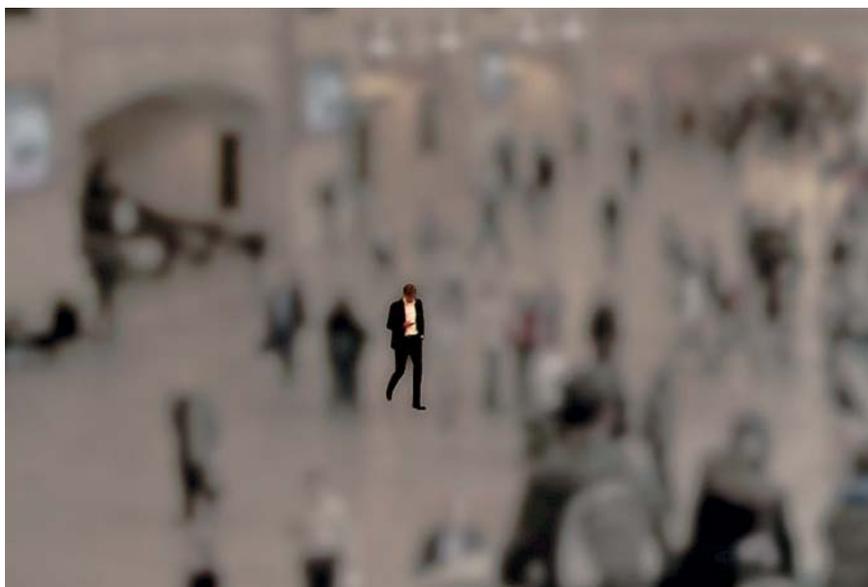


BIG DATA

So what?

Métadonnées, « pour ou contre » ?



La collecte et l'utilisation à grande échelle des métadonnées sont devenues non seulement possibles mais très bon marché.

Le 6 juin 2013, un nouveau mot est (ré)apparu dans la langue française : métadonnées ! Pas de nouvelle édition du *Petit Robert* à l'horizon, mais bien les révélations inédites d'un ancien consultant de la NSA, l'agence américaine de renseignement. En quelques jours, le mot a fait la une de tous les grands quotidiens.

Métadonnées, littéralement « données à propos des données ». Bien que le terme ne soit pas nouveau – il est utilisé dans les systèmes de classification des bibliothèques – l'avènement du numérique lui donne un nouveau sens et surtout une nouvelle portée. Les métadonnées modernes sont les traces numériques que nous laissons tous derrière nous, en permanence. Lorsque nous téléphonons, lorsque nous naviguons sur Internet, lorsque nous payons avec notre carte bancaire. Les métadonnées de nos téléphones portables ressemblent à une facture très détaillée : appels ou textos reçus, dates et heures, antennes GSM auxquelles nous sommes connectés. Ces métadonnées comportementales sont, avec les données textuelles, un des deux grands types de « big data », ces très grands ensembles de données dont la collecte et l'utilisation à grande échelle sont récemment devenues non seulement possibles mais (très) bon marché.

Est-ce parce que le grand public a appris leur existence par les agences de renseignement aux États-Unis ou en France ? Parce qu'elles sont collectées de manière passive ? Ou encore parce qu'elles sont plus difficiles à appréhender que leurs équivalents textuels ? Il est en tout cas certain que ces métadonnées inquiètent.

Du positif...

Mais d'abord, revenons sur le côté positif de ces métadonnées. Elles facilitent notre vie quotidienne : quel est le meilleur chemin pour éviter les bouchons ? Quel sera mon film préféré ? Quelle page web répond exactement à ma question ?

Les métadonnées sont également cruciales pour l'ingénieur : gérer et améliorer le réseau téléphonique, lutter contre la fraude bancaire, optimiser un réseau de distribution.

Enfin, pour la recherche scientifique, ces métadonnées sont une révolution. Un récent article dans la revue *Science* compare leur impact scientifique à l'invention du microscope. En épidémiologie, les données de mobilité sont utilisées pour étudier la propagation d'un virus comme la malaria. En économie du développement, les chercheurs travaillent à l'utilisation des données téléphoniques pour comprendre et mieux combattre la pauvreté. Autre exemple : les métadonnées font avancer la recherche en management et en sciences sociales. Comment la productivité d'un employé est-elle influencée par ses liens sociaux les plus forts, quel découpage rationnel pour un territoire comme la France, comment la diversité de notre réseau social est-elle liée à notre pouvoir d'achat, ou encore comment nos connaissances et amis influent-ils sur nos opinions ?

Que du positif ? Même si nous sommes bien loin des références orwelliennes ou kafkaïennes, l'utilisation commerciale et gouvernementale à grande échelle de métadonnées comportementales soulève trois grandes questions.

Des interrogations/questions

Tout d'abord, celle de l'anonymat : nos traces numériques contiennent beaucoup d'informations et sont fondamentalement personnelles et privées. C'est pourquoi les chercheurs comme les entreprises n'utilisent que des bases de données anonymisées,



Il est possible de prédire la personnalité d'une personne en observant la manière dont celle-ci utilise son téléphone.

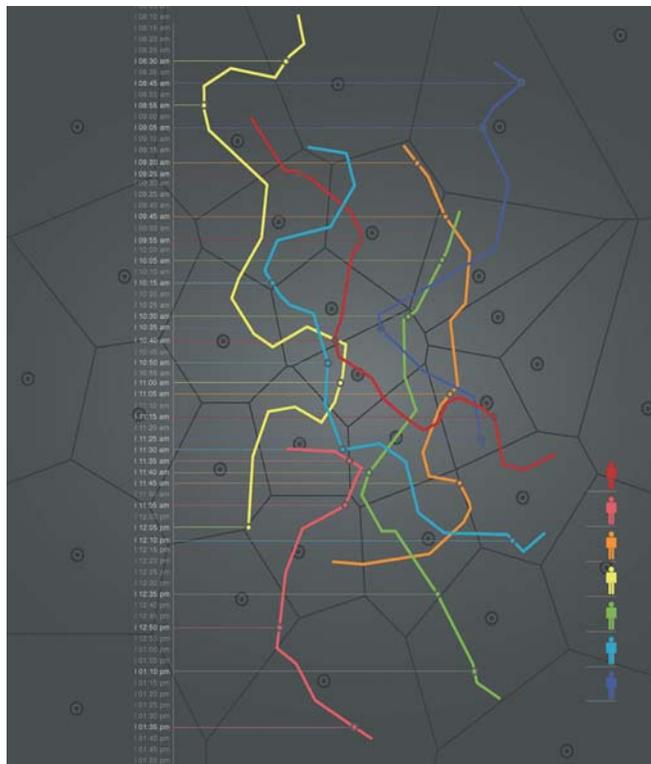
desquelles on a retiré les identifiants d'un utilisateur : son nom, son numéro de téléphone, son adresse... Cependant, dans le cas des métadonnées, cela n'est absolument pas suffisant. Un récent article en collaboration avec l'université de Louvain montre par exemple que notre manière de nous déplacer est très régulière, unique, et comparable à des empreintes digitales. Il suffit en effet de connaître quatre points, quatre endroits et temps approximatifs, où un utilisateur était pour le retrouver dans une base de données pourtant apparemment anonyme de 1,5 million de personnes. Les métadonnées sont riches, leurs usages multiples et il est très peu probable qu'il soit jamais possible de les anonymiser. Il est donc temps d'oublier, légalement et techniquement, la notion d'anonymat au profit d'une quantification du risque de ré-identification.

Deuxième interrogation : les révélations indirectes sur l'individu. Les métadonnées téléphoniques contiennent beaucoup plus d'informations qu'il n'y paraît. Une étude en collaboration avec des chercheurs de l'ENS de Lyon a montré qu'il est possible de prédire la personnalité d'une personne en observant la manière dont celle-ci utilise son téléphone. En calculant un certain nombre d'indicateurs à partir des métadonnées téléphoniques, la durée moyenne qu'un utilisateur prend pour répondre à un texto, la distance moyenne qu'il parcourt par jour ou encore la diversité de ses contacts, des algorithmes de *machine learning* peuvent prédire le score d'un utilisateur dans chacun des cinq grands facteurs de personnalité : l'extraversion, le neuroticisme, l'ouverture à l'expérience, la conscienciosité ou encore l'agréabilité. La vraie question à se poser pour les métadonnées n'est donc pas ce qu'elles révèlent directement mais bien ce qu'un algorithme pourrait, raisonnablement, révéler sur une personne en les utilisant.

Enfin, troisième question, la propriété et l'accès aux métadonnées. Bien qu'utilisées à

bon escient, leur collecte et leur utilisation sont malheureusement souvent faites de manière peu transparente. Ce manque de transparence nourrit les fantasmes. L'utilisateur, celui qui génère les données, doit au minimum y avoir accès. Seul l'accès aux métadonnées brutes permet de comprendre ce qu'elles contiennent et l'usage qui peut en être fait, directement ou indirectement. De même seul cet accès aux données brutes permet à l'utilisateur de les utiliser pleinement.

Il ne s'agit donc pas d'être « pour ou contre » les métadonnées mais de les expliquer, de se poser les bonnes questions et de choisir les réponses que nous voulons y apporter, en tant qu'ingénieurs ou que simples citoyens.



Notre manière de nous déplacer est régulière, unique et comparable à des empreintes digitales.



Yves-Alexandre de Montjoye (08)

@yvesalexandre est chercheur en mathématiques appliquées au MIT Media Lab. Il développe des méthodes stochastiques pour l'analyse de métadonnées comportementales :

données de mobilité, transactions financières, communications dans les réseaux sociaux. Ses recherches ont reçu une couverture médiatique dans *BBC News*, *CNN*, *The New York Times*, *Wall Street Journal*, *Foreign Policy*, *Le Monde*, *Der Spiegel*, dans les rapports du World Economic Forum et des Nations unies.

Avant de rejoindre le MIT, Yves-Alexandre était chercheur au Santa Fe Institute (Nouveau-Mexique).

Il est titulaire d'un master en mathématiques appliquées de l'université de Louvain et d'un master en ingénierie mathématique de la KU Leuven (Belgique).