TELECOM
REVUE DE L'ASSOCIATION TELECOM PARISTECH ALUMNI

TELECOM ParisTech
ALUMNI

N°169 - JUILLET 2013

BIG DATA
NOUVEAUX DÉFIS

Editorial

Fleur Pellerin, Ministre Déléguée chargée des PME,
de l'Innovation et de l'Économie numérique

Cher lecteur,

Je remercie Télécom ParisTech de me donner l'occasion, à travers cet éditorial, de partager avec vous mes convictions sur ce sujet stratégique des Big data.

Derrière le « buzz », je suis convaincue que nous avons devant nous une nouvelle révolution du numérique, avec des implications sociétales et économiques.

Nous vivons au quotidien un « déluge de données », que l'explosion des terminaux mobiles et l'émergence attendue de l'Internet des Objets vont amplifier. Mais les enjeux sont devant nous : comment créer de la valeur (une valeur d'usage, une valeur économique) à partir du traitement de ces grandes masses de données, tout en respectant des contraintes de confidentialité ?

Des standards de fait ont déjà émergé et permettent de « passer à l'échelle », je pense à Hadoop. Sur cette base, nos chercheurs, nos entrepreneurs, nos startups ont une vraie carte à jouer pour faire de la France la référence pour cette nouvelle économie.

Pourquoi suis-je si optimiste ?

Plus que dans d'autres domaines de l'économie numérique, la barrière technologique reste un élément important pour les Big data, par exemple avec la nécessité de disposer d'algorithmes performants de *crawling* ou de *text mining* (en bon français). Avec nos compétences en mathématiques, en statistiques ou en informatique, nous avons tous les atouts pour faire de cette barrière un avantage concurrentiel.

Un autre élément majeur est la capacité à disposer de données, le carburant de cette nouvelle économie. Parmi les sources de données, il y a bien sûr les données ouvertes (« open data »), largement ancrées dans nos territoires, donc à activités induites moins délocalisables. Avec le travail d'Etalab, sous la direction renouvelée d'Henri Verdier, nous avons un autre atout dans notre manche.

Enfin, les secteurs économiques et industriels « classiques » sont en train de réaliser que leurs modèles vont changer avec les Big data. C'est toute ma responsabilité que de veiller à ce qu'aucun domaine ne rate le train en marche.

La politique que je mène pour soutenir l'économie numérique est à l'écoute des acteurs que vous êtes. Du nouvel appel à projet du FSN dédié aux Big data (à partir de septembre) aux quartiers numériques que nous sommes en train de mettre en place pour dynamiser les écosystèmes numériques de croissance, en passant par ma politique en faveur de l'innovation, nous construisons les éléments dont vous avez besoin pour réaliser vos ambitions. Et les Big data sont un axe central de cette politique.

Ce numéro montre tout le dynamisme de Télécom ParisTech, plus largement de la communauté française du numérique, je pense par exemple à des entreprises comme Talend ou Exalead, à des entrepreneurs comme Tariq Krim, et à bien d'autres encore !

Bonne lecture !

BIG DATA : NOUVEAUX DÉFIS

Notre réseau

LE PROCHAIN NUMÉRO SERA CONSACRÉ AUX TRANSPORTS INTELLIGENTS.

TÉLÉCOM n°169 - Juillet 2013
est édité par l'Association Télécom ParisTech alumni.
Dépôt légal à parution.

Directeur de la publication : Dominique Jean (73)
Secrétaire de rédaction : Amélie Pageard
Numéro préparé par : Gérard Cambillau (73) et Louis-Aimé de Fouquières (82)
Comité de rédaction :
Marylin Arndt (81), Céline Beillouin (2011), Skander Benattia (2009), Christine Chardon (95), Michel Cochet (73), Jean-Pierre Dardayrol (77), Stéphane Debarbieux (84), Hélène Delahousse (93), David Fayon (93), Ayoub Figuigui (2011), Grégoire Galievsky (2000), Philippe Hilsz (80), Michel Mairal (97), Mejdi Mrad (2011), Pascal Verveur (88)

Direction artistique et réalisation : Valérie Mounier
Dessinateur : Gédéon
Banque d'image : Thinkstock
Les illustrations des articles sont fournies par les auteurs, sous leur responsabilité concernant les droits de reproduction. Les idées exprimées dans cette revue engagent la seule responsabilité de leurs auteurs. Reproduction autorisée avec mention d'origine après accord de la publication.

Rédaction & Abonnements :
46 rue Barrault
75634 Paris Cedex 13
Tél. 01 45 81 74 77
Courriel : revue@telecom-paristech.org
Site : www.telecom-paristech.org

Régie publicitaire : FFE
15 rue des Sablons - 75116 Paris
Directeur de la publicité :
Patrick Sarfati - Tél. 01 53 36 20 40
Chef de publicité : David Sellam
Tél. 01 48 05 26 65 - Email : david.sellam@ffe.fr
Assistante de fabrication :
Aurélie Vuillemin - Tél. 01 53 36 20 35

Imprimé en france

CCP n° 29854 N Paris
Abonnements annuels 2011 : 53 € TTC
Prix au numéro : 21 € TTC
ISSN 0040-2478

Enjeux économiques du Big data

par Patrick WAELBROECK

Les big data adoptent une approche macro-économique d'un problème en cherchant des structures statistiques basées sur des données massives et globales issues de l'Internet. Les big data permettront aux entreprises et aux gouvernements de mieux analyser les tendances de marché et les changements d'opinions. Cet article explore les enjeux économiques du big data liés à l'utilisation de données personnelles. Or ces données personnelles sont liées à la manière dont une personne se rend visible sur Internet et donc à son identité. Pour comprendre les enjeux économiques des big data, il est important de faire un détour sur la notion d'identités numériques. En effet, la manière dont les gens gèrent leurs identités peut avoir un impact sur des entreprises telles que Google ou Facebook qui exploitent les données personnelles à travers les big data.

Big data et identités numériques

Les big data sont étroitement liés à la manière dont les internautes construisent et gèrent leurs identités numériques, car même si beaucoup d'informations ne sont pas *a priori* personnelles, elles peuvent le devenir a posteriori grâce au data mining (exploitation de données pour détecter des régularités) et aux big data (mise en relation de données publiques, privées, d'entreprises) qui permettent de recouper des informations.

Deux visions des identités numériques s'opposent dans la littérature sociologique. La première pré-suppose que les gens utilisent les outils internet pour construire leur identité, ils utilisent donc des avatars et adoptent des comportements différents de leur personnalité réelle. Les internautes utiliseraient des personnages différents de leur véritable identité. Ces identités sont fonction de la technologie, du contexte social (par exemple issu des normes sociales établies dans les communautés en ligne) et du contexte culturel. Les outils de communication électronique représenteraient une forme de laboratoire d'identités. La seconde vision considère, au contraire, que les internautes se représentent en ligne comme ils sont dans la vie réelle, et qu'ils gèrent de manière active les informations qu'ils divulguent aux autres membres d'une ou plusieurs communautés. L'utilisateur est alors incité à brouiller son identité : anonymat, pseudo, rétention d'informations, déclarations mensongères. Ces deux approches décrites ne sont pas nécessairement contradictoires, car les internautes passent de l'une à l'autre, dans un processus dynamique impliquant la construction et la projection de soi.

Enjeux des big data

L'un des principaux enjeux économiques des big data est de mieux cibler, analyser et prévoir des comportements et des tendances. La manière dont les internautes gèrent leurs traces et contributions ainsi que leurs identités numériques, a donc un impact de premier ordre sur la fiabilité et la qualité des outils statistiques issus des big data.

Discrimination par les prix

Lorsque les entreprises ont plus d'informations sur leurs clients, elles peuvent pratiquer la discrimination par les prix, à savoir tarifer un même produit ou service à des prix nets différents. Le prix net s'entend net de frais de livraison. Ainsi, pour les produits numériques, la forme de discrimination la plus répandue est celle où les entreprises identifient plusieurs groupes de consommateurs et leur proposent des versions différentes d'un même produit ou service. Par exemple, un fabricant de logiciel propose plusieurs versions de son produit avec des fonctionnalités différentes : une version professionnelle complète et une version plus simple (ou pour étudiants) d'où certaines fonctionnalités sont absentes. La collecte d'informations permet donc de personnaliser les offres à des coûts souvent très faibles pour les entreprises fournissant des biens numériques facilement modulables. Quel est l'effet de cette pratique sur les profits des entreprises et la satisfaction des consommateurs ? De manière surprenante, le surplus total ne diminue pas nécessairement lorsqu'une entreprise en monopole pratique la discrimination par les prix, en particulier lorsque les ventes augmentent. Ainsi, une augmentation de la collecte d'informations n'est pas nécessairement mauvaise pour les consommateurs qui se

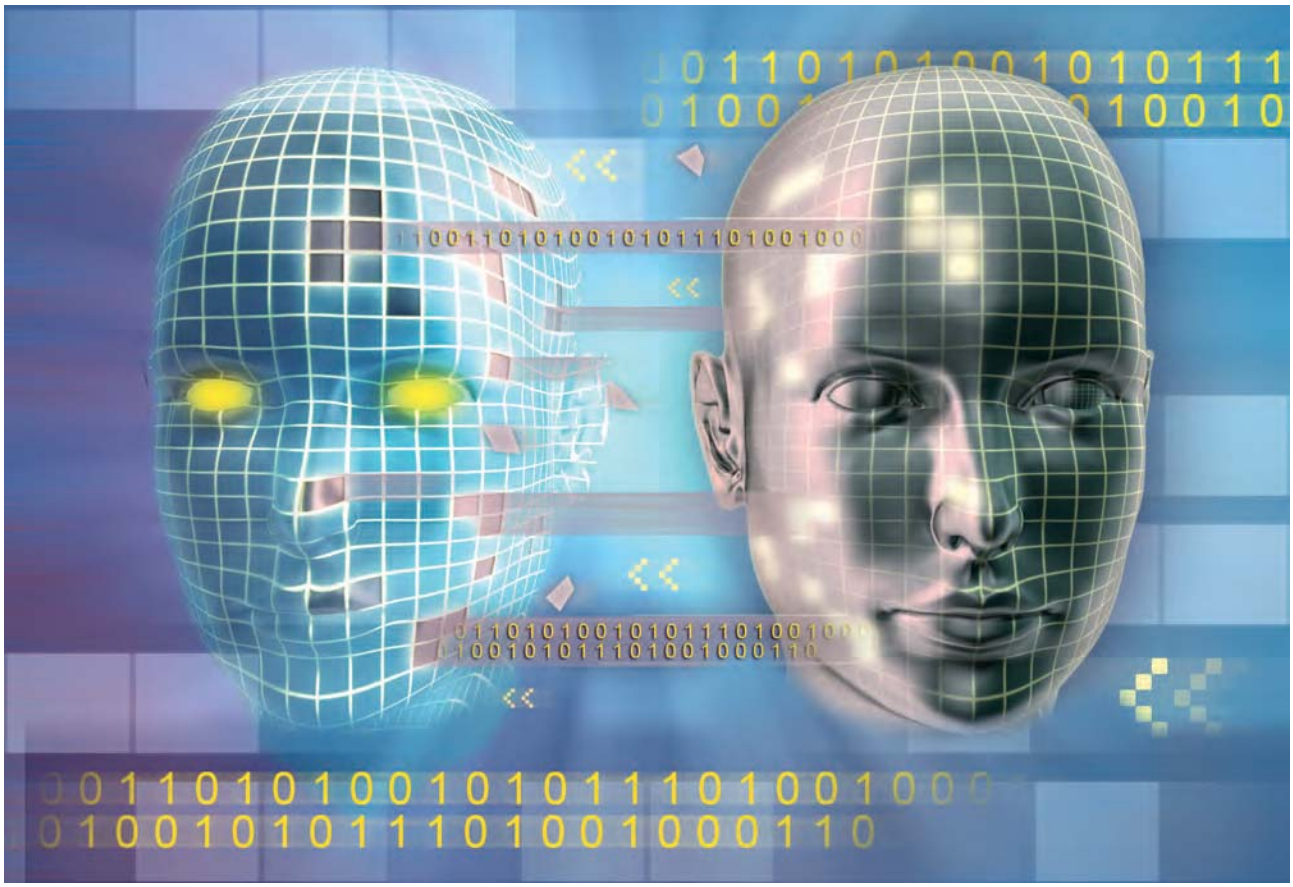
voient proposer des offres ciblées, même s'il existe un risque de clientélisme (un vendeur propose des prix à ses clients en fonction de critères non objectifs) et de pratiques anti-concurrentielles liées au bundling, aux contrats exclusifs de distribution et aux ventes liées.

Marchés à plusieurs versants, valorisation des données et le gratuit

De nombreuses plateformes de réseaux sociaux, d'actualités et de recherche d'informations peuvent être caractérisées par ce que la littérature économique a appelé des marchés à deux versants. Ceux-ci sont caractérisés par des externalités croisées entre différents groupes d'agents. Par exemple, sur un site de recherche d'informations tels que Google.com, les internautes accèdent à des contenus gratuitement, et le site se finance par la publicité. Le site met donc en relation des consommateurs potentiels et des annonceurs. La satisfaction d'un consommateur potentiel lorsqu'il effectue une recherche sur un site dépend du nombre de publicités ciblées qu'il trouvera pertinentes par rapport à ses recherches. De même, pour un annonceur, le bénéfice d'une publicité dépend du nombre d'utilisateurs du moteur de recherche. Il y a donc une externalité croisée positive entre les internautes et les annonceurs. La dynamique des marchés à deux versants où internautes et publicitaires/fournisseurs de contenu se rencontrent implique que les plateformes qui parviennent à gagner un petit avantage comparatif initial puissent se retrouver dans une boucle positive, alors que les autres subissent une spirale négative. Ainsi, les marchés des moteurs de recherche et des sites d'actualités où l'internaute est à même de découvrir de nouveaux produits risquent d'être très concentrés, avec l'émergence d'un intermédiaire dominant. La théorie des marchés à plusieurs versants montre que l'intermédiaire a intérêt à subventionner le côté du marché qui est fondamental pour le développement de la plateforme. Ainsi les services de Google et de Facebook sont gratuits pour les utilisateurs, alors que les annonceurs paient pour accéder à leurs audiences. Cette pratique explique également pourquoi les utilisateurs ont du mal à donner une valeur à leurs données personnelles, puisqu'ils ne paient pas pour le service, alors que, prises dans leur ensemble, ces données personnelles sont valorisées à des centaines de milliards de dollars par les investisseurs sur les marchés financiers.

Frontière entre propriété intellectuelle et protection des données personnelles

Un internaute qui met en ligne des données, comme par exemple un commentaire laissé sur un forum Internet, est théoriquement protégé par le droit d'auteur/copyright. Cependant, l'utilisateur renonce à ce droit par la signature d'un contrat électronique qui transfère la propriété intellectuelle au propriétaire du site. Par exemple, il est difficile de contrôler des informations personnelles sur une photo référencée par des moteurs de recherche. Lorsque plusieurs tags sont associés à une photo, la question de savoir qui possède le contrôle final sur les tags (règle de décision) est fondamentale pour que des internautes continuent de partager des photos sur des sites tels que Flickr ou Facebook. Ceci soulève l'importante question du marché des données personnelles. L'utilisateur pourrait devenir offreur/producteur de données personnelles, et les annonceurs deviendraient demandeurs de ces données. Même si les données personnelles semblent très hétérogènes, un marché pourrait se créer où s'établirait un prix de transaction payé par l'annonceur directement à l'utilisateur, et non plus en passant par un intermédiaire comme Google



et Facebook. Les données deviendraient alors « autonomes ». Un tel marché s'autorégulerait-il ou au contraire nécessiterait-il une intervention gouvernementale ?

Big data, empowerment et relation au pouvoir

Les consommateurs seront mieux informés aussi grâce au pouvoir décisionnel que leur procurent les informations issues de ces données massives. L'open data va permettre au citoyen de mieux évaluer les actions politiques. Il y gagnera en autonomie et le processus démocratique pourrait en sortir renforcé.

Les limites des big data

Cet article soulève d'importantes limites à l'utilisation des données issues de collectes d'identités numériques et aux «big data» de manière générale : en voici une liste non exhaustive.

Biais de sélection et anticipations rationnelles

Les individus peuvent se présenter sous de fausses identités ou des identités partielles ou « espérées » ou s'identifier au minimum pour rester le plus anonyme possible. De plus, ils peuvent contribuer beaucoup ou peu à des projets collectifs (open source, wikipedia, forums et autre communautés de savoir). Dès lors, certaines caractéristiques et opinions seront sur-représentées, créant ainsi un biais de sélection dans l'analyse statistique de ces données. Le biais de sélection peut également provenir d'un marché de la réputation où des agents proposent de bien noter contre rémunération. On peut également acheter des « followers » sur Twitter ou des « like » sur Facebook. De manière plus générale, les modèles de prévision et d'anticipation du comportement des individus ne sont opérationnels que si ces derniers sont relativement passifs. Au contraire, lorsque les individus anticipent eux-mêmes les règles qu'on leur applique, ils peuvent manipuler leurs données pour rendre l'algorithme de prévision inefficace. Cette critique des modèles de prévision basée sur les anticipations rationnelles dans un contexte macro-économique a été énoncée par le prix Nobel Robert Lucas.

Défiance

Dans un contexte de défiance envers les réseaux socio-numériques et les entreprises qui les ciblent, les internautes auront recours à des outils de plus en plus sophistiqués pour garantir le respect de leurs données personnelles. L'utilisation de réseaux privés (VPN) et d'outils cryptographiques déclenchera en réponse une réaction proportionnelle des annonceurs qui se tourneront vers des outils de ciblage de plus en plus intrusifs, menant à une guerre de protection. Dans un scénario extrême, on peut envisager que les internautes se déconnectent et n'échangent plus leurs données qu'à travers des disques durs portables.

Big data et small data

Selon Alan Mitchell (2012), l'un des principaux enjeux de productivité portera sur la logistique de l'information, ou comment garantir que l'information pertinente arrive au bon endroit au bon moment. Il s'agit d'une approche micro-économique. L'efficacité des big data dépendra de celle de la combinaison entre les approches micro- et macro-économiques d'un problème.

Surveillance, sous-veillance et les limites de l'empowerment

L'historique des données de navigation et des traces volontairement ou involontairement laissées sur Internet facilitent la surveillance par une autorité centrale qui pourrait ainsi renforcer son pouvoir politique. Ce phénomène est déjà largement connu sous le terme « Big Brother ». Un autre phénomène plus insidieux appelé par Castells (2001) « Little Sisters » renvoie à un phénomène de « sous-veillance » où les actions, les traces et les posts d'un internaute sont sans cesse suivis par une multitude de « petites soeurs » qui rendent la frontière entre le monde physique et le monde en ligne de plus en plus poreuse. Cette double évolution aura des conséquences sur l'évolution de l'autonomie des individus, et sur la manière dont ils s'éloignent ou non d'une norme sociale imposée par un consensus.

Gouvernementalité algorithmique

Dans un article du Wall Street Journal ("Meet the new boss: Big data", 20 septembre 2012), l'auteur montre comment les logiciels utilisant les données massives peuvent fournir des recommandations à un employeur d'un centre d'appel, non pas basées sur l'expérience professionnnelle passée pour embaucher de nouvelles recrues, mais sur des tests de personnalité (attitude envers l'alcool, distance du domicile au travail, personnalité inquisitrice, ...) qui garantiraient qu'un nouvel employé reste au moins six mois. Antoinette Rouvroy (2012) soulève les risques liés à une utilisation systématique du data-mining et d'algorithmes statistiques utilisant des données massives. La création de catégories et la rationalité algorithmique éloignent les personnes de la véritable communication et remettent en question les notions de sens critique, de justice et de protection des minorités. Comment réagir lorsqu'un algorithme vous a assigné une catégorie qui ne reflète évidemment pas la complexité de votre personnalité ?

Bibliographie

Castells M. (2001), The Internet Galaxy: Reflections on the Internet, Business and Society. Oxford: Oxford University Press

Mitchell A. (2012), "Big data, big dead end", http://www.ctrl-shift.co.uk/index.php/news/2012/01/17/big-data-big-dead-end/

Rouvroy A. (2012), "Face à la gouvernementalité algorithmique, repenser le sujet de droit comme puissance", Document de travail

L'AUTEUR

Patrick WAELBROECK est titulaire d'une thèse en économie de l'Université de Paris 1 Panthéon-Sorbonne. Il a fait une partie de ses études à l'Université de Yale aux Etats-Unis pour lesquelles il a obtenu une bourse Fulbright. Ses travaux portent sur l'économie de l'innovation, l'économie de la propriété intellectuelle, l'économie de l'Internet et des données personnelles. Patrick Waelbroeck est membre du comité éditorial du Journal of Cultural Economics. Il a publié de nombreux travaux très largement cités sur le sujet du piratage dans les industries culturelles, qui ont influencé le débat public en France, en Europe et aux Etats-Unis. Il est président de l'association European Policy for Intellectual Property (2013-2014). Patrick Waelbroeck est également membre fondateur de la chaire « Valeurs et politiques des informations personnelles » qui aborde les problèmes des données personnelles et du Big data sous différents angles : juridique, économique, technique et philosophique.