2016-1321 du 7 octobre 2016 pour une République numérique, dite 'Loi Lemaire' introduit notamment l'ouverture par défaut des données publiques, la neutralité du net, une obligation de loyauté des plateformes en ligne, ainsi qu'une protection accrue pour les données personnelles des usagers du net. La loi pour une République numérique prévoit également les conditions d'un Internet accessible au plus grand nombre, au travers de l'accélération de la couverture du territoire en très haut débit et en téléphonie mobile, de mesures pour un meilleur accès des personnes handicapées aux services en ligne, et de la création d'un droit au maintien de la connexion internet en cas d'impayé pour les foyers en difficulté.

A Saint Malo une démarche de concertation des citoyens - sous forme d'un vote – a été réalisée en vue d'ouvrir certaines données et de les 'exposer'.

Quid du 'Pass Navigo' (*question de l'animateur*) qui se créé sans consentement des voyageurs ? Les informations sur les itinéraires des voyageurs ne sont pas exposées. C'est une des réglementations de la CNIL.

Quid des 'compteurs Linky' ? (*question de Fabrice Benaut*) Ils peuvent tout tracer par foyer et les données personnelles sont gardées par l'exploitant… Par ailleurs, des municipalités ont accepté de placer de données personnelles de citoyens sur des plateformes situées à l'étranger… alors que c'est interdit.

C'est vrai que la tendance est à faire héberger la donnée là où c'est simple et pas cher.

Autre exemple, à Nantes, où il y a une dynamique forte sur les transports en commun alors qu'une étude récente montre que 40 % des déplacements se font encore en voiture pour des distances comprises entre 1 et 10 Km. D'où des bouchons et de la pollution. Pour arriver à convaincre de privilégier les déplacements en vélo, en particulier via les vélos partagés (Bicloo), il va falloir déterminer les circuits les plus adaptés en fonction de l'heure pour éviter d'avoir au même moment des stations Bicloo vides et d'autres inutilisées : pour cela il faut trouver les bons algorithmes pour gérer de façon optimale l'utilisation de ces vélos.

Dominique CARDON

Ma vision sociologique du Big Data tend de plus en plus à être un peu 'blasée' par rapport à des modes qui vont et qui viennent. Le Big Data suppose des consultants et des marchés. Mais ces données accompagnent des 'promesses' de gens qui n'ont jamais mis la main dans les données. Il ne suffit pas de croiser un nombre très important de données (notion de Variété) pour en sortir automatiquement quelque-chose. En fait on fabrique de la promesse et surtout…des craintes. C'est du techno-déterminisme mais les deux types d'attitude sont dans l'erreur. Le Big Data ne marche pas si bien que ça : i) la donnée « n'est pas donnée »…elle a un prix, ii) elle est 'sale' et il faut d'abord la nettoyer puis lui donner du sens, iii) il ne faut pas

aborder de la même façon les différentes familles de données, iv) des données 'silotées' depuis longtemps commencent tout juste à être utilisées.

On revient toujours à « quels seront les bons calculs pour donner du sens à ces données ? ».On ne voit pas arriver de choses très performantes, par exemple, dans le domaine de 'l'hyperdata'. La vraie révolution, en fait, ce sont les algorithmes.

Une famille de données particulièrement sollicitée ce sont les données 'bayésiennes' qui sont prédictives par comparaison des probabilités d'occurrence d'évènements (Vélocité). L'ingénieur construit une infrastructure 'qui apprend' et qui est donc peu différente d'une 'intelligence artificielle' (IA). C'est là qu'il y a une vraie révolution scientifique. Cependant, les courbes obtenues à partir des données collectées par Linky ne donnent qu'une bonne information sur… la Météo ! Le 'Deep Learning' est plus original car il permet une vision du monde à partir des données de faible granularité. Est-ce une connaissance de la population par catégories ? Plutôt par des 'traces' (navigation, tickets de caisse, heures de sommeil) qui vont permettre de prédire les services plus ou moins utiles à mettre en place. En matière de choix musicaux des applications comme Deezer ou Spotify ne feront une bonne prédiction que si on leur a fourni beaucoup de données et de façon régulière.

Oui (*intervention de Fabrice Benaut*) il faut beaucoup de données pour construire l'algorithme. En dépit de ce que l'on croit le 'Digital Data' enferme dans des fonctionnements : on ne propose que ce qu'on a l'habitude de consommer. Il y a l'exemple connu de Microsoft qui, lorsqu'il a lancé son algorithme avec IA, s'est aperçu que l'application avait des choix 'raciste' parce qu'elle avait été insuffisamment paramétrée.

QUESTIONS DU PUBLIC

Question 1 :

En fait une salve de trois questions d'un ex-hacker très motivé : i) Existe-t-il des garde-fous dans le contrôle du Big Data vis-à-vis de la NSA et du secteur privé ? ii) Quid de l'extrémisme technologique ? et iii) Doit-on considérer que l'empreinte écologique du Big Data est nulle ?

Réponse de Dominique CARDON : il y a deux façons de considérer la première question i) même si on n'est pas 'Geek' on ne peut s'extraire de cet environnement et il n'est pas très réaliste 'd'en sortir' et ii) en réalité on domestique vite ces technologies et on peut toujours mettre des 'régulateurs' plus tard (voir loi 'Lemaire'). Il y a ce qu'on appelle l'obfuscation : on est dans ce monde et on ne peut en sortir mais on peut en adapter les usages. (Définition de Wikipedia que le rédacteur est allé chercher : obfuscation = obscurcissement ou stratégie de gestion de l'information qui vise à obscurcir le sens qui peut être tiré d'un message. Cette stratégie peut servir en matière de protection de la vie privée - par exemple, pour la protection des données personnelles ou la gestion de la réputation numérique).

<u>Réponse de Valentina FERREOL</u> : à l'origine, les 'Data Centers' qui hébergeaient les données tournaient doucement et consommaient beaucoup d'énergie. Aujourd'hui, avec les nouvelles technologies, le matériel a progressé. Il consomme beaucoup moins d'énergie et stocke davantage de données. Par ailleurs les produits très polluants comme le plomb et d'autres métaux lourds sont recyclés. C'est plutôt positif vis-à-vis du développement durable. Quant à la 'réalité augmentée' c'est plutôt un mythe.

<u>Réponse de Nicolas GUY</u> : chez Facebook les images sont stockées sur DVD et c'est un robot qui va les chercher parce que si on les stockait sur des machines actives ce serait très énergivore.

Question 2 :

Depuis la prise de conscience due à l'affaire Edward Snowden[1] on a découvert des choses qu'on ne soupçonnait pas et on sait maintenant ce qu'on peut craindre. Récemment on a appris par les informations qu'en Turquie le président Erdogan a choisi de couper l'accès à Wikipedia. Quel est le pouvoir réel du secteur privé et quel est le pouvoir réel du politique (exemple de la Turquie) ? Il est intolérable que dans un pays on puisse bruler l'équivalent de la moitié des bibliothèques sans que personne ne s'en émeuve. Qui maîtrise le robinet ?

<u>Réponse de Fabrice BENAUT</u>: il arrive en effet qu'un état décide de 'couper' l'accès aux données d'Internet (c'est arrivé également en Chine). C'est tout le problème de la concentration sur des acteurs tout puissants. La réglementation Européenne n'est pas excessive : tout ou presque a été 'volé' en France et en Belgique, par exemple. A cause des données gmail Lafarge a pu être racheté par le premier cimentier américain…

Question 3 :

Je suis fondateur d'une société qui travaille sur la place de la donnée. Pour moi la question essentielle est : Qui contrôle <u>ma</u> Data ? C'est la vraie source de nos peurs. En France nous avons la CNIL et la loi de 1978. La réglementation européenne à venir sera très protectrice et les français auront été vraiment 'pilotes' dans ce projet. Mais comment chacun peut-il 'contrôler' ses données ? Le plus grand mensonge d'Internet c'est le fameux clic « j'ai lu les conditions générales et je les accepte ».

<u>Réponse de Dominique CARDON</u> : oui, c'est vrai, tout repose sur un consentement de l'utilisateur qui a été quelque-part 'extorqué'. Au départ on accepte – ou pas – que le site dépose un cookie. Certains cookies vont se déposer sur notre navigateur et c'est comme cela que nos déplacements vont être captés et revendus à des sites commerciaux. C'est le problème du cookie 'tiers' auquel on n'a pas consenti. Mais cela fait tenir tout le marché de la publicité

[1] À partir du 6 juin 2013, Snowden rend publiques par l'intermédiaire des médias, notamment The Guardian et The Washington Post, des informations classées top-secrètes de la NSA concernant la captation des métadonnées des appels téléphoniques aux États-Unis.

numérique qui, autrement, aurait depuis longtemps périclité. A noter que l'on peut utiliser les services de la startup Privony® qui protège nos données.

Réponse de Nicolas GUY : à partir du moment où on est sur Internet on sait que l'on va être 'tracé' et il faut l'accepter.

Question 4 :

On est passé de grosses machines couteuses à plein de petites machines pas chères et cela a été le départ du Big Data. N'est-ce pas l'Asie - en particulier la Chine – qui est à l'origine du Big data ? Apple peut vendre 10 millions de téléphones en une semaine : ce n'est pas possible sans la capacité de production chinoise.

Réponse de Nicolas GUY : s'il n'y avait pas eu la Chine ce serait de toute façon la robotisation qui l'aurait fait (la production en masse). Il y a à côté du web connu le 'dark web' où les informations ne sont pas référencées par des opérateurs privés.

Réponse de Dominique CARDON : Oui, la Chine a un impact sur le 'hardware' mais en ce qui concerne le Big Data tous les algorithmes sont publics, sous forme de librairies informatiques.

Question 5 :

La philosophe Cynthia Fleury craint que l'informatique ne se substitue totalement aux hommes dans la conduite de leurs affaires publiques autant que privées. Elle pose le problème « l'Homme qui se machinise ». Toutes les données sur nos achats en grande surface, par exemple, sont stockées et des machines décident de notre consommation future. Il n'y a plus de choix humain, malgré la percée des produits 'Bio'. Chaque fois qu'on paie en monnaie digitale on est esclave de ceux qui pilotent ces informations. Par ailleurs, sur Facebook, il est connu que les 'amis de os amis' piratent nos données…

Réponse de Fabrice BENAUT : on devient de plus en plus des guichets avec plus personne à qui parler. Il n'y a plus de contacts humains. Par ailleurs, le e-commerce, contrairement à ce qui avait été dit, n'a pas tout remplacé. A part pour quelques acteurs très concentrés il a surtout entraîné des pertes financières. La transformation digitale ne doit pas 'mettre l'Homme sur le bord du trottoir'.

Réponse de Dominique CARDON : un algorithme c'est un ensemble de paramètres réglés pour être 'efficaces'. Il suffirait de changer ces paramètres pour obtenir quelque-chose de plus attrayant et de plus humain, changer leurs fonctions d'objectif pour ne plus être axé sur une utilisation optimale.

Question 6 :

Comment récupérer les photos mises sur Facebook, quel est notre droit d'image ? J'ai fait le choix, personnellement, de tout payer en espèces, de téléphoner avec un vieux Nokia sans

application, de repérer sur Internet ce que je veux acheter puis d'aller acheter sur place dans le magasin…

Réponse de Dominique CARDON : La photo mise sur Facebook peut être récupérée mais il ne faut pas oublier que Facebook peut l'utiliser comme il veut car… on lui en a donné le droit en nous inscrivant.

Question 7 :

Qu'est-ce qui manque à la France pour se positionner sur le marché de la Big Data ?

Réponse de Fabrice BENAUT : l'écosystème des startups françaises est très dynamique. Mais on se fait souvent racheter par les grands groupes. Pour garder sa souveraineté il faut arriver à conserver ses 'routeurs' (les aiguillages par lesquels passent obligatoirement les données). On avait cette indépendance avec Alcatel (France), jusqu'à ce qu'il se fasse 'essorer' par Lucent (USA). Nous n'avons donc plus de routeurs français.

Réponse de Valentina FERREOL : on a une bonne capacité d'innovation en France, surtout au niveau des PME. Peu de startups ont réussi à se maintenir –comme on vient de le dire – parce qu'elles ont été rachetées. Il faut arrêter de s'autocensurer, reconquérir le terrain perdu : il n'est pas trop tard.

Patrick Lassus le 8 05 2017