Big Data et anonymisation

Publié le 5 octobre 2017 – Mis à jour le 26 janvier 2018

Le recueil et l'anonymisation des données personnelles constituent depuis l'apparition de l'informatique des années 70 un sujet de préoccupation.

La notion de nom, de personnalisation en elle-même, constitue une valeur extrêmement forte pour l'individu : citons en vrac la célèbre malédiction envers Olrik du Grand Prêtre égyptien « *que ton nom ne soit plus*» du *Mystère de la Grande Pyramide* (Blake et Mortimer), les vols d'identité innombrables dans les romans populaires (Fantômas, Arsène Lupin...) et au cinéma. Citons par exemple les histoires de vols d'empreintes digitales et d'iris des yeux de la personne dont on veut dérober l'identité dans *Minority Report* de Steven Spielberg. N'oublions jamais les hideuses expériences des camps de la mort nazis du siècle dernier, où les déportés n'étaient plus qu'un numéro tatoué sur leur bras.

L'informatique c'est les *data*, l'informatique c'est les fichiers, l'informatique c'est tout de suite la relation aux personnes, clients d'une entreprise le plus souvent. Une personne dans un fichier a une certaine valeur, liée aux relations commerciales qu'une entreprise peut entretenir avec elle. Dès les années 70, Tore Dalenius travaille sur les variables présentes dans les fichiers informatiques (ex : nom, adresse, âge), et propose en 1986 la notion de quasi-identifiant, sorte de clef regroupant dans un ordre précis plusieurs variables. Il montre déjà que certaines combinaisons de variables permettent de ré-identifier de manière unique (« *singletons*») certains individus des fichiers. Une autre étape est franchie avec Latanya Sweeney [1] qui dévoile la maladie du gouverneur du Massachusetts en croisant deux fichiers disponibles aisément, l'un « anonymisé », l'autre non : un quasi-identifiant constitué des variables (zipcode, date de naissance, sexe) a suffi. C'est la curée : Yves-Alexandre de Montjoye du *Massachasetts Institute of Technology (MIT)* démontre en 2013, avec plusieurs co-auteurs, que quatre positions spatiales et temporelles d'un smartphone suffisent à identifier à 95 % près 1,5 millions d'Américains. Puis, en 2015, des résultats similaires sont trouvés pour leurs tickets de caisse dans les super-marchés, etc.

Parallèlement, des parades sont trouvées : les concepts de floutage (on rajoute du bruit statistique aux variables, on utilise des regroupements) permettent d'abaisser le taux de ré-identification des quasi-identifiants. Une métrologie des méthodes et des risques de ré-identification est validée en 2010 dans un important colloque à Washington sur les données de santé, qui précédera la prise de conscience française et européenne au tournant de l'année 2014, elle-même simultanée du *Privacy Report*demandé par le Président Obama.

Les procès en utilisation inappropriée des données privées des fichiers, émanant souvent de « *class-action suits*» de clients excédés, mais aussi des institutions européennes, se multiplient. Ils constituent aujourd'hui probablement la meilleure régulation – financiarisation et judiciarisation des offenses – de cet équilibre délicat entre le respect de la vie privée des individus et l'utilité économique et sociétale de l'utilisation des données qui proviennent des fichiers d'aujourd'hui que sont les Big Data. En bon adepte du principe de Le Chatelier (1884 : la Nature tend à s'opposer aux modifications d'un équilibre qu'un système nouveau entend lui apporter), je pense que le mythe du Big Brother du roman *1984*ne sera jamais réalité. Comme le disait le mathématicien du film *Jurassic Park* : « *la Nature trouve toujours un chemin* ».

Par Michel Bera,
Professeur du Cnam,
chaire de Modélisation statistique du risque.

[1] Professeure et directrice du laboratoire *Data Privacy* de l'Université d'Harvard

| Informatique | Numérique | Sécurité

Écoutez sa leçon inaugurale

Leçon du 14 décembre 2010 "Modélisation statistique du risque : réalités de l'histoire, utopies pour le futur"

http://blog.cnam.fr/anciennes-rubriques/grand-angle/les-big-data/big-data-et-anonymisation-947632.kjsp?RH=1506929