Ces masses de données que nous semons sur Internet

Préambule **2**
Remarque 2
Structure du document 2

Cinq points pour mieux comprendre les données numériques **3**
1. De quoi parle-t-on ? Quelles sont ces fameuses données ? 3
De l'origine des données 3
2. Comprendre les mégadonnées 5
3. Qui sont les acteurs du monde des données numériques ? 7
4. Quelle est la finalité des mégadonnées ? 8
Une finalité avant tout économique 8
Données et publicité 8
Mégadonnées et ciblage publicitaire 8
Le reciblage 9
Un ciblage des produits et des contenus 9
Un ciblage très controversé 9
Du ciblage à la manipulation 10
5. Pourquoi se préoccuper des mégadonnées ? 11
Une explosion des mégadonnées qui va s'accentuer 11
L'Internet des objets est un Internet des données 11
Une invasion de la sphère professionnelle 12

Trois points pour comprendre les principaux enjeux des données **12**
1. Les mythes du consentement et de la liberté des utilisateurs 12
Une opposition fondamentale 12
L'impossible consentement 13
L'absence de réel choix 13
Un accroissement des risques 14
La dimension collective du risque 14
Des systèmes pas si neutres 15
2. La concentration des données 15
Les risques de l'hyperconcentration 15
Un profilage affiné 15
Une sécurisation impossible ? 16
Entre données marketing et surveillance généralisée des populations 16
3. La transparence asymétrique 16
Asymétrie, panoptique et liberté 16
Un cadre défini unilatéralement par des privés 17

Quelles pistes de « solutions » face aux mégadonnées ? **18**
L'absence de solutions simples 18
Prise de conscience 18
Résistance 18
Un cadre contraignant 18
Des mesures à prendre également par domaine 19
Les organisations peuvent-elles faire preuve de responsabilité ? 19
Depuis la vision jusqu'à la mise en œuvre 19
Des compromis protégeant l'essentiel 19

Quelques références **20**
Sites web 20
Livres 20
Notes 20

Préambule

Nos données constituent un enjeu majeur de la transformation digitale dans laquelle le monde s'est engagé. Jour après jour, nous semons ces données, souvent à notre insu, et des systèmes de plus en plus gigantesques les collectent, les enregistrent, les traitent, les recoupent, les échangent et les valorisent...

Le domaine du marketing, et en particulier celui de la publicité, sont friands de ces masses de données exploitées dans l'espoir de mieux comprendre, de mieux satisfaire, et de mieux influencer le comportement des consommateurs.

Ce domaine est loin d'être le seul à s'intéresser aux mégadonnées, terme désormais consacré pour remplacer en français l'expression anglaise *big data*[1]. La plupart des domaines et des organisations s'affairent à développer des projets de collecte et d'exploitation de données, les uns pour améliorer le recrutement ou l'efficacité des salariés, les autres pour optimiser la logistique, améliorer leurs produits, prévoir la météo, détecter précocement des épidémies, prédire l'évolution des marchés financiers, comprendre le cancer, imaginer des médicaments, espionner des ennemis, etc.

C'est peut-être ce développement tout azimut qui rend difficile la compréhension des questions, des enjeux et des risques liés aux données numériques. Il se peut aussi que ce soit leur caractère abstrait, immatériel, invisible, flou même, ou la technicité des systèmes informatiques associés à la gestion de ces données. La relative récence du phénomène est peut-être aussi un facteur explicatif de cette incompréhension, à moins qu'il ne s'agisse d'une volonté délibérée de la part de tous ceux qui œuvrent au développement de ces systèmes et ont tout intérêt à pouvoir continuer à les construire sans question ni entrave aucune.

Remarque

Ce document aborde la question des mégadonnées sous l'angle du marketing et des usages des mégadonnées par les entreprises. Il ne fait qu'évoquer ponctuellement les enjeux associés aux pratiques de surveillance et d'espionnage, non que ces questions ne soient pas cruciales, bien au contraire. Il est d'ailleurs important de comprendre que l'ensemble des mégadonnées collectées par des entreprises privées à des fins marketing, sont potentiellement intégrées dans les programmes de surveillance. Les révélations d'Edward Snowden sur l'ampleur de la collecte et du stockage de données par la NSA sont parfaitement explicites à ce sujet. En Europe, les récentes lois sur le renseignement légitiment d'ailleurs en grande partie ce genre d'interception de données et de surveillance à très large échelle.

Cette collusion entre les entreprises privées et la surveillance d'État au niveau des mégadonnées brouille la compréhension du sujet et rend ce sujet particulièrement sensible, au regard de la protection des libertés individuelles et collectives[2].

Structure du document

Ce document a donc pour objectif de lever un coin du voile sur les questions relatives aux mégadonnées telles qu'elles sont utilisées dans le domaine du marketing. Il commence par définir les données numériques, pour ensuite préciser ce que sont les mégadonnées et leurs caractéristiques, puis par rappeler qui sont les acteurs de ce domaine, pour ensuite discuter certains des enjeux qui leur sont associés et proposer quelques pistes de réflexion par rapport aux problèmes qu'elles posent.

Cinq points pour mieux comprendre les données numériques

1. De quoi parle-t-on ? Quelles sont ces fameuses données ?

La source de notre difficulté à appréhender les questions relatives aux données apparaît dès lors qu'il s'agit de définir précisément ce que sont ou ce que ne sont pas ces fameuses données. Or, la définition même d'une donnée n'est pas si simple.

Dans sa leçon inaugurale au Collège de France, Serge Abiteboul[3] définit la donnée comme étant « une description élémentaire, typiquement numérique pour nous, d'une réalité ». Il rappelle, quelques lignes plus loin, la proximité des notions de donnée, d'information et de connaissance.

Différentes sources établissent d'ailleurs une forme de hiérarchie à quatre niveaux, partant des données brutes, pour aller vers la sagesse, en passant par l'information et la connaissance (en anglais : *data* → *information* → *knowledge* → *wisdom*). L'information s'appuie sur les données auxquelles elle ajoute un contexte, une structure, des relations. Les connaissances reposent à leur tour sur les informations auxquelles elles confèrent un sens, une capacité à soutenir une action, une prise de décision, par rapport à une finalité spécifique. La sagesse, pourrait alors émerger, à travers une mise en relation de multiples connaissances.

Suivant les sources, les frontières entre ces différents concepts varient sensiblement. D'un point de vue marketing, ces frontières ne sont pas essentielles et d'un point de vue informatique, l'ensemble de ces notions sont en réalité confondues sous forme de données stockées et traitées par les systèmes. D'ailleurs, les métadonnées, à savoir les données sur les données, entrent également dans ce magma informationnel, et en constituent même un part significative. Pour résumer, et afin de clarifier nos propos, nous considèrerons dans la suite de cet article que tout est *donnée*.

De l'origine des données

Pour comprendre les questions relatives aux données, il est utile d'abord de se pencher sur leur origine, leur source, leur génération. Dans ce domaine, et en particulier dans le contexte du numérique, nous pouvons distinguer plusieurs sources de données :

- la **déclaration** : lorsque l'utilisateur remplit un formulaire, et communique un certain nombre de données en répondant à des questions. A priori, ces données sont transmises de façon consciente et de plein gré par l'utilisateur. Cela étant, nous avons tous rencontré des formulaires dont certains champs sont *obligatoires*, quand bien même les données demandées semblent superflues ou sans rapport direct avec le contexte. Dans ce cas-là, il est certainement permis de s'interroger sur la liberté de l'utilisateur face à la fourniture de ses données. De la même façon, l'utilisateur peut avoir conscience qu'il communique des données, sans forcément comprendre l'utilisation qui peut en être faite ou les déductions qui peuvent en découler. Nous reviendrons ultérieurement sur ces questions.

- l'**observation** : lorsqu'un utilisateur consulte un site web, il est possible d'observer son comportement et d'enregistrer des données à propos de celui-ci. Cette observation peut générer un très grand nombre de données, d'une incroyable précision. Il est techniquement très simple d'observer par exemple les pages consultées, les produits étudiés, les recherches effectuées, les achats en ligne, les articles lus, ou tout autre comportement dépendant du contexte d'usage du site web. Ces observations peuvent intégrer des métadonnées liées à cette consultation, par exemple le lieu, l'heure, l'appareil utilisé, ses caractéristiques, le temps passé sur les différentes opérations, etc. Ces exemples sont issus du web, mais la même typologie peut parfaitement s'appliquer dans d'autres domaines, comme celui des applications mobiles,

des réseaux sociaux, des achats par carte de crédit, etc. Le fait d'apprécier – d'*aimer* dans la terminologie consacrée par Facebook - une personne, une page ou un billet constitue également un comportement que le système enregistre comme une nouvelle donnée à propos de l'utilisateur. Le fait de visionner une vidéo sur YouTube ou d'effectuer une recherche sur Google relève de la même catégorie d'intrant par rapport au profil d'un utilisateur. Par analogie, les banques ou les fournisseurs de carte de crédit peuvent observer les transactions effectuées par leurs clients, les opérateurs télécom sont capables d'étudier les appels des abonnés, les publicitaires enregistrent la réponse des cibles face aux annonces, les constructeurs de voitures connectées commencent eux-aussi à analyser les parcours des conducteurs, etc. S'il semble relativement facile de comprendre qu'un fournisseur de services numérique peut observer le comportement de ses utilisateurs, il est nettement moins évident d'avoir conscience de la masse de données que ces observations comportementales vont générer, sans parler de la façon dont elles vont être stockées ou traitées…

– la **combinaison**, le rapprochement : le terme n'est peut-être pas idéalement choisi, mais l'idée est de décrire le regroupement d'informations provenant de différentes sources afin de compléter un jeu de données, sans forcément créer de nouvelles données. A titre d'exemple, nous pourrions penser à la décision prise par Google[4] en 2012 d'unifier la gestion des données de ses produits et services (environ 70 à l'époque). A travers cette décision, Google a regroupé les comptes de ses utilisateurs et s'est autorisé à traiter de façon combinée l'ensemble des données associées à chacun de ces comptes. Le but affiché était alors de fournir de meilleurs services et de faciliter la gestion des données aux utilisateurs. Il a en réalité surtout permis une combinaison de données sans précédent et la constitution de profils utilisateurs extrêmement détaillés car traversant les frontières de l'ensemble de produits de Google. La firme a ainsi pu regrouper autour d'un profil unifié une quantité de données comme les recherches effectuées, les messages électroniques échangés, les photos réalisées et admirées, les amis connectés, les articles de blog publiés ou consultés, les vidéos regardées, etc. Google n'est pas le seul à procéder à ce type d'opération. Beaucoup d'entreprises combinent notamment des données provenant du monde en ligne avec des informations issues du monde réel. Les administrations publiques font aussi régulièrement état de projets de cette nature et dans certains pays l'interconnexion ou le rapprochement de fichiers est réglementé.

– la **déduction**, l'inférence : les catégories de données qui précèdent peuvent être traitées pour en déduire de nouvelles. Ces données déduites s'ajoutent alors aux données de base. Les déductions peuvent être construites à partir des données d'un individu, ou être développées en croisant les données d'un individu avec celles d'autres individus (par exemple des données statistiques). Nous verrons plus loin que ces déductions posent des questions spécifiques, notamment par rapport à leur nature, par essence probabiliste. Par ailleurs, ce type de données, généré par des algorithmes, des règles d'inférences ou des modèles statistiques, est généralement entouré d'une part de mystère. Dans la plupart des cas, ni les algorithmes, ni leurs paramètres, ne sont connus des utilisateurs.

Ces différentes catégories décrivent donc l'origine des données, leur source. Elles qualifient d'une certaine façon la nature des intrants d'un système qui stocke et traite ensuite ces éléments comme on traite une matière première. Leur masse a donné naissance à l'expression **big data**, traduite en français par *mégadonnées*.

2. Comprendre les mégadonnées

Dans le secteur de l'informatique, cela fait plus de 15 ans que le big data est évoqué, et déjà plusieurs années qu'il fait l'objet d'une intense promotion de la part des acteurs du domaine[5]. Pour tenter de décrire ce courant, aux frontières floues, il est souvent fait référence au modèle dit « des 5 V »[6]. Ces cinq V, **Volume**, **Vitesse**, **Variété**, **Véracité** et **Valeur** sont autant de clés pour mieux comprendre les phénomènes associés aux mégadonnées et les questions qu'elles posent.

Pour les entreprises, et pour le marketing en particulier, ces dimensions permettent de mieux appréhender la nature des différences entre les bases de données relationnelles classiques (apparues dans les années 1970) et les mégadonnées d'aujourd'hui. Or, c'est à cette époque que les premières réglementations sur les données personnelles ont vu le jour en Europe[7]. Il est facile dès lors d'imaginer à quel point celles-ci sont dépassées actuellement, et pourraient le devenir encore plus, au vu de l'explosion des méga données que promet l'Internet des objets (*Internet of Things*).

Penchons-nous donc quelques instants sur les cinq caractéristiques fondamentales des mégadonnées :

- le **Volume** : le volume est le premier terme qui vient à l'esprit lorsque l'on évoque les mégadonnées. Dans ce domaine, les chiffres donnent le vertige et l'esprit humain a bien du mal à imaginer ce que peut représenter la masse de données générées par les systèmes informatiques. En l'espace de quelques décennies, la taille des bases de données est passée de quelques mégaoctets (10^6) à quelques pétaoctets (10^{15}). Cisco estime même qu'à fin 2016, le seuil d'un zettaoctet (10^{21}) de trafic annuel sera franchi par l'Internet[8]. Le volume des données transmises et stockées est immense et ne cesse de croître. Avec lui, les technologies permettant de gérer et d'exploiter de telles masses de données émergent, notamment chez les plus grands acteurs de l'Internet comme Facebook ou Google.

- la **Vitesse** : dans le domaine de la vitesse, l'objectif qui prévaut est d'absorber et de traiter les données en temps réel, ou de s'en approcher. Là encore, les changements ont été très profonds. En quelques dizaines d'années, des traitements qui étaient réalisés périodiquement, par exemple une fois par mois, par semaine ou par jour ont vu leur fréquence accélérer pour tendre vers le temps réel et la gestion des données en flux. Cette quête de vitesse et de fluidité n'est d'ailleurs pas l'apanage des mégadonnées, c'est une tendance que l'on retrouve dans bien d'autres domaines comme celui du transport de marchandises par exemple.

- la **Variété** : historiquement, l'informatique a commencé par gérer des données essentiellement composées de textes et de nombres, représentés sous forme binaire. Avec l'arrivée du multimédia, ce sont ajoutés les images, les sons, les vidéos, puis toutes sortes d'autres types de données, des documents, des animations, des coordonnées géographiques, des plans, des représentations 3D, etc. Beaucoup d'entreprises sont encore incapables de gérer une telle diversité d'information. Leur organisation et leurs systèmes ne sont pas adaptés pour gérer les images, les tweets, les vidéos et les autres types de données susceptibles de rejoindre leur système d'information. Cette diversité des types de données est par ailleurs conjuguée avec le volume et la vitesse. Lorsqu'il s'agit par exemple de traiter des vidéos, et en particulier en haute-résolution, les fichiers sont très volumineux et les capacités de stockage et traitement requises sont importantes. La variété entraîne également une grande complexité dans le traitement des données. Les systèmes informatiques actuels étant, pour leur grande majorité, incapables de comprendre la sémantique d'un document, d'une image ou d'une vidéo. Les recherches en matière d'intelligence artificielle se focalisent précisément sur les domaines de la compréhension du langage naturel ou la reconnaissance vocale et visuelle, afin de doter les

machines d'une capacité à décoder le sens des textes, des sons, des images et des vidéos. Dans ces différents domaines les progrès sont aussi spectaculaires qu'inquiétants.

– la **Véracité** : ce quatrième V est sans doute un des plus complexes à maîtriser pour l'instant. Il décrit l'idée d'associer aux données un certain niveau de probabilité, un niveau de confiance. Rien de tel qu'un exemple pour bien comprendre de quoi il s'agit. Imaginez un supermarché en ligne qui analyse les achats de ses clients. Le fait qu'un client commande par exemple plusieurs fois des croquettes ou de la litière pour chat permet de déduire avec un très haut niveau de probabilité que ce client possède un chat, sans toutefois en avoir la certitude absolue. La nature des achats, mais aussi leur quantité et leur fréquence permet même sans doute de déterminer statistiquement le nombre de chats possédés, leur âge, peut-être même leur sexe ou leur type. Il n'est pas nécessaire de poser la question au client, il suffit d'observer son comportement d'achat pour en inférer des données supplémentaires, avec un niveau de confiance suffisant pour que la donnée en question soit utile car valorisable commercialement. Dans certains cas, il n'est même plus besoin de définir les règles établissant des liens entre des prémisses et des conclusions particulières. Ce sont les systèmes eux-mêmes qui vont traiter les données afin d'identifier des corrélations statistiquement valables entre différentes données, lesquelles sont parfois insoupçonnables. La création et la gestion de ces données probabilistes est en plein développement, par exemple dans le domaine des recommandations d'achat. Gérer la véracité est cependant problématique car la notion même de probabilité implique qu'il existe une possibilité d'erreur. L'information inférée peut être fausse. Dans l'exemple du propriétaire de chat, les conséquences sont sans doute minimes, mais imaginez ce qu'il peut advenir lorsque l'on infère des éléments plus sensibles comme l'état de santé[9], la grossesse[10], la perte d'emploi, l'orientation sexuelle ou politique, etc. Ces exemples peuvent sembler fictifs mais sont déjà une réalité. Si l'on y réfléchit, la véracité rejoint là la dimension de vitesse, en débouchant sur une vitesse qui est en quelque sorte « supérieure » au temps réel, puisque non contents d'analyser le passé ou le présent, les systèmes se mettent à anticiper et à prédire l'avenir, qu'il s'agisse de vos prochains achats, du solde de votre compte en banque à la fin du mois, ou de la trajectoire d'un piéton qui pourrait croiser celle d'un véhicule automatique.

– la **Valeur** : cette dimension justifie en grande partie les investissements consentis dans le domaine des mégadonnées. Les données ont une certaine valeur, que l'on peut analyser de différentes façons. Pour commencer, les données ont un coût direct, relatif à leur collecte, à leur stockage et à leur traitement. Ensuite, elles sont exploitables pour générer de la valeur, notamment dès lors qu'on les utilise pour augmenter les ventes ou réduire les coûts d'une organisation. Pour prendre un exemple courant, l'historique d'achat d'un client est souvent utilisé comme base pour lui proposer des achats complémentaires. Si ces suggestions sont efficaces, c'est-à-dire qu'elles se traduisent par des achats, elles confèrent bel et bien une valeur aux données sous-jacentes. La valorisation des données est aussi apparente dans le domaine de la publicité sur Internet. Lorsqu'un annonceur souhaite accroître la précision du ciblage de la diffusion d'une publicité en l'adressant à des cibles présentant des caractéristiques – c'est-à-dire des données – particulières, le site support va généralement facturer un surcoût qui matérialise le lien entre ciblage et efficacité, autrement dit le lien entre données et valeur. De telles observations sont courantes, par exemple diffuser un bandeau publicitaire Rectangle sur le site 20min.ch[11] en français coûtera Fr. 35.- pour mille affichages non ciblés et Fr. 45.- pour un affichage ciblé dans une rubrique (donnée comportementale implicite). L'ajout d'un ciblage géographique basé sur la localisation de l'utilisateur coûtera Fr. 5.- additionnels pour mille affichages. Ces différences tarifaires témoignent bien de la valeur que les données

comportementales ou observables (localisation) représentent pour le site 20min.ch (et indirectement pour ses annonceurs).

3. Qui sont les acteurs du monde des données numériques ?

Identifier de façon exhaustive les acteurs du monde des données numériques, ceux qui produisent, extraient et valorisent ce fameux « or gris » comme certains le nomment, représente un défi tant le domaine est vaste et mouvant. La généralisation des bases de données a aussi pour effet de brouiller les frontières entre les véritables acteurs et leurs clients, dans la mesure où aujourd'hui beaucoup d'organisations, à commencer par les plus grandes, gèrent des méga bases de données.

Il est cependant possible de déterminer de grandes familles d'acteurs[12], par exemple :

- les **fournisseurs d'infrastructures** matérielles, logicielles et réseaux. Ces sociétés développent et commercialisent les outils permettant principalement de stocker et de manipuler de grandes quantités de données. On trouve dans cette catégorie les éditeurs de solutions de base de données qu'elles soient relationnelles ou non, distribuées ou centralisées, hébergées dans les entreprises ou dans le nuage (*cloud*).

- les **fournisseurs de solution de traitement et d'analyse** de ces données. Cette catégorie regroupe des acteurs variés dont les outils et les compétences permettent de faire des recherches dans les données, de les analyser, de les traiter statistiquement, de les représenter sous forme visuelle et graphique, etc.

- les **fournisseurs d'applications**, souvent spécialisées dans un domaine spécifique, qui sont capable de valoriser des données pour en faire un usage particulier. Dans cette catégorie, on pourra aussi bien trouver des acteurs spécialisés dans le marketing, que d'autres spécialisés dans les ressources humaines, la logistique, la sécurité, la santé ou l'urbanisme.

- les **sources** et les **agrégateurs de données**[13], qui fournissent des données spécifiques, par exemple sur un territoire (données cartographiques), une population (données sociodémographiques) ou des domaines particuliers (données financières, médicales, météo, médiatiques, etc.).

- les **organisations**, dont le but n'est pas la collecte de données en soi, mais qui constituent des méga bases de données dans le but de servir leur activité. Une entreprise qui décide par exemple de créer une base de données client entre dans cette catégorie. Elle sera susceptible de travailler avec les fournisseurs d'infrastructure, de solutions d'analyse, d'applications, et même éventuellement des sources de données particulières. Dans le cas d'une base client, il serait par exemple imaginable d'utiliser une base d'adresses postale afin d'améliorer la qualité des adresses enregistrées ou de s'appuyer sur une base d'évaluation des risques-clients afin de mieux gérer ces derniers (*credit scoring*). Différentes organisations, généralement de grande taille, gèrent traditionnellement d'importantes quantités de données. On peut par exemple penser aux banques, aux opérateurs de téléphonie, aux grandes entreprises privées (chaîne de supermarchés par exemple), aux institutions publiques (services fiscaux par exemple), et bien sûr aux acteurs du monde digital, et notamment à ses géants comme Google, Amazon, Facebook, Apple, Microsoft ou Alibaba[14] qui gèrent chacun des données sur des centaines de millions d'individus, c'est-à-dire des masses d'informations colossales.

- Les **utilisateurs** eux-mêmes, qui fournissent de plein gré ou non, de nombreuses données, et qui à ce titre participent à cet écosystème.

4. Quelle est la finalité des mégadonnées ?

Une finalité avant tout économique

La question peut paraître étrange, mais lorsque l'on y réfléchit, la réponse n'est pas si évidente. Dans la plupart des entreprises, l'objectif (avoué ou non) de la mise en place de méga bases de données est de générer ou le plus souvent d'accroître le **profit**, en augmentant le chiffre d'affaires ou la marge.

Dans le domaine du marketing, les bases de données sont utilisées pour **développer les ventes**, par exemple en proposant aux clients des produits ou des services (publicité, offres, *cross-selling, up-selling…*) ou réduire les coûts par exemple en **ciblant** mieux les publicités ou les actions de marketing direct grâce aux données clients.

Certaines entreprises collectent aussi des données dans le but d'améliorer leurs produits ou leurs services. En général il s'agit là encore de réduire les coûts (liés aux problèmes de qualité et au service après-vente), et d'accroître les ventes (en mettant en avant la qualité accrue ainsi obtenue).

La constitution de méga bases de données sans but lucratif direct est aussi possible. La mise en place d'une base d'archives historiques, par exemple, pourrait entrer dans une telle catégorie. Les cas existent donc, mais il est vrai qu'ils sont rares, et que le caractère vénal des mégadonnées ne peut être ignoré.

Données et publicité

Les données sont très utilisées dans le domaine de la **publicité** digitale. Le fait de pouvoir cibler la diffusion des annonces en fonction des caractéristiques des utilisateurs est un avantage majeur de l'univers numérique. Le ciblage publicitaire s'appuie donc sur des données de plus en plus nombreuses et précises permettant de qualifier la cible. Il s'agit par exemple de données sociodémographiques ou comportementales. Sur un site éditeur, on s'intéressera aux caractéristiques d'un abonné et aux articles qu'il consulte ; sur un moteur de recherche ces seront les recherches que l'utilisateur effectue ; sur un site e-commerce enfin, les produits que la personne consulte, achète ou a acheté dans le passé. Lorsqu'un annonceur réalise une campagne sur des réseaux sociaux comme Facebook ou LinkedIn, la plupart des informations du profil sont utilisables comme critère de ciblage[15].

Mégadonnées et ciblage publicitaire

Toujours dans le domaine de la publicité, le marché évolue depuis plusieurs années d'un modèle dans lequel l'annonceur achetait de l'espace publicitaire sur des médias (par exemple une page de publicité dans tel magazine), vers un modèle dans lequel l'annonceur achète une **audience**, ou plus exactement le fait que sa publicité soit affichée en face de cette audience.

Historiquement, l'audience était globalement déterminée en réalisant des enquêtes, par exemple sur le lectorat d'un magazine. L'annonceur savait qu'en choisissant tel support dans son plan média, il pouvait espérer toucher par exemple des hommes d'un certain âge ou des femmes d'une certaine catégorie socio-professionnelle.

Aujourd'hui, dans le monde digital, le média, c'est-à-dire le support publicitaire, engrange des données sur chacun de ses visiteurs (par exemple sa localisation géographique, son âge, ses préférences, ses centres d'intérêt, etc.). Ces données sont stockées et gérées dans des systèmes que l'on nomme des DMP pour *Data Management Platforms*.

Les DMP sont spécialement conçues pour gérer des masses de données, en provenance de différents systèmes sources. Un éditeur de média pourra donc y combiner les données de ses abonnés (dont il connait par exemple le sexe, l'adresse postale, etc.), les données comportementales de ses utilisateurs (abonnés ou non qui visitent le site), et même y adjoindre d'autres données provenant de sources tierces. Ces masses de données sont ensuite traitées pour qualifier l'audience et la proposer aux annonceurs.

De son côté l'annonceur pourra par exemple imaginer une campagne qu'il souhaite diffuser aux hommes habitant à Genève et portant un intérêt à un sport ou un domaine spécifique, par exemple le tennis ou la finance. A chaque visite sur le site, différents systèmes vont être activés pour reconnaître et identifier le visiteur et le rapprocher de son profil (des cookies, mais aussi des identifiants mobiles, voire l'empreinte des appareils utilisés[16] ou tout simplement une identification avec nom d'utilisateur et un mot de passe). Le profil va ensuite être analysé afin de voir s'il correspond à ceux que les annonceurs cherchent à atteindre à cet instant, en fonction des campagnes publicitaires en cours.

Si une campagne existe, la publicité sera affichée, ce qui sera dûment enregistré dans la DMP (tel profil s'est vu proposer telle campagne), tout comme le sera la réaction éventuelle de l'internaute par rapport à la publicité en question (visionnement, clic, etc.).

Dans le cas où aucune campagne ne s'adresse au profil qui s'est présenté, l'espace publicitaire, ou plus exactement l'opportunité d'atteindre ce profil, sera proposé sur des plates-formes de vente aux enchères d'espace publicitaire (les *ad-exchange*[17]). Ces marchés électroniques existent aujourd'hui et commercialisent une part significative de l'espace publicitaire sur Internet.

Les annonceurs eux-mêmes travaillent de plus en plus avec des logiciels qui se chargent d'acheter en temps réel des espaces publicitaires correspondant aux profils d'audience visés, et ce, à un niveau individuel. On utilise pour cela des logiciels dits de RTB *Real-Time Bidding*, à savoir la gestion des enchères en temps réel. Cette approche de l'achat d'espace, que l'on nomme l'achat programmatique, est en plein développement. Il s'appuie massivement sur les données et sur le profilage des internautes.

Le reciblage

Le **reciblage** marketing est une application spécifique qui se base sur le comportement de l'internaute. Il s'agit d'un exemple emblématique de l'application des données dans la publicité. Il consiste à afficher des annonces personnalisées en fonction par exemple des derniers produits recherchés, consultés ou non-achetés sur un site e-commerce. La plupart des internautes ont déjà expérimenté ce phénomène au départ un peu surprenant. Il suffit d'avoir consulté des produits sur un site, et a fortiori de les avoir ajoutés dans son panier virtuel (sans conclure l'achat), pour que des publicités portant précisément sur ces produits apparaissent ensuite comme par enchantement sur d'autres sites visités par la suite. Des sites comme Booking.com ou Amazon.com[18] en sont friands depuis plusieurs années.

Cette approche a démontré son efficacité commerciale et des acteurs comme Criteo délivrent chaque année des centaines de milliards de publicités ainsi personnalisées[19] à plus d'1.1 milliards d'internautes et de mobinautes. En Suisse, le réseau adwebster permettrait d'atteindre de façon personnalisée 6.8 millions d'individus soit 94% des internautes[20]. En parallèle, cela signifie que des acteurs comme Criteo ou adwebster – et ce ne sont là que quelques exemples – enregistrent et scrutent le comportement de masses considérables d'individus.

Un ciblage des produits et des contenus

L'application des techniques de ciblage ne se limite pas à la seule publicité. La plupart des grands sites web profilent leurs utilisateurs pour leur présenter des contenus adaptés, qu'il s'agisse de produits, d'articles éditoriaux, de nouvelles sur les réseaux sociaux ou autre.

Un ciblage très controversé

Le débat est intense entre ceux qui voient dans ces systèmes une solution permettant de proposer aux internautes **des publicités mieux ciblées** et donc **plus intéressantes ou utiles**, et ceux qui les considèrent comme étant des **systèmes particulièrement manipulatoires et intrusifs** par rapport à la vie privée des individus… notamment parce qu'ils enregistrent des données très détaillées sur les sites consultés (qui sont leurs partenaires) et les publicités (affichées, cliquées ou ignorées).

L'autre dimension du débat concerne le **rôle de la publicité** dans l'économie digitale. Nul ne peut nier qu'elle contribue à financer au moins en partie certains services en ligne, à commencer par ceux de Google par exemple, puisque la publicité représente près de 90% du chiffre d'affaires de l'entreprise[21]. Les défenseurs de ce modèle le justifient en mettant en avant la qualité des services dont bénéficient *gratuitement* les internautes en échange de cette exploitation publicitaire. Ses détracteurs dénoncent au contraire une monétisation généralisée des utilisateurs qui selon eux sont en réalité le « produit » que Google commercialise à ses annonceurs. Le développement de l'offre publicitaire de Facebook, comme celle des autres grands acteurs du domaine des réseaux sociaux pose le même type de questions.

En terme quantitatif, pour Facebook les revenus publicitaires sont d'environ 1.05 USD/mois/utilisateur au niveau mondial[22]. Techniquement on utilise la notion d'ARPU – *Average Revenue per User*, ou Chiffre d'Affaires Moyen par Utilisateur. Notez que ce chiffre varie très fortement d'un marché à l'autre. Pour la région US & Canada, il avoisine les 4 USD.

Du ciblage à la manipulation

Dans bien des cas, les mégadonnées ont pour objectif de réduire des incertitudes, des risques. L'idée qui prévaut est qu'en ajoutant des données, il est possible de **prévoir** avec plus de précision des phénomènes ou des comportements. En marketing, cette philosophie encourage la création de base de données consommateurs très détaillées et offrant une vision panoramique[23] des relations avec le client (achats, service après-vente, interactions, historique, etc.).

Cette recherche de prédictibilité des comportements s'appuie sur une **vision fondamentalement déterministe** de l'humain, laquelle est difficilement compatible avec la vision d'un homme libre. Les détracteurs de ce courant dénoncent d'une part l'incapacité des systèmes à intégrer toutes sortes d'éléments non quantifiables, et de l'autre une réification de l'humain qui – devenu prévisible – se transforme en littéralement en chose, en marionnette contrôlable par des algorithmes capables de prédire et d'orienter son comportement.

Au niveau marketing, elle remet également sur la table la question du **potentiel manipulatoire** des actions menées, ce d'autant plus que la possibilité d'une manipulation efficace semble parfaitement possible, ce qu'avait attesté l'expérience très controversée menée en 2012 et portant sur la manipulation de l'état psychologique des utilisateurs à travers la modification du contenu de leur mur Facebook[24]. On se souvient que les retombées de cette recherche ont suscité un grand émoi et ont entraîné la définition d'un nouveau cadre pour ce genre d'activités chez Facebook[25].

Sans aller jusqu'à ces extrêmes, il semble que l'exploitation massive du ciblage et du reciblage peuvent enfermer les utilisateurs dans une sorte de bulle informationnelle personnalisée. Cette bulle réduit le champ de vision de l'utilisateur, qu'il soit acheteur (il ne verra plus que des produits liés à son profil et ses recherches), ou lecteur (il ne verra plus que des articles sur les sujets susceptibles de l'intéresser). Le tout est décidé par des algorithmes, des machines logicielles ingérant et disséquant les données et les profils pour décider de ce que ces derniers vont voir ou non. La myopie ainsi induite peut être interprétée comme étant de nature manipulatoire.

Antoinette Rouvroy, chercheure au Fond National de la Recherche Scientifique Belge, analyse les conséquences de ces approches et relève que « l'**hyperpersonnalisation** des environnements numériques, des offres commerciales, voire des interactions administratives, porte moins la menace d'une disparition de la vie privée que celle d'une **hypertrophie de la sphère privée au détriment de l'espace public**. D'une part, il devient de plus en plus rare, pour l'individu, d'être exposé à des choses qui n'ont pas été prévues pour lui, de faire, donc, l'expérience d'un espace public, commun ; d'autre part, les critères de profilage des individus échappent à la critique et à la délibération collective »[26].

Face à ces questions, il est important de garder en tête que l'objectif de la plupart des grandes plates-formes captatrices de données est avant tout commercial. L'objectif est donc souvent plus de satisfaire l'annonceur en lui proposant une audience la plus large, transparente et captive que possible. C'est précisément pour cette raison que Facebook par exemple insiste tant sur ses métriques de succès que sont le nombre d'utilisateurs quotidien et surtout le temps passé chaque jour sur ses plates-formes (Facebook, mais aussi Messenger, Instagram, Whatsapp…). La bataille est une lutte pour **l'attention** des utilisateurs, pour la conquête du « temps de cerveau humain disponible »[27]. Il s'agit de capter l'attention, pour ensuite pouvoir la commercialiser à des annonceurs.

D'autres exemples témoignent aussi du possible transfert de ces prédictions dans le monde réel. Le plus célèbre d'entre eux est sans doute le brevet accordé à Amazon.com concernant une méthode et un système permettant l'expédition anticipée de colis[28], c'est-à-dire un **système logistique** optimisant les délais de livraison en fonction des possibles achats futurs des clients. Là encore, la vision sous-jacente aussi déterministe que stupéfiante.

5. Pourquoi se préoccuper des mégadonnées ?

Une explosion des mégadonnées qui va s'accentuer

Les éléments évoqués précédemment esquissent les contours des questions que pose la massification des systèmes de méga bases de données. En réalité le phénomène explose sur ses cinq dimensions fondamentales (les « cinq V ») et la combinaison de ces effets entraîne une croissance globale encore plus importante des questions posées.

Le Volume par exemple explose sur plusieurs axes. D'abord le **nombre de personnes connectées** à l'Internet à travers des ordinateurs ou des téléphones intelligents ne cesse de croître[29], entraînant en parallèle une croissance du nombre des individus fichés dans les méga bases de données. Déjà les nombres donnent le vertige. A titre d'exemple, le 27 avril dernier, Facebook annonçait ses résultats pour le premier trimestre 2016, avec 1'654'000'000 utilisateurs mensuels[30]. De mémoire d'homme, jamais une telle quantité d'informations sur les humains n'a été regroupée, qui plus est par une seule et même entité.

Ensuite, les **usages continuent à augmenter**. Dans le cas de Facebook, la firme estime que ses utilisateurs passent désormais plus de 50 minutes par jour sur ses différents sites et applications (hors Whatsapp). Or chacun des usages, chaque *connexion*, chaque *clic*, chaque *j'aime*, chaque *partage*, chaque *commentaire*, chaque *publication*, chaque nouvelle *amitié*, génère ses données et ses métadonnées qui sont enregistrées dans les serveurs de l'entreprise. Là encore les volumes donnent le vertige. La plate-forme Facebook[31] ingère un volume de 600 téraoctet par jour, elle stocke plus de 300 pétaoctets de données et en traite plus de 10 pétaoctets par jour. Les autres géants de l'Internet font de même et les données s'accumulent… Les exigences techniques posées par ces masses colossales de données et la nécessité de les traiter efficacement poussent l'innovation technologique comme jamais. Les géants de l'Internet travaillent d'ailleurs de concert avec les universitaires du monde entier pour résoudre les questions posées par ces infrastructures hors-normes et trouver des solutions (propriétaires et *open source* d'ailleurs).

La Variété des données contribue à l'accroissement du Volume. Elle augmente en parallèle avec une explosion des messages, des images ou des vidéos créés, échangés et stockés.

L'Internet des objets est un Internet des données

Avec **l'arrivée des objets connectés**, ces chiffres vont croître encore plus fortement. Prenez par exemple l'automobile. A ses débuts, elle ne produisait aucune donnée. Puis l'électronique embarquée a commencé à générer des données qui sont mémorisées dans le véhicule et lisibles par exemple par le garagiste avant d'effectuer un contrôle de la voiture. A l'heure actuelle, les voitures connectées qui

arrivent sur le marché sont reliées en permanence à des méga bases de données qu'elles alimentent avec toute sorte d'information sur leur position géographique (coordonnées, trajets, etc.), leur état (vitesse, fonctionnement, etc.), le comportement de leur conducteur ou passager, etc. Chaque capteur intégré dans le véhicule produit des informations qui peuvent être transmises et stockées[32]. Cela peut aller du trajet planifié, aux arrêts effectués, en passant par la musique écoutée, les informations affichées sur les écrans intégrés, jusqu'aux images prises par les caméras ou micros embarquées. Avec l'arrivée des véhicules automatiques, qu'il s'agisse de voitures ou de camions, les données générées seront sans doute encore plus importantes…

L'exemple de la voiture est emblématique, mais de très nombreux autres exemples d'**objets connectés** sont concernés. Pensez aux montres intelligentes, aux bracelets captant l'activité, aux appareils d'électronique domestique connectés, aux systèmes médicaux, aux dispositifs installés dans les magasins, ou même aux millions de caméras de surveillance installées dans nos villes et nos bâtiments, etc. En parallèle, l'évolution des titres de transport[33] vers le digital, comme celui des paiements vers les paiements électroniques (par carte à puce, carte sans contact ou téléphone mobile), sont autant de systèmes qui **vont produire massivement des données** sur les individus et leurs comportements.

Une invasion de la sphère professionnelle

La généralisation de ces pratiques dans le **monde professionnel**, notamment par l'intermédiaire des solutions en nuage va également ajouter des masses considérables de données couvrant tout le spectre de l'activité professionnelles de centaines de millions d'individus. Des entreprises comme Microsoft (Office 365[34]), Google (Apps for Work[35]), Salesforce[36] ou Facebook (@Work[37]) affichent des ambitions vertigineuses dans ce domaine.

Simultanément, les solutions techniques permettant de capter, stocker et traiter les données imaginées par les géants du digital deviennent de plus en plus accessibles aux entreprises traditionnelles. A l'heure actuelle, nombre d'organisations publiques et privées sont loin derrière les géants du Net en matière de mégadonnées. Certaines en sont encore à mettre en place leur base de données clients (CRM). Néanmoins, les forces concurrentielles les poussent à exploiter ces mêmes approches de la publicité ciblée et du marketing personnalisé, au point de les inciter fortement à s'équiper d'outils similaires et à construire elles-aussi des méga bases de données.

Les années qui viennent promettent donc une multiplication extrêmement rapide et tout azimut des sources et des quantités de données. C'est notamment pour cette raison qu'il est fondamental de réfléchir aux questions critiques posées par ces phénomènes et leur évolution si rapide qu'ils demeurent encore souvent dans les limbes de l'impensé.

Trois points pour comprendre les principaux enjeux des données

L'objectif ici n'est pas de polémiquer sur le thème des données, mais d'essayer de montrer à travers des éléments très concrets, en quoi les mégadonnées posent des questions aussi importantes que spécifiques.

1. Les mythes du consentement et de la liberté des utilisateurs

Une opposition fondamentale

Antoinette Rouvroy résume parfaitement la **divergence entre les principes** de protection des données et la philosophie du big data. Elle rappelle que « ces quantités massives de données et les Big data entrent en opposition frontale avec les grands principes de la protection des données : la *minimisation* (on ne collecte que les données nécessaires au but poursuivi), la *finalité* (on ne collecte les données qu'en vue d'un but identifié, déclaré, légitime), la *limitation dans le temps* (les données doivent être effacées une fois le but atteint, et ne peuvent être utilisées, sauf exceptions, à d'autres fins que les fins

initialement déclarées). Les Big data, c'est au contraire une collecte maximale, automatique, par défaut, et la conservation illimitée de tout ce qui existe sous une forme numérique, sans qu'il y ait, nécessairement, de finalité établie a priori. L'utilité des données ne se manifeste qu'en cours de route, à la faveur des pratiques statistiques de datamining, de machine-learning, etc. »[38]

L'impossible consentement

Dans de nombreux pays, et en particulier en Suisse[39], des lois sur la protection des données encadrent les pratiques en matière de données concernant des personnes physiques et morales. Ces textes prévoient en général le **consentement éclairé** des personnes, en l'occurrence des utilisateurs, des clients, etc. Or la plupart des politiques de confidentialité destinées à expliquer aux utilisateurs la nature et l'utilisation des données collectées sont purement et simplement incompréhensibles pour le commun de mortels. En vocabulaire technique et expressions juridiques, il est difficile de réellement comprendre la portée du consentement. Elles sont en outre rédigées en des termes extrêmement larges, lesquels permettent aux entreprises une grande souplesse dans l'utilisation ultérieure des données. Ajoutez à cela qu'elles sont susceptibles d'évoluer dans le temps, ce qui n'arrange rien[40], sans oublier que les utilisateurs recourant aux services de plusieurs organisations vont devoir accepter autant de politiques de confidentialité.

Pour prendre un cas concret, observez par exemple l'insistance répétée avec laquelle les réseaux sociaux tentent d'obtenir votre numéro de téléphone « pour vous aider à sécuriser votre compte » comme ils l'affirment. L'objectif présenté est louable et réel. On peut parfaitement imaginer qu'un utilisateur accepte de donner cette information en échange de cette promesse de sécurité. En réalité, l'objectif premier est de capter une donnée à très fort **pouvoir identifiant**. En effet, s'il est très facile de créer des adresses e-mails fictives et gratuites, il est plus difficile de le faire pour un numéro de téléphone. En ajoutant cette précieuse information au profil, l'entreprise gagne une clé fiable d'identification, clé qui va lui permettre d'effectuer des recoupements, par exemple en la croisant avec les numéros récoltées dans les carnets d'adresses partagés par ailleurs par ses utilisateurs.

A la décharge des organisations de bonne volonté, le défi est important. Comment expliciter tous les aspects d'une politique de confidentialité, de façon claire, complète, et synthétique ? Comment concilier les intérêts divergents entre l'organisation et ses clients ? Les acteurs concernés semblent faire des efforts, ou font semblant d'en faire… mais le résultat est encore loin d'être satisfaisant. Paradoxalement peut-être, les géants de l'Internet ne sont pas les moins bons élèves dans ce domaine. Google[41] par exemple propose une rubrique détaillée sur ses pratiques et les outils proposés aux utilisateurs. Facebook[42] fait de même, ce qui ne l'empêche pas d'avoir été mis en demeure en France par la CNIL en février 2016 pour ses pratiques en matière de données personnelles[43]. Les entreprises plus traditionnelles déploient en général beaucoup moins de pédagogie quand il s'agit d'expliciter leurs pratiques en matière de confidentialité. La plupart du temps, les utilisateurs restent réduits à des choix contraints, à savoir accepter d'un clic et les yeux fermés des conditions obscures et définies unilatéralement.

L'absence de réel choix

Face aux questions relatives aux données, l'argument souvent avancé est le **libre choix** offert aux utilisateurs de renoncer à utiliser tel ou tel service. Or, la liberté de consentement dépend de la réalité de ce choix. Si les utilisateurs disposent de véritables alternatives, il est facile de décider s'ils souhaitent ou non utiliser un service. Mais face aux monopoles ou aux oligopoles des géants de l'Internet, il est bien difficile de prétendre disposer d'alternatives dignes de ce nom.

Dans le domaine de la publicité ciblée, l'industrie a proposé des systèmes[44] permettent aux utilisateurs de renoncer au ciblage. Cette **autorégulation** s'avère être en réalité un leurre, car elle s'appuie sur des mécanismes de cookies, et devient donc inopérante dès que l'utilisateur les efface. Les éditeurs de

navigateur ont proposé d'intégrer le renoncement directement dans leurs produits[45]. C'est ce que certains éditeurs appellent le paramètre « *Do not track* » ou « Interdire le suivi », mais en réalité les sites n'en tiennent pas compte correctement et l'industrie publicitaire fait pression pour ne pas y être forcée.

De la même façon, lorsqu'un utilisateur recourt à un service depuis longtemps, qu'il l'utilise pour gérer ses données, ses contacts, ses souvenirs, etc. comme c'est le cas sur Facebook ou d'autres réseaux sociaux, la liberté de choix se réduit considérablement si le fournisseur souhaite modifier sa politique de confidentialité. L'utilisateur devient en quelque sorte otage de la plate-forme à laquelle il est liée par de très **puissants mécanismes d'infofidélisation** (sans évoquer même la dépendance psychologique que semblent susciter ces mêmes plates-formes). Or, les forces de rétention augmentent avec la quantité d'information que chaque utilisateur confie aux plates-formes et par effet réseau[46] avec le nombre d'utilisateur participants. Elle va donc s'accroître avec le temps, puisque les systèmes stockent toujours plus de données. Le nouveau règlement général européen prévoit la possibilité de récupérer ses données (pour les transmettre à un autre prestataire), mais ce droit reste encore très théorique.

Parallèlement, la liberté réelle de **refuser le changement des règles** d'utilisation quand elles sont modifiées se réduit au point de devenir fictive. Et même dans les cas où des alternatives payantes sont proposées à ceux qui souhaitent échapper à la publicité, de nombreuses données sont encore collectées.

Dans tous ces cas, il est indispensable que des options permettant une gestion fine des paramètres de confidentialité soient intégrées directement au cœur des outils afin que le choix ne se résume pas au choix binaire usage / non-usage, mais qu'il autorise un usage adapté aux souhaits des utilisateurs.

Un accroissement des risques

Là encore, ces questions risquent de prendre une importance grandissante dans les années qui viennent. D'abord parce que les plates-formes deviennent tellement massives qu'elles en deviennent **incontournables**, ensuite parce qu'elles s'appuient de façon croissante sur des organisations qui incitent ou contraignent leurs membres à les rejoindre. C'est précisément ce qui se produit lorsqu'une association ou en entreprise utilise les services de Google Apps, Microsoft, Facebook at Work... ou par extension ce qui se passe quand un commercial reçoit un téléphone intelligent ou une voiture de fonction connectée. Dans le même registre, les généreux accords conclus par ces mêmes entreprises avec les Etats, par exemple les ministères de l'éducation[47], peuvent contraindre des masses d'élèves et d'enseignants à les rejoindre, même à leur corps défendant. Similairement, lorsque c'est l'état qui met en place de tels systèmes de mégadonnées, la question du libre choix des citoyens est centrale. Une consultation est par exemple en cours en France sous l'égide du Ministère de la Santé[48] sur le thème « Partager ses données de santé : pour quels bénéfices et à quelles conditions ? ».

La dimension collective du risque

Il faut aussi avoir conscience que la liberté de choix peut également concerner les personnes qui n'utilisent pas un service donné, mais dont des données sont malgré tout enregistrées, le plus souvent à leur insu. C'est un des points que la CNIL en France a soulevé par rapport à Facebook et en particulier à ses widgets que les responsables de sites ajoutent dans leurs pages afin de permettre aux visiteurs d'aimer leurs sites ou d'en partager l'information sur le réseau social. Lorsqu'un non-utilisateur de Facebook visite un de ces sites, des données le concernant sont enregistrées par les serveurs de Facebook[49]. C'est aussi le cas pour tous les sites qui utilisent Google Analytics, Google AdSense et une multitude d'autres services disponibles sur Internet. En théorie, tout cela est couvert par l'acceptation de politiques de confidentialité, dont nous avons vu plus haut qu'elles sont incompréhensibles pour le commun des mortels. Cette question des profils fantômes (*shadow profiles*), ces profils de non-utilisateurs des plates-formes, reste très sensible[50].

Un autre aspect important est la **dimension collective**[51] de la protection des données personnelles. Il est assez rare que l'on y fasse référence, mais celle-ci est en fait très présente. Par exemple dès lors qu'un utilisateur d'un réseau social ou d'une application mobile quelconque autorise l'accès à son carnet d'adresse, il expose et partage de facto des données personnelles concernant d'autres personnes, utilisatrice ou non du service concerné, sans que celles-ci aient le moins du monde consenti à ce partage. Similairement, si une application demande l'accès à un compte de messagerie ou aux photos contenues dans un téléphone mobile, elle peut obtenir des données sur des tiers. Le même phénomène se produit implicitement dès que vous envoyez un mail à un utilisateur de Google Mail. Le destinataire a accepté en théorie la politique de confidentialité de Google, mais l'expéditeur ne l'a pas fait, et pourtant toutes les données associées à son message finiront bel et bien par être traitées par Google[52].

Des systèmes pas si neutres

Le dernier point est peut-être celui du mythe de la **neutralité** des technologies, voire des données elles-mêmes. Cet argument est souvent utilisé pour dédouaner les concepteurs des systèmes de leurs responsabilités et de proche en proche dédouaner des secteurs entiers des leurs. En réalité, dans le domaine des mégadonnées, les systèmes ne sont pas neutres. Ils sont conçus, optimisés et utilisés dans un but essentiellement tourné vers la maximisation du profit de ceux qui les détiennent. La protection de la vie privée y est vécue comme une simple contrainte qu'il s'agit d'intégrer et de respecter *a minima*, de façon à éviter les problèmes juridiques ou des réactions trop négative des utilisateurs, ces risques pouvant nuire précisément à la maximisation des profits (risques risque d'amende, risque réputationnel, risque de renoncement des utilisateurs…).

2. La concentration des données

Les risques de l'hyperconcentration

Au-delà des 5 V, la question de la concentration des mégadonnées est centrale. La philosophie de l'Internet, au moins à ses débuts, était, on s'en souvient, basée sur une forte **décentralisation** des systèmes, laquelle était supposée garantir la résilience du *tout* en cas de défaillance de certaines *parties*. L'évolution de l'économie de l'Internet est fort différente puisqu'elle aboutit à des **concentrations** apparemment inéluctables de services, d'utilisateurs, et de données. Les maîtres du secteur regroupent plus d'un milliard de profils et des milliards d'informations sur ces profils.

Dans le domaine des services secrets, les révélations d'Edward Snowden ont montré que les états ne sont pas en reste et ont eux aussi constitué de gigantesques bases de données sur leurs citoyens comme sur ceux d'autres états[53]. La collaboration imposée aux acteurs de l'Internet soulève d'ailleurs à ce sujet des questions fondamentales[54]. Les différents textes de loi approuvés ou en projet en Europe, développés notamment suite aux attentats terroristes légitiment la collecte d'information. En Suisse le texte de la Loi fédérale sur le renseignement[55] (LRens) va dans le même sens. Ce texte sera soumis à un référendum qui aura lieu sans doute en septembre prochain. Parallèlement, un autre référendum a été lancé contre la révision de la Loi fédérale sur la surveillance de la correspondance par poste et télécommunication[56] (LSCPT). Le délai référendaire court jusqu'au 7 juillet 2016. L'objectif de ce document étant plutôt de se concentrer sur les aspects marketing des mégadonnées, nous n'allons pas creuser plus en détail cet aspect de la question, ce qui ne signifie pas qu'il n'est pas important, bien au contraire.

Un profilage affiné

Cette concentration pose de nombreuses questions. D'abord parce que la **masse** diversifiée d'information récoltée permet d'établir des **profils** toujours plus précis sur les individus (y compris quand les données sont censées être anonymes, mais qu'en réalité elles permettent, par croisement entre-elles, d'identifier des personnes précises). Ensuite parce que la combinaison des différentes

données permet d'en inférer d'autres, qui elles aussi augmentent la finesse de profils. La dimension collective des données joue également son rôle ici puisque la combinaison des données de différentes personnes peut entrer dans la création de ces nouvelles données. Pour prendre quelques exemples concrets. La personne A partage une photo datée et géoréférencée sur laquelle la personne B apparaît. Par déduction on sait que la personne B se trouvait à l'endroit en question à la date donnée.

Une sécurisation impossible ?

La seconde question concerne **la sécurisation des données** recueillies, car la concentration des données concentre simultanément la valeur, et donc l'attractivité des systèmes pour qui voudrait s'y attaquer, sans parler même des risques d'erreurs humaines ou d'incompétence qui peuvent tout à coup exposer une grande quantité de données.

La concentration des informations incite les pirates, espions et ennemis à redoubler d'effort pour déjouer les mécanismes de sécurité, puisqu'elle promet une récompense proportionnelle.

Et dans ce domaine les exemples foisonnent et inondent les médias, depuis les Panama Papers[57] ou les LuxLeaks[58], jusqu'aux célèbres cas Ashley Madison[59], en passant par les récents cas des registres de citoyens turcs[60] ou mexicains[61].

Dans tous ces exemples, on parle de millions de documents ou de données dévoilées. Les cas connus vont d'ailleurs sans doute se multiplier car différentes législations vont consacrer **le droit d'être informé** en cas de piratage des données. Ce point est intégré dans le futur Règlement général européen sur la protection des données, alors qu'un texte allant dans ce sens est en discussion aux États-Unis[62].

Entre données marketing et surveillance généralisée des populations

Les questions relatives à la sécurité informatique sont complexes, mais ces exemples démontrent que ni les entreprises privées, ni les états ne peuvent aujourd'hui **garantir la sécurité** des systèmes. Par ailleurs, il faut bien comprendre qu'entreprises privées et états n'ont même pas forcément intérêt à ce que les systèmes soient **trop sécurisés**, les uns pour des questions de coût, de complexité ou de performance, les autres parce qu'une sécurisation accrue empêche - ou tout au moins gêne – les activités d'espionnage et de surveillance.

Les citoyens ont d'ailleurs une attitude que certains jugent ambiguë à propos de leurs données personnelles. D'un côté ils luttent vigoureusement contre « l'état fouineur », tout en acceptant de l'autre de livrer volontairement ou négligemment, des quantités très importantes de données personnelles à des entités privées, souvent étrangères et dont le but avoué est purement commercial.

3. La transparence asymétrique

Asymétrie, panoptique et liberté

Il existe une grande **asymétrie** de pouvoir et de transparence entre les parties concernées par ces questions de protection des données. Lorsqu'un internaute partage une information, ou accepte une politique de confidentialité pour accéder à un service en ligne, il ne dispose que d'une information très vague (voire inexistante) sur des questions aussi importantes que :

- la nature exacte des données collectées
- la façon dont elles sont stockées
- les traitements qui leur sont appliqués (inférence, algorithmes, intelligence artificielle)
- les utilisations, notamment commerciales, qui en sont faites pour valoriser

- avec qui elles seront, le cas échéant, échangées
- la manière dont elles seront sécurisées
- la durée de leur rétention
- ou même les personnes ou les organismes qui auront accès aux données en question.

Le manque de transparence est réel et quasi généralisé. Cette opacité n'est pas favorable au respect des règles destinées à protéger les individus et leurs droits fondamentaux. Il s'agit par exemple des droits définis dans les articles 6, 7 et 8 de la Charte des droits fondamentaux de l'Union Européenne qui définissent le droit à la liberté et à la sûreté, le respect de la vie privée et familiale, et la protection des données à caractère personnel[63].

La transparence asymétrique induit également des problèmes en matière de liberté individuelle et d'autocensure. Ces questions ont été réfléchies et théorisées notamment par Michel Foucault dans son analyse de notion de **panoptique** telle que proposée par Jeremy Bentham. Ce « type d'architecture carcérale […] permet à un gardien, logé dans une tour centrale, d'observer tous les prisonniers, enfermés dans des cellules individuelles autour de la tour, sans que ceux-ci puissent savoir s'ils sont observés »[64].

Comme l'explique Pierre Ropert sur le site de France Culture[65], « le panoptique, c'est finalement faire de la visibilité la prison. On cesse d'enfermer pour mettre en pleine lumière. L'essentiel, c'est que l'on se sache surveillé. Le pouvoir est automatisé et désindividualisé, puisqu'il n'est pas vu ». Michel Foucault écrit « l'effet du panoptique est d'induire chez le détenu un état conscient et permanent de visibilité qui assure le fonctionnement automatique du pouvoir. (...) La surveillance est permanente dans ses effets, même si discontinue dans son action ». C'est précisément ce qui se passe dans le monde digital et qui **influence** très insidieusement les comportements.

Vous l'avez peut-être ressenti par exemple en signant une pétition digitale. Si la liste des signataires est affichée publiquement, il se peut que vous hésitiez à la valider. De la même façon l'utilisateur peut hésiter à aimer, commenter ou partager des contenus sur les réseaux sociaux s'il ne sait pas précisément qui pourra voir ces actions (par exemple son employeur, ses collègues). Similairement, l'utilisateur choisira peut-être de partager certains contenus sur une plate-forme plutôt que sur une autre pour tenter de gérer la visibilité et donc la confidentialité relative des éléments en question (par exemple publier sur Facebook ou sur LinkedIn). De façon tout à fait équivalente, la communication orale sera parfois préférée au mail écrit, car elle ne laisse pas ou peu de trace. Et si vous souhaitez d'autres exemples, imaginez le cas des personnes qui *de bonne foi*, peuvent souhaiter consulter des sites djihadistes pour mieux comprendre le phénomène afin de le combattre. La France pourrait bien interdire pénalement une telle consultation. Le sujet est évidemment aussi sensible que controversé, quelle que soit l'issue du processus législatif[66].

Un cadre défini unilatéralement par des privés

Associée à la concentration des données, l'asymétrie et l'autocensure qu'elle induit, entraîne les utilisateurs vers un **conformisme** dont le moule, le modèle, est défini par un nombre de plus en plus restreint d'organisations. Une sorte de pensée unique, se met en place, appuyée sur la définition privée, unilatérale, imposée, a-territoriale et universelle de **l'acceptable**, de ce que l'on peut dire, publier ou faire. Une illustration emblématique de ce phénomène est le cas de la censure du « téton féminin » et du nu en général imposée par Facebook à ses utilisateurs[67]. Dans ce cas précis, il s'agit bien d'une règle imposée par la plate-forme, face à laquelle l'utilisateur ne peut rien. Or il existe d'autres approches, comme celle de Google SafeSearch[68] qui permet d'activer ou non un filtre destiné à masquer certains

contenus ou liens « explicites ». La solution n'est peut-être pas parfaite, mais elle a le mérite d'offrir un choix à l'utilisateur.

Si l'on pousse la réflexion, l'autocensure des utilisateurs peut aller jusqu'au renoncement à utiliser tel ou tel service, par manque de confiance. Le paradoxe sur ce point est que le renoncement lui-même devient suspect[69] dans un monde où ces pratiques se sont généralisées.

On peut penser que ces questions vont devenir de plus en plus sensibles, à mesure que l'éventail des domaines dans lesquels des données sont collectées s'élargit (domaine médical, paiements, déplacements, etc.), et que les outils de recherche et d'analyse s'affinent. Les utilisateurs comprennent, à force d'exemples[70], l'impact que leurs traces numériques peuvent avoir, ce qui renforce l'autocensure et contribuera – on peut l'espérer- à un regain de contrôle sur les données personnelles.

Quelles pistes de « solutions » face aux mégadonnées ?

L'absence de solutions simples

Face au développement tout azimut des mégadonnées et face à la quantité de traces que les systèmes génèrent, il est difficile de proposer des solutions simples, d'autant plus que ces solutions sont généralement contraires aux objectifs de maximisation des profits des acteurs concernés.

Laissons de côté les mesures, souvent **illusoires**, d'effacement des cookies ou de blocage des publicités, tant elles sont loin de résoudre vraiment les problèmes.

Prise de conscience

La première véritable étape passe par la prise de conscience de l'importance des données personnelles et de leur protection. Cette prise de conscience, qui peut s'appuyer sur de la formation et de l'information adéquate, doit s'accompagner d'une attention portée aux choix des logiciels et des services utilisés, notamment ceux qui sont proposés dans le nuage. Ces derniers sont en effet souvent plus risqués, puisqu'ils impliquent en général de confier les fichiers et les données au producteur du service, sans parler de l'usage du service, qui devient observable en temps réel.

Lorsqu'il existe des **alternatives** moins intrusives, ou disons plus respectueuses de ces questions, il est bien entendu recommandé d'y souscrire. Néanmoins, certains des services sont difficilement contournables, tant ils sont généralisés. Dans ces cas, il est important de faire l'effort de prendre connaissance de leurs pratiques en matière de protection des données, et d'étudier les possibilités qu'ils proposent pour en gérer les éventuels paramètres. Les grands acteurs comme Google ou Facebook offrent à ce titre des options insuffisantes, certes, mais intéressantes malgré tout.

Résistance

Lors d'une récente conférence[71] à Lausanne, un représentant de la petite association Framasoft[72] a présenté le **projet d'alternative libre** aux grands services en ligne[73]. La vision proposée est celle d'une re-décentralisation et d'une réappropriation des outils et services numériques par les individus et des entités locales réparties dans les territoires. En matière de protection des données, c'est une approche intéressante qui répond comme en écho à d'autres mouvements, orientés par exemple vers la consommation de produits alimentaires produits localement et distribués en circuit court. L'analogie avec la philosophie des locavores est pleine de sens et montre qu'il est possible de faire des choix différents, même s'ils sont partiels, imparfaits et à petite échelle.

Un cadre contraignant

Pour dépasser les actions individuelles ou associatives et répondre d'égal à égal aux très grandes entreprises qui sont à l'œuvre dans les mégadonnées, il faut un **cadre juridique fort**. Ce cadre juridique est un élément fondamental en matière de protection des données, notamment parce que c'est un

moyen de poser des limites à la toute-puissance des grandes organisations. Dans ce domaine, c'est le citoyen électeur et militant qui peut influer sur les règles proposés par le parlement, et lorsque ces règles ne sont pas appropriées, la démocratie participative suisse offre la chance de pouvoir proposer ou soutenir des initiatives ou des référendums pour faire changer les choses ou les définir. De puissants lobbies sont à l'œuvre dans ces domaines et il est essentiel que la société civile soit présente dans les débats pour faire entendre ses avis. De nouveau, le préalable s'inscrit dans la formation et l'**éveil des consciences** à la gravité des phénomènes et des enjeux.

Par rapport au cadre, cela peut concerner les limites définissant la capture des données en amont, mais aussi leur stockage, leur traitement, leur transmission, etc. Toutes les étapes de la chaîne de valeur des données sont concernées. Le cadre peut aussi définir les règles de transparence à respecter, la façon dont il est nécessaire de proposer de vrais choix aux utilisateurs afin que ceux-ci puisse contrôler leur données et l'usage qui en est fait, sans être pénalisés injustement dans leur usage des produits et services. Des mécanismes de contrôle et des sanctions proportionnelles aux enjeux doivent aussi être définies.

Des mesures à prendre également par domaine

Les questions relatives aux données vont s'immiscer dans de nombreux pans du droit, la protection des données bien sûr, mais aussi la protection de la vie privée, les règles de droit du travail, le droit de la consommation, les conditions relatives aux services publics (en particulier dans le cadre des projets dits d'*open data*), etc. La vigilance est de mise dans tous ces domaines. D'une certaine façon, l'objectif devrait être de **renforcer la transparence des organisations**, et notamment des grandes entreprises, car c'est là que se concentrent les enjeux et les risques, tout en **accroissant la protection des données et de la sphère privée des individus**. La directive européenne protégeant le « secret des affaires » adoptée par les eurodéputés le 13 avril 2016, malgré une pétition signée par plus de 500'000 personnes[74], ne va hélas pas dans ce sens et c'est fort regrettable.

Les organisations peuvent-elles faire preuve de responsabilité ?

Au niveau des organisations, qu'il s'agisse d'entreprises ou d'autres organisations, notamment publiques, il va également falloir faire preuve de retenue et de **responsabilité** par rapport à la collecte et au traitement de données personnelles. Dans leurs pratiques en matière de marketing, de gestion de la relation client, de vente, de gestion des ressources humaines ou d'informatique, pour ne prendre que quelques exemples, les organisations feront face à des choix dans lesquelles elles devront réaliser des arbitrages en matière de protection des données. Espérons qu'elles aussi se comporteront de façon pleinement responsable dans ce domaine.

Depuis la vision jusqu'à la mise en œuvre

Ces choix seront à faire au niveau stratégique, sur la philosophie générale, mais aussi au niveau opérationnel dans l'implémentation et la gestion des systèmes qui manipulent effectivement les données. Dans ce domaine, la prise au sérieux de la sécurité, tant au niveau de l'analyse des risques qu'au niveau des mécanismes de protection, sera déterminante. L'adoption de **méthodes de conception** intégrant la sécurité et le respect de la sphère privée sont essentielles (approche dite de *security by design* et *privacy by design*).

Des compromis protégeant l'essentiel

Enfin, protéger les données, respecter l'identité numérique des individus et garantir la sphère privée des personnes, c'est aussi, sans doute faire des **compromis**. C'est-à-dire trouver ensemble **les justes équilibres** capables de mettre en balance les **droits humains fondamentaux** face à des forces aussi puissantes que la maximisation des profits, l'efficacité et le rationnel des systèmes, l'hygiénisme, le principe de précaution, le mythe sécuritaire ou le solutionnisme technologique. C'est à ce prix qu'il sera

possible de protéger l'humain contre ces systèmes afin qu'il reste puisse rester libre et préserver ce qui le différencie des machines.

Quelques références

Sites web

Préposé fédéral à la protection des données et à la transparence (PFPDT)
<http://www.leprepose.ch>

Commission Nationale de l'Informatique et des Libertés (France)
<https://www.cnil.fr/>

Commission européenne – Protection des données
<http://ec.europa.eu/justice/data-protection/index_fr.htm>

La Quadrature du Net
<https://www.laquadrature.net/fr/vie_privee>

Le guide en ligne, Contrôle tes données
<https://controle-tes-donnees.net/>

Electronic Frontier Foundation
<https://www.eff.org/issues/privacy>

Livres

BABINET, Gilles - *Big Data, penser l'homme et le monde autrement*, Le Passeur, février 2015.
MOROZOV, Evgeny - *Pour tout résoudre cliquez ici - l'aberration du solutionnisme technologique*, FYP Editions, septembre 2014.
MOROZOV, Evgeny - *Le mirage numérique : Pour une politique du Big Data*, Les Prairies Ordinaires, octobre 2015
SADIN, Eric - *La vie algorithmique*, Critique de la raison numérique, L'échappée, collection "Pour en finir avec", 288 pages, mars 2015.
SADIN, Eric – *L'humanité augmentée*, L'administration numérique du monde, L'échappée, collection "Pour en finir avec", mai 2013.

Notes

[1] Journal officiel du 22/08/2014.
[2] Ces thèmes sont cruciaux et l'on ne peut qu'encourager le lecteur à se reporter à des analyses critiques comme :
http://www.humanrights.ch/fr/droits-humains-suisse/interieure/protection/securite/suisse-loi-renseignement
https://www.amnesty.ch/fr/themes/surveillance/docs/2015/raisons-surveillance-de-masse
http://jenairienacacher.fr/
https://www.laquadrature.net/fr/Vie_privee
[3] ABITEBOUL, Serge. *Sciences des données : de la logique du premier ordre à la Toile : Leçon inaugurale prononcée le jeudi 8 mars 2012 In : Sciences des données : de la logique du premier ordre à la Toile : Leçon inaugurale prononcée le jeudi 8 mars 2012* [en ligne]. Paris : Collège de France, 2012 (généré le 24 avril 2016). Disponible sur Internet : <http://books.openedition.org/cdf/529>. ISBN : 9782722601710.
[4] GOOGLE. *Updating our privacy policies and terms of service*. January 24, 2012 (consulté le 24 avril 2016). Disponible sur Internet : <https://googleblog.blogspot.ch/2012/01/updating-our-privacy-policies-and-terms.html>.
[5] Wikipedia, (consulté le 24 avril 2016). <https://fr.wikipedia.org/wiki/Big_data>
[6] Ce modèle trouve ses racines dans une version à 3V : Volume, vélocité et variété en 2001, avec LANEY, Doug. *3D Data Management : Controlling Data Volume, Velocity, and Variety*. 6 Février 2001 (consulté le 24 avril 2016). Disponible sur Internet : <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Ce modèle a évolué et s'est vu ajouter deux V supplémentaires pour Véracité et Valeur, comme dans les infographies d'IBM : <http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg> et <http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data>.

[7] Rappelons qu'en France, la CNIL – Commission nationale de l'informatique et des libertés – a été créée en 1978. Voir à ce sujet <https://fr.wikipedia.org/wiki/Commission_nationale_de_l%27informatique_et_des_libert%C3%A9s>.
[8] CISCO. *The Zettabyte Era—Trends and Analysis*. 23 Juin 2015, (consulté le 24 avril 2016). Disponible sur Internet : <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html>.
[9] Lire par exemple Adam GAFFNEY, *Google Knows You're Sick*, March 22, 2016, (consulté le 2 mai 2016), <https://newrepublic.com/article/131843/google-knows-youre-sick>.
[10] Nous faisons référence notamment à un cas célèbre relaté dans les médias. Kashmir HILL, *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did* , Forbes, FEB 16, 2012, (consulté le 2 mai 2016), <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>.
[11] Exemple basé sur les données Médias de 20 minutes (consulté le 27 avril 2016). Disponible sur Internet : < http://donneesmedias.20min.ch/20-minutes/onlinemobile-fr/tarifs/desktop/>. CPM Prix W-CH 2016 pour un bandeau Rectangle en ROS : 35.- contre 45.- pour un Channel particulier. Option Geo-Targeting Desktop : plus 5.– CPM.
[12] Cette segmentation est inspirée de TURK, Matt, *Is Big Data Still a Thing? (The 2016 Big Data Landscape)*, 1er février 2016, (consulté le 27 avril 2016). Disponible sur Internet : <http://mattturck.com/2016/02/01/big-data-landscape/> et en particulier de l'infographie <http://mattturck.com/wp-content/uploads/2016/03/Big-Data-Landscape-2016-v18-FINAL.png>.
[13] Le terme « *data broker* » ou « *information borker* » est employée pour décrire ces agrégateurs (et commerçants) de données. Acxiom est un exemple de telles sociétés. Voir <http://www.acxiom.com/data-packages/>. Voir aussi <https://www.privacyrights.org/content/data-brokers-and-your-privacy>
[14] Lire à ce sujet les très intéressantes études sur les GAFAs éditées par FaberNovel, GAFANOMICS, New Economy, New Rules, 2014, (consulté le 27 avril 2016), < http://www.fabernovel.com/work/study-gafanomics-new-economy-new-rules/> et GAFANOMICS Season 2, Four superpowers to outperform in the network economy, 2015, (consulté le 27 avril 2016), <http://www.fabernovel.com/work/study-gafanomics-2-4-superpowers-network-economy/>. On ajoute en général aux GAFAs occidentales d'une part Microsoft, et d'autres les « clones » chinois des GAFAs que sont des entreprises commes Alibaba, Baidu, Sina Weibo, Youku, etc.
[15] Voir à titre d'illustration les Options de ciblage publicitaire de Facebook, <https://www.facebook.com/business/help/433385333434831>
[16] Des techniques existent pour identifier statistiquement les ordinateurs ou les téléphones mobiles sur la base des caractéristiques techniques dévoilées à travers un navigateur ou une connexion Internet. Des solutions sont commercialisées pour calculer ces empreintes et les comparer. Elles sont utilisées par exemple dans le domaine de la sécurité pour détecter une connexion depuis un appareil inhabituel. A titre d'illustration, la société IOvation affirme déterminer les empreintes de plus de 3 milliards d'appareils (source < https://www.iovation.com/>). Lire également < http://motherboard.vice.com/blog/device-fingerprinting-can-track-you-without-cookies-your-knowledge-or-consent>.
[17] Par analogie au *stock exchange*, il s'agit de bourses, de marchés, permettant l'achat et la vente d'espace publicitaire.
[18] Lire par exemple *Amazon Begins Display Ad Retargeting*, AdExchanger, April 11th, 2011, < http://adexchanger.com/ad-exchange-news/amazon-begins-display-ad-retargeting/>
[19] En 2015, Criteo aurait diffusé 710 milliards de publicités (source : Criteo, http://www.criteo.com/fr/about-us/).
[20] Source : adwebster, http://www.adwebster.com/Pages/advertiser/why-adwebster.aspx
[21] Source : https://abc.xyz/investor/news/earnings/2016/Q1_alphabet_earnings/
[22] Il s'agit d'une estimation grossière, basée sur le chiffre d'affaire trimestriel annoncé pour Q1 2016, rapporté au nombre d'utilisateur actif mensuel de Facebook. Source : < http://investor.fb.com/results.cfm>
[23] L'expression « vision à 360° » est consacrée pour décrire cette approche.
[24] MEYER, Michelle N., *Everything You Need to Know About Facebook's Controversial Emotion Experiment*, Wired Magazine, 30.06.2014, (consulté le 27 avril 2016), <http://www.wired.com/2014/06/everything-you-need-to-know-about-facebooks-manipulative-experiment/>. Et l'article publié à cette occasion, par Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, Proceedings of the National Academy of Sciences of the United States of America (PNAS), first published June 2, 2014, (consulté le 27 avril 2016), <http://www.pnas.org/content/111/24/8788.full.pdf>.
[25] Mike SCHROEPFER, Chief Technology Officer, Facebook, *Research at Facebook*, October 2, 2014, (consulté le 27 avril 2016), <http://newsroom.fb.com/news/2014/10/research-at-facebook/>.
[26] Lire dans la série « Entretien autour de l'informatique », celui d'Antoinette Rouvroy, *Big data : l'enjeu est moins la donnée personnelle que la disparition de la personne*, 22 janvier 2016, par Serge Abiteboul et Christine Froidevaux, (consulté le 2 mai 2016), < http://binaire.blog.lemonde.fr/2016/01/22/le-sujet-de-droit-au-peril-de-la-gouvernementalite-algorithmique/ >.
[27] Une expression utilisée notamment par Patrick Le Lay en 2004 alors qu'il était président-directeur général du groupe TF1. Voir <https://fr.wikipedia.org/wiki/Temps_de_cerveau_humain_disponible>.
[28] United States Patent N° US8,615,473 B2, en date du 24 décembre 2013, *Method and system for anticipatory package shipping*. (consulté le 27 avril 2016), <http://pdfpiw.uspto.gov/.piw?PageNum=0&docid=08615473&IDKey=28091688238A%0D%0A&HomeUrl=http%3A%2F%2Fpatft.uspto.gov%2Fnetacgi%2Fnph-Parser%3FSect1%3DPTO2%2526Sect2%3DHITOFF%2526p%3D1%2526u%3D%25252Fnetahtml%25252FPTO%25252Fsearch-bool.html%2526r%3D1%2526f%3DG%2526l%3D50%2526co1%3DAND%2526d%3DPTXT%2526s1%3D%252522anticipatory%252Bpackage%252522%2526OS%3D%252522anticipatory%252Bpackage%252522%2526RS%3D%252522anticipatory%252Bpackage%252522>. Lire également BENSINGER , Greg, *Amazon Wants to Ship Your Package Before You Buy It*, The Wall Street Journal, 17 Janvier 2014, <http://blogs.wsj.com/digits/2014/01/17/amazon-wants-to-ship-your-package-before-you-buy-it/>
[29] La volonté des GAFAs de faciliter d'ailleurs l'accès au réseau est sans équivoque. Facebook ne fait pas mystère de ses projets de fourniture d'accès *gratuit* à travers notamment son projet internet.org (voir < https://info.internet.org/en/>). Google met en place des propositions semblables au travers de projets comme l'accès Internet à travers des ballons, Google Loon (voir < https://www.google.com/loon/>), à travers la fibre optique, Google Fiber (voir < https://fiber.google.com>) ou à travers des bornes WiFi publiques déjà testées à New York (voir < https://www.intersection.com/> et < http://www.sidewalklabs.com/>, une entité qui appartient à Alphabet, la maison mère de Google < https://abc.xyz/>).
[30] Source : < http://investor.fb.com/results.cfm>.
[31] Regarder à ce sujet la conférence de mars 2015, *F8 2015 - Lifting the Curtain: the Data Infrastructure Behind Facebook Apps*, (consulté le 28 avril 2016), <https://www.youtube.com/watch?v=T9YbeJ4DMuM> et <http://www.opencompute.org/>.
[32] A titre d'exemple, voir MUSK, Elon, *A Most Peculiar Test Drive*, February 13, 2013. (consulté le 28 avril 2016), <https://www.teslamotors.com/blog/most-peculiar-test-drive>.
[33] A titre d'exemples les systèmes comme Navigo en Ile de France (http://www.navigo.fr/) ou le plus récent SwissPass des CFF sont emblématiques (https://www.swisspass.ch).
[34] Voir <https://products.office.com/fr/business/explore-office-365-for-business>
[35] Voir <https://apps.google.com/>
[36] Voir <https://www.salesforce.com/fr/>
[37] Voir <https://work.fb.com/>
[38] Op. cit.
[39] En Suisse, la Loi fédérale sur la protection des données (LPD), <https://www.admin.ch/opc/fr/classified-compilation/19920153/index.html> ou en France la Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, <https://www.cnil.fr/fr/loi-78-17-du-6-janvier-1978-modifiee>. Au niveau européen, la Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:FR:HTML> et au niveau Européen le Règlement général sur la protection des données, adopté en avril 2016 par le Conseil européen et le Parlement européen. Voir <http://www.consilium.europa.eu/fr/policies/data-protection-reform/>. Le règlement et la directive associée devraient entrer en vigueur au printemps 2016 et seront d'application à compter du printemps 2018. Lire également <https://www.cnil.fr/fr/adoption-du-reglement-europeen-par-le-parlement-europeen-un-grand-pas-pour-la-protection-des-donnees>.
[40] Le site de Google par exemple affiche 17 versions historiques de ses règles de confidentialité publiées entre 2005 et 2016, (consulté le 28 avril 2016), <https://www.google.com/policies/privacy/archive/>

[41] Voir < https://www.google.com/policies/>
[42] Voir < https://www.facebook.com/full_data_use_policy> et < https://www.facebook.com/about/privacy/>
[43] Voir < https://www.cnil.fr/fr/la-cnil-met-publiquement-en-demeure-facebook-de-se-conformer-dans-un-delai-de-trois-mois-la-loi> et l'article de UNTERSINGER, Martin, Données personnelles : le virulent réquisitoire de la CNIL contre Facebook, Le Monde.fr, 09.02.2016, (consulté le 28 avril 2016), <http://www.lemonde.fr/pixels/article/2016/02/09/donnees-personnelles-le-virulent-requisitoire-de-la-cnil-contre-facebook_4861621_4408996.html>
[44] Voir < https://www.aboutads.info/>.
[45] Voir < https://en.wikipedia.org/wiki/Do_Not_Track>. Pour Google Chrome, voir *Activer ou désactiver la fonctionnalité Interdire le suivi*, <https://support.google.com/chrome/answer/2790761?p=settings_do_not_track&rd=1 >. Pour Firefox, *Comment puis-je activer l'option « ne pas me pister » ?* < https://support.mozilla.org/fr/kb/comment-activer-option-ne-pas-pister >.
[46] Voir < https://fr.wikipedia.org/wiki/Effet_de_r%C3%A9seau>.
[47] Pour la France, voir par exemple l'accord controversé avec Microsoft < http://www.education.gouv.fr/cid96030/numerique-a-l-ecole-partenariat-entre-le-ministere-de-l-education-nationale-et-microsoft.html> et <http://cache.media.education.gouv.fr/file/Partenaires/17/7/convention_signee_506177.pdf>
[48] Voir <http://www.faire-simple.gouv.fr/bigdatasante>
[49] Voir <https://www.facebook.com/help/cookies/update>. Facebook aurait même déposé une demande de brevet sur le suivi des activités des internautes en dehors du réseau social. Voir United States Patent N° US 2011/0231240 A1, en date du 22 septembre 2011, *Communicating information in a social network system about activities from another domain*. (consulté le 29 avril 2016), <http://pdfaiw.uspto.gov/.aiw?PageNum=0&docid=20110231240&IDKey=9D6001649EF2&HomeUrl=http%3A%2F%2Fappft.uspto.gov%2Fnetacgi%2Fnph-Parser%3FSect1%3DPTO2%2526Sect2%3DHITOFF%2526u%3D%25252Fnetahtml%25252FPTO%25252Fsearch-adv.html%2526r%3D1%2526p%3D1%2526f%3DG%2526l%3D50%2526d%3DPG01%2526S1%3D20110231240.PGNR.%2526OS%3Ddn%2F20110231240%2526RS%3DDN%2F20110231240>
[50] Lire par exemple <https://www.propublica.org/article/its-complicated-facebooks-history-of-tracking-you>.
[51] Des chercheurs de l'Ecole Polytechnique Fédérale de Zurich ont par exemple étudié ce phénomène dans leur publication : Emre Sarigol, David Garcia et Frank Schweitzer, ETH Zurich, *Online Privacy as a Collective Phenomenon*, COSN'14, October 1–2, 2014, Dublin, Ireland, (consulté le 29 avril 2016), <http://arxiv.org/pdf/1409.6197v1.pdf>
[52] D'ailleurs certaines applications de messageries ne s'en cachent pas et vous proposent par exemple de *mettre à jour* les coordonnées d'un contact, montrant bien qu'elles sont capables de lire et de « comprendre » le contenu d'un message, d'y détecter la présence d'information de contact, et de les rattacher à un contact présent dans le carnet d'adresse. C'est notamment le cas chez Apple avec iOS 9 et OS X El Capitan, voir par exemple <http://www.idownloadblog.com/2015/10/22/how-to-disable-mail-contact-suggestions-iphone-mac/>.
[53] Pour une synthèse très pédagogique, voir la courte vidéo réalisée par Le Monde, *Comment la NSA vous surveille (expliqué en patates)*, sur <http://www.lemonde.fr/technologies/video/2013/10/21/comment-la-nsa-vous-surveille-explique-en-patates_3499887_651865.html>
[54] Le sujet est au cœur de l'actualité récente, avec l'affaire du décryptage de l'iPhone par le FBI ou l'activation du cryptage de certaines applications.
[55] Le texte de la loi est disponible sur <https://www.admin.ch/opc/fr/federal-gazette/2015/6597.pdf>. Voir aussi <http://www.vbs.admin.ch/internet/vbs/fr/home/themen/ndb/uebersicht.html>
[56] Le texte de la loi est disponible sur <https://www.admin.ch/opc/fr/federal-gazette/2016/1821.pdf>. Voir aussi les sites du comité référendaire <https://www.lscpt.ch/> et <https://stopbuepf.ch/>.
[57] Voir <https://panamapapers.icij.org/>
[58] Voir <https://www.icij.org/project/luxembourg-leaks>
[59] Pour mémoire, à l'été 2015, un pirate a mis la main sur plus de 30 millions de comptes d'un célèbre site de rencontres extra-conjugales. Les données ont été publiées sur le réseau. Voir à ce sujet <http://media.ashleymadison.com/> et <https://www.theguardian.com/technology/2016/feb/28/what-happened-after-ashley-madison-was-hacked>.
[60] En avril 2016, les données d'état-civil de près de 50 millions de citoyens turcs sont piratées <http://www.reuters.com/article/us-turkey-cyber-idUSKCN0X31ZK>.
[61] En avril 2016, les données personnelles de plus de 90 millions d'électeurs mexicains ont fuité <http://www.databreaches.net/personal-info-of-93-4-million-mexicans-exposed-on-amazon/>. Et ce ne sont que des exemples récents parmi une longue liste de cas.
[62] Voir *The Personal Data Notification & Protection Act* proposé par le président Obama en janvier 2015 <https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/updated-data-breach-notification.pdf >, et une analyse sur <https://iapp.org/resources/proposed-personal-data-notification-and-protection-act/ >
[63] Voir la *Charte des droits fondamentaux de l'Union Européenne*, < http://www.europarl.europa.eu/charter/pdf/text_fr.pdf>.
[64] Source : Wikipedia, <https://fr.wikipedia.org/wiki/Panoptique>.
[65] Lire et écouter < http://www.franceculture.fr/philosophie/la-societe-de-surveillance-de-foucault>.
[66] Voir < http://www.numerama.com/politique/142821-la-condamnation-de-la-visite-de-sites-terroristes-adoptee-au-senat.html> et l'article en question <http://www.senat.fr/amendements/textes/2015-2016/336.html#AMELI_SUB_4_1453915629906_4940>
[67] Voir < https://www.facebook.com/communitystandards>et <http://newsroom.fb.com/news/2015/03/explaining-our-community-standards-and-approach-to-government-requests/>
[68] Voir < https://support.google.com/websearch/answer/510?source=gsearch&hl=fr>
[69] Pour exemple, dans la traque d'Oussama ben Laden, « La résidence fortifiée n'est pas reliée au téléphone ni à Internet. Elle présente des mesures de sécurité hors-normes, y compris les murs de quatre à six mètres qui l'entourent. Cette résidence, en concluent les spécialistes du renseignement des États-Unis, « a été spécialement construite pour cacher quelqu'un d'importance » » (Wikipedia, <https://fr.wikipedia.org/wiki/Mort_d%27Oussama_ben_Laden>).
[70] Voir le tout récent cas lié au procès d'Antoine Deltour dans le cadre du procès LuxLeaks, où l'enquête s'appuie sur les publications de l'accusé sur Facebook (cité dans <http://www.lemonde.fr/international/article/2016/04/27/au-proces-luxleaks-le-lanceur-d-alerte-accuse-d-anticapitalisme_4909822_3210.html>) ou celui de la Charte du Recrutement Responsable proposée par le Syntec en France, dans laquelle à l'article 6 on peut lire « Le conseil en recrutement […] s'interdit d'utiliser des informations d'ordre personnel concernant les candidats collectées sur les réseaux sociaux. ». <http://www.syntec-recrutement.org/fichiers/20130809104649_SCR_Charte_du_Recrutement_Responsable_06_2013_modif_integrees.pdf>
[71] Dans le cadre du cycle de conférences autour d'un thème commun consacré aux multinationales : « est-ce qu'au nom du profit, tout est permis ? » <http://www.autrementcava.ch/>
[72] Voir <https://framasoft.org/>
[73] Voir <https://degooglisons-internet.org/>
[74] Voir <http://www.lemonde.fr/les-decodeurs/article/2016/04/19/ce-qu-il-faut-savoir-de-la-directive-sur-le-secret-des-affaires_4904548_4355770.html> et le texte de la directive <http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:52013PC0813&from=FR>