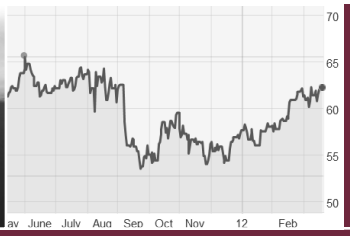
CIRANO

Allier savoir et décision

Centre interuniversitaire de recherche en analyse des organisations

RAPPORTBOURGOGNE



2014RB-01 > Juin 2014

Un état des lieux sur les données massives

Thierry Warin (HEC Montréal et CIRANO)

Nathalie de Marcellis-Warin (École Polytechnique de Montréal et CIRANO)

Avec les contributions de :

Antoine Troadec (CIRANO)

William Sanger (École Polytechnique de Montréal et CIRANO)

Bertrand Nembot (École Polytechnique de Montréal et CIRANO)

Les Rapports bourgogne sont des documents de synthèse portant sur des questions d'intérêt général produits par des chercheurs du CIRANO. Ils contribuent à alimenter la réflexion et le débat public sur des questions d'actualité.

Le CIRANO est un centre de recherche multidisciplinaire qui a pour mission l'accélération du transfert des savoirs entre le monde de la recherche et celui de la pratique.

ISSN 1701-9990

Partenaire financier

Enseignement supérieur, Recherche, Science et Technologie
Québec

Sous la direction de

Thierry Warin

Professeur, HEC Montréal et Vice-président Stratégie et économie internationales, CIRANO.

Thierry Warin est l'auteur de plus de 30 publications académiques et de 9 livres.

Avant de rejoindre HEC Montréal, Thierry Warin a enseigné à l'École polytechnique de Montréal et a eu différentes positions dans plusieurs institutions académiques (Middlebury College, ESSEC Business School, HEC Paris, La Sorbonne, UIBE Beijing, Sun-Yat-Sen University Canton).

Thierry Warin s'intéresse aux thématiques de finance internationale, macroéconométrie et économie politique internationale. Alumnus du Centre Minda de Gunzburg sur les études européennes à l'université de Harvard, Thierry a obtenu son Ph.D en économie monétaire et financière à l'ESSEC (France). warint@cirano.qc.ca

Nathalie de Marcellis-Warin

Professeure agrégée à l'École Polytechnique de Montréal et vice-présidente au CIRANO des groupes Risques et Développement durable.

Nathalie de Marcellis-Warin, Ph. D., est diplômée de l'École Normale Supérieure (Cachan, France) en sciences de la gestion.

Ses intérêts de recherche portent sur la gestion des risques et la théorie de la décision dans un contexte de risque et d'incertitude. Madame de Marcellis-Warin a publié de nombreux articles scientifiques et plus de 20 rapports de recherche pour des organismes gouvernementaux et d'autres organisations. Elle participe actuellement à des projets de recherche sur la perception des risques, la communication des risques et l'acceptabilité sociale des grands projets publics.
Nathalie.De_Marcellis@cirano.qc.ca

Avec les contributions de :

Antoine Troadec (CIRANO)

William Sanger (Polytechnique Montréal et CIRANO)

Bertrand Nembot (Polytechnique Montréal et CIRANO)

Les auteurs remercient Industrie Canada pour l'aide financière accordée à ce projet de recherche. Notez cependant que les opinions exprimées dans ce rapport ne sont pas nécessairement celles d'Industrie Canada ou du gouvernement du Canada.
Financial support from Industry Canada to conduct the research on which this report is based is gratefully acknowledged. Note however the views expressed in this report are not necessarily those of Industry Canada or of the Government of Canada.

Table des matières

Les données massives

Les données massives sont partout. Elles sont de plus en plus identifiées et collectées à des fins stratégiques pour les entreprises ou de mesure des risques pour les gouvernements. L'exploitation des données massives introduit des changements importants dans nos sociétés. Nous sommes aujourd'hui au début d'une véritable révolution industrielle. Avec de nouvelles opportunités apparaissent aussi de nouveaux enjeux et de nouveaux risques. Les données massives ont une citoyenneté juridique, politique, morale et économique. Ce rapport souligne que la territorialité des données est un enjeu important en ce début de XXIe siècle.

Ce rapport est un état des lieux au début de l'année 2014. Il ne se veut pas exhaustif et ne peut l'être. Les avancées technologiques vont tellement vite qu'il faudra continuer à contempler la révolution en marche. Les références bibliographiques retenues sont celles que l'on pouvait regrouper sous les thématiques choisies dans ce rapport. C'est une toute nouvelle littérature qui n'en est qu'à ses débuts et l'avenir est sans aucun doute prometteur. Les données massives sont une innovation radicale de procédé.

INTRODUCTION

De janvier à juin 2014, ce sont plus de 500 millions de messages qui se sont échangés quotidiennement sur Twitter entre 241 millions d'utilisateurs. Et ce n'est qu'une toute petite partie des données qui sont disponibles sur les réseaux sociaux et plus largement sur Internet.

Les données massives sont le résultat de la rencontre de trois éléments : Internet, les réseaux sociaux et les appareils intelligents (ordinateurs, téléphones, tablettes, etc.). L'Internet permet la transmission de l'information quelle que soit sa forme créée sur les appareils intelligents et partagée sur les réseaux sociaux. L'outil de création des données est l'appareil intelligent, le consommateur est l'utilisateur des réseaux sociaux et le vecteur de transmission est Internet.

Le terme anglais « Big Data » a été proposé par John Mashey, alors expert scientifique chez Silicon Graphics (Diebold, 2012). « Big Data » est souvent défini en utilisant l'acronyme VVV pour volume, vitesse et variété (Laney, 2001) : « Volume » se réfère à des quantités massives de données qui sont disponibles, « Vitesse » se réfère à la vitesse nécessaire pour traiter, analyser et utiliser les données, et « Variété » désigne les deux types de données disponibles : les données structurées (prix, dates, températures, poids, données boursières, etc.) et les données non structurées qui sont générées sur Internet (vidéos, images, données audio, textes, etc.).

Les données massives sont le résultat de la rencontre de trois éléments : Internet, les réseaux sociaux et les appareils intelligents.

Les données structurées

Les données structurées sont les données que l'on peut clairement codifier et identifier. Les données d'un tableur sont typiquement des données structurées. On peut comprendre leur signification en croisant les titres de la ligne et la colonne dans laquelle se trouvent les données. Ces données répondent à une codification qui permet de les classer et d'en tirer une information. Les systèmes d'analyse algorithmique ont depuis toujours été développés afin de traiter ce genre de données. L'ère du Big Data permet surtout un traitement de grande ampleur et en temps réel de ces données.

Les données non structurées

Comme leur nom l'indique, les données non-structurées ne répondent pas à une codification qui permet d'en tirer mécaniquement une information. Il n'y a pas d'autre moyen que de lire les gazouillis (terme français pour identifier les tweets) pour en extraire le sens. C'est ensuite en analysant le contenu des messages que l'on structure l'information. Les exemples de types de données non-structurées sont les fichiers textes, les fichiers audio ou vidéo et toute autre information issue d'un signal analogique. Les données non structurées représentent actuellement la grande majorité de l'information que l'on côtoie et surtout la partie la moins exploitée. En mettant en place des outils d'analyse de données massives efficaces, cette information devient exploitable.

D'autres définitions du « Big Data » sont proposées dans la littérature (Warin & Sanger, 2014), nous nous limiterons ici à définir les données massives comme ayant les caractéristiques suivantes :

1. elles sont soit structurées, soit non structurées;
2. elles sont disponibles en grande quantité, presque infinie;
3. elles sont disponibles en temps réel;
4. elles sont (le plus souvent) longitudinales.

Ce n'est pas la première fois que les sciences de l'information, les statisticiens ou les économètres ont affaire à des données en grande quantité. Par exemple, en finance, il est évident que les volumes d'information qui transigent sur les serveurs des institutions financières correspondent aux caractéristiques des données massives énoncées ci-dessus. En effet, les données financières sont bien identifiées (structurées), elles sont disponibles en quantité extrêmement importante, en temps réel et souvent pour des longues périodes. Toutefois, avec la « variété » des données, d'autres informations - non structurées - sont maintenant disponibles et circulent sur les réseaux sociaux. Par exemple, les gazouillis donnent des informations qualitatives sur ce que pensent les utilisateurs, leur analyse des données financières structurées, leurs réactions ou interprétations après la publication d'un rapport annuel d'une compagnie ou ses données financières, etc. Ces données non structurées, une fois qu'elles sont « restructurées », peuvent permettre de mieux expliquer certains phénomènes et réactions sur les marchés financiers. Elles sont donc une nouvelle source d'information massive.

Il est certain que l'informatisation ou la numérisation des informations concernant nos vies, nos perceptions et nos sentiments aura un impact sur nos sociétés. Cette nouvelle source de données un peu moins neutre – car de nature qualitative – va s'ajouter aux données massives dont nous disposons déjà. Le croisement de ces données nous donnera de nouvelles informations. Dans le langage des économètres, les interactions entre ces variables, au sein de bases de données plus complètes, permettront la genèse d'informations extrêmement importantes dont nous ne disposions pas auparavant. Ces interactions peuvent aussi être reliées aux métadonnées qui sont ces informations annexes associées.

Les métadonnées

Les métadonnées sont des données structurées qui accompagnent principalement des données non-structurées. Les métadonnées caractérisent partiellement certaines données non-structurées mais n'aident pas réellement à traiter l'information de la donnée non-structurée. Pour revenir à l'exemple des gazouillis, les métadonnées liées à un gazouillis sont : la date, l'expéditeur, le ou les suiveurs, la présence ou non d'un lien, si c'est un gazouillis re-transféré et si oui, de qui, etc. Bien que ces informations caractérisent parfaitement le gazouillis, le sens de ce dernier ne peut être dévoilé en exploitant seulement les métadonnées.

C'est donc l'ensemble de ces données, structurées et non structurées, qui donne aujourd'hui une information de meilleure qualité. En termes d'implications, cela veut dire par exemple que nous pourrions faire une meilleure gestion des risques. Néanmoins, les enjeux sont encore mal définis. Les écoutes de la NSA par exemple, viennent poser la question du point limite d'entrée dans la vie privée pour lutter contre des risques tels que le terrorisme. Pourtant, nous nous rendons bien compte de l'importance de ces données. Elles recèlent une quantité d'informations dont nous ne disposions pas avant l'apparition des premiers téléphones intelligents en 2006 ou des tablettes en 2010, et avant l'apparition des réseaux sociaux comme Facebook ou Twitter. Le matériel et le logiciel ont commencé leur convergence au début de ce XXIe siècle. Et c'est là où se trouve le point de rupture.

C'est donc l'ensemble de ces données, structurées et non structurées, qui donne aujourd'hui une information de meilleure qualité.

En effet, cette convergence constitue ce que les économistes appellent soit une innovation radicale de procédé (Carlton & Perloff, 2005) soit une technologie générique (General Purpose Technology) (Rousseau, 2008).

Une autre catégorisation est celle du Manuel d'Oslo - un projet conjoint entre l'OCDE et l'Union européenne (OECD & Eurostat, 2005). Dans cette catégorisation des innovations, il y a les innovations de produit, les innovations de procédé (dues par exemple à des changements technologiques) et les innovations organisationnelles et de marketing. Les données massives sont donc ici à la fois des innovations de procédé et des innovations organisationnelles.

Les données massives sont donc ici à la fois des innovations de procédé et des innovations organisationnelles.

De plus, il est coutume d'ajouter une autre dimension qui est celle de l'importance de l'innovation : nous parlons alors d'innovation incrémentale ou d'innovation radicale. Une innovation radicale est une innovation qui n'est pas une amélioration graduelle ou continue mais qui fait faire un bond en termes de nouveau produit (le premier iPhone par exemple) ou de nouveau procédé (le Fordisme par exemple). Cette convergence entre le matériel et le logiciel rentre dans la catégorie des innovations radicales de procédé. Les façons de travailler, de communiquer, de collecter des informations ou de mesurer ne sont plus les mêmes, que ce soit pour des entreprises privées, des gouvernements ou les acteurs/nouveaux acteurs de la société civile.

Dans l'histoire de l'humanité, les révolutions industrielles ont toujours eu comme base l'apparition d'innovations radicales de procédé ou de technologies génériques. Ces dernières pouvaient venir de l'apparition de nouvelles technologies comme les machines ou l'électricité, mais ces nouvelles technologies avaient un impact sur les fonctions de production, c'est-à-dire sur les façons de faire les choses. La conséquence était la mise en place de nouveaux procédés de production, plus efficaces et donc plus rentables. Les pays qui faisaient ces adaptations les premiers conservaient leur avantage comparatif très longtemps.

Les données massives mettent à notre disposition de nouvelles informations qui ont déjà impacté nos façons de travailler. De plus en plus d'organisations (entreprises, gouvernements, organisations internationales gouvernementales ou non gouvernementales) ont déjà commencé à s'intéresser aux informations disponibles dans les données massives. Les opportunités sont énormes, il est clair que nous n'en sommes qu'au début du potentiel d'utilisation de l'information générée par les données massives. Pourtant, il y a une dimension intéressante : il ne s'agit pas

seulement d'utiliser l'information issue des données massives, il s'agit aussi de la propriété de ces données. C'est là où les données massives perdent leur caractère virtuel et peuvent être comprises comme une véritable ressource physique. En effet, les données massives sont bien souvent collectées, exploitées et valorisées par des entreprises privées. Même si elles sont collectées auprès d'utilisateurs résidant dans des pays différents, elles finissent par être exploitées par une entreprise qui a elle-même un siège social dans un pays. Les données ont une définition territoriale. L'appropriation de ces données par le collecteur fait que ces données virtuelles se voient conférer un ensemble de caractéristiques juridiques, notamment la propriété intellectuelle, qui va leur donner une quasi existence physique. Nous montrerons d'ailleurs dans ce rapport que les données massives ont une citoyenneté qui s'accompagne d'éléments tels que la dimension juridique associée à cette citoyenneté, la dimension politique, la dimension morale ou la dimension économique. Pour simplifier, nous parlerons alors de citoyenneté juridique, morale, politique et économique. Cette catégorisation nous servira de grille d'analyse.

Pour prolonger ce raisonnement et mesurer ses implications, si dans les années 1970 le pétrole était considéré comme le nouvel or et a fini par se faire appeler « l'or noir, » nous pouvons penser que les données massives sont le « nouvel or noir. »

PREMIÈRE PARTIE

La nature des données et les méthodes d'analyse

Les modèles d'analyse développés dépendent grandement de la qualité de l'accès aux données pertinentes. Les problématiques d'optimisation interne pourront souvent être résolues grâce à l'analyse des données propres à l'entreprise. Cependant, dès que les analyses sortent du périmètre de l'entreprise (clients, fournisseurs, nouveaux marchés,…), de nombreuses barrières à l'accès à l'information et aux données peuvent se dresser.

1.1 LA CARACTÉRISATION DES DONNÉES

Le McKinsey Global Institute propose 4 variables qui permettent de caractériser les données concernant leur utilisation (McKinsey & Company, 2014) :

1. Le degré d'accès aux données;
2. La compatibilité informatique des données;
3. Le coût des données;
4. Le droit d'utilisation et de diffusion des données.

Ces quatre caractéristiques permettent de déterminer à quel point ces données sont faciles d'accès et pleinement utilisables. De nombreuses organisations rendent disponibles leurs données à tous (« données ouvertes ») en utilisant leur site internet et contribuent ainsi à la génération de données massives. Une autre caractéristique qui pourrait être rajoutée est le type de données et l'utilisation directe possible ou non (données structurées ou non).

1.2 LA NATURE JURIDIQUE DES DONNÉES

Les données massives sont générées de façon virtuelle, mais ont également d'autres caractéristiques qu'il faut considérer. Il est important par exemple de bien définir la nature juridique des données : propriétaire, personnelle ou ouverte.

1.1.1. ***Les données propriétaires ou personnelles***

Propre à chaque entreprise, les données propriétaires sont celles issues directement de l'activité d'une entreprise. On retrouve ici les données clients, fournisseurs, les données transactionnelles, les données de production, des employés, les résultats financiers. Ces données ne sont généralement pas accessibles pour les acteurs externes à l'entreprise à moins de mettre en place un contrat permettant un échange. Les données personnelles des individus (par exemple leur numéro d'assurance sociale ou encore le nombre de comptes bancaires, etc.) leur appartiennent et ne sont pas accessibles. Certaines données ne doivent tout simplement pas se retrouver sur des sites ouverts à tous.

1.1.2. ***Les données ouvertes***

A l'opposé, certaines données peuvent être partagées librement par les gouvernements, certaines entreprises ou organisations non-gouvernementales, mais aussi les données issues de la recherche. Par exemple, les données ouvertes gouvernementales sont disponibles à tous et permettent notamment d'avoir accès à de l'information pertinente sur les marchés du pays. Le site américain, data.gov, donne accès à plus de 85 000 banques de données. Plus de 40 pays dans le monde ont mis en place ce genre de plateforme afin de donner libre accès à des données fondamentales. Selon le *Open Data Index* de la *World Wide Web Foundation*, les 5 pays les plus performants dans la diffusion de leurs données sont les États-Unis, le Mexique, Singapour, le Royaume-Uni, et la Nouvelle-Zélande. Le Canada est 8e.[1]

Un autre exemple est le portail des données ouvertes de la Ville de Montréal. La Ville « a ouvert ses données à tous et permet qu'elles soient réutilisées à différentes fins, incluant des fins commerciales. Les résultats de cette réutilisation peuvent ensuite être partagés dans la communauté, ce qui crée un effet démultiplicateur. Les données libérées et réutilisées génèrent ainsi des bénéfices à la fois dans les sphères économiques, culturelles, sociales et technologiques » (http://donnees.ville.montreal.qc.ca). Par exemple, La Direction du greffe a libéré des données sur les contrats et les subventions versées par la Ville mais aussi d'autres types de données sont disponibles dont la liste des emplacements de tous les feux de circulation dont au moins une traverse est munie d'un feu sonore pour

1 http://www.webfoundation.org/2012/09/introducing-the-open-data-index/od-index/

malvoyants, la géolocalisation des arceaux à vélo sur le territoire de la Ville de Montréal, des photographies aériennes de la ville, etc.

Il y a aussi toutes les données transmises par les utilisateurs sur les réseaux sociaux qui peuvent être accessibles à tous.

1.3 LES MÉTHODES D'ANALYSE DES DONNÉES MASSIVES

La grande question est de savoir si nos outils quantitatifs sont suffisants pour analyser toutes ces données. Y a-t-il une difficulté supplémentaire avec les données non structurées? Commençons par une tentative de définition des données non structurées. Ce sont en réalité des données qui ont à l'origine plutôt une nature qualitative et qui sont néanmoins traduites en code informatique pour pouvoir être utilisées sur nos appareils électroniques (téléphones intelligents, tablettes, ordinateurs, etc.). Le fait que des informations qualitatives soient traduites en code informatique va permettre d'analyser de façon quantitative cette information. Il faudra pour cela structurer ces données. Par exemple, le fait d'accepter une personne en tant qu'amie sur un réseau social relève de l'émotionnel. Pourtant, choisir une personne comme amie va créer un champ dans la base de données du réseau social qui associera ces deux personnes. Un autre exemple plus complexe est celui de l'analyse sémantique : il s'agit par exemple de permettre la conversation dans une langue entre deux personnes et d'être capable d'analyser si cette conversation est de nature positive, négative, joyeuse, agressive, etc.

Des modèles d'analyse des données quantitatives existent déjà depuis les premiers travaux sur les probabilités et l'analyse stochastique et n'ont eu de cesse d'être améliorés à travers le temps. Des outils d'analyse de très grosses bases de données avec une puissante capacité de calculs ont été développés.

De plus, pour aller explorer et chercher des données, les méthodes de fouille de données sont un point de départ évident (Domingos & Hulten, 2000), mais il faut aussi tenir compte des spécificités des données (Gama, Sebastião, & Rodrigues, 2009). On pourra retrouver avec intérêt quelques revues de littérature sur le sujet (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Les méthodes de *machine learning* sont un autre point de départ évident. Très populaires dans les années 1970, elles ont été aussi complétées par les modèles d'économétrie linéaires de première génération (Choi & Varian, 2012; Shapiro & Varian, 1999). La disponibilité d'une puissance de calcul importante avec le développement de nouveaux ordinateurs avait permis ce développement. Aujourd'hui, considérant la nature des données et

leur disponibilité massive, les méthodes issues du *machine learning* trouvent des domaines d'application intéressants, par exemple les arbres de régressions ou les analyses factorielles sont parfois plus efficaces que les modèles d'économétrie de première génération. Les modèles d'économétrie de deuxième génération avec des estimations non linéaires s'ajoutent à la panoplie des outils à la disposition des analystes de données massives.

Pourtant, un autre défi analytique aujourd'hui est d'être capable de structurer toutes ces nouvelles informations à notre disposition, qu'elles soient de nature structurée ou non structurée. Lorsqu'une personne est équipée d'un téléphone intelligent est qu'elle est active sur les réseaux sociaux, elle génère en effet un nombre considérable d'informations à la seconde : son emplacement géographique, ses commentaires sur Twitter, ses interactions sur Facebook, ses photos sur Instagram, ses lectures sur iBook, sa musique sur iTunes, etc. Pour l'instant, un tiers de l'humanité est connectée et génère ce genre d'informations, mais bientôt ce sera toute l'espèce humaine qui sera connectée et qui participera à la genèse continuelle de ces données. Ainsi, il sera possible de comparer des groupes d'êtres humains entre eux, mais aussi l'évolution à travers le temps. Néanmoins, pour structurer ces données et les rendre analysables, il faut développer des méthodes de structuration et c'est là où se trouve ce nouveau champ disciplinaire qui est celui de l'analyse des données massives. En effet, il ne s'agit pas seulement d'utiliser une série de techniques économétriques ou d'algorithmes provenant des sciences de l'information, mais il faut aussi s’assurer que les meilleures méthodes des différents champs disciplinaires intéressés soient connues et utilisées pour analyser les données massives avec toutes leurs spécificités. Le champ disciplinaire portant sur les données massives est multi-disciplinaire par essence, mais il est aussi inter-disciplinaire, c'est-à-dire à la croisée de plusieurs disciplines, créant ainsi son existence propre. De nouveaux projets de recherche et des nouveaux centres de recherche se créent actuellement pour développer des outils qui permettront d'exploiter le potentiel des données massives.

Il faut s’assurer que les meilleures méthodes des différents champs disciplinaires intéressés soient connues et utilisées pour analyser les données massives.

Nous le comprenons donc bien, il y a des opportunités encore non explorées avec l'analyse des données massives. Mais il y a aussi des risques par exemple pour la protection des données propriétaires ou personnelles et la protection de la vie privée. Allons-nous assister à de grands changements de paradigme dans nos sociétés?

DEUXIÈME PARTIE

Les données massives : innovation ou révolution?

Les données massives correspondent à la notion d'innovation dans sa grande définition : elles sont le résultat d'innovations de produits et elles constituent une innovation de procédé. Elles ont un impact sur les opérations dans les entreprises, sur les procédés d'affaires, sur les stratégies des entreprises, mais aussi sur les institutions gouvernementales et sur les choix des individus.

Les données massives sont utilisées dans un nombre de secteurs de plus en plus important. Nous allons présenter des exemples d'applications en finance, en santé publique, en politique et en éducation. Il est important aussi de bien comprendre que les données massives créent de nouveaux marchés pour des entreprises qui gèrent ces données massives mais aussi pour les entreprises existantes qui peuvent utiliser ces données pour mieux définir leurs stratégies, cibler leur campagne marketing, connaître leurs marchés, etc. Dans ce qui suit, nous présentons une revue de la littérature portant sur les secteurs auxquels la littérature académique s'est intéressée. Ceci ne veut pas dire que seuls ces secteurs sont intéressants, il s'agit simplement du fait que la littérature est vraiment toute nouvelle sur ces sujets.

2.1 LES OPPORTUNITÉS POUR LES ENTREPRISES

Les données massives sont en elles-mêmes un marché. La Silicon Valley est un des exemples – et certainement le plus probant – de la convergence entre les produits de haute technologie et les services autour de ces produits. Aujourd'hui, des entreprises initialement différentes comme Apple et Google se retrouvent en réalité sur les mêmes marchés. À l'origine, Apple fabrique des ordinateurs et Google rend un service de recherche sur Internet. Rien ne les opposait. Aujourd'hui, leur modèle d'affaires repose de plus en plus sur leur capacité à générer, attirer, collecter et analyser des données à des fins de revenus soit publicitaires (Google), soit en direction de la vente de produits (Apple). Nous pouvons ajouter des

entreprises comme Amazon qui font leur virage vers les données massives, mais aussi de façon peut-être moins intuitive des entreprises comme Target ou Wal-Mart. L'objectif est la capture de ces données afin de les utiliser soit de manière à générer des revenus directs (ventes ou publicité), soit des revenus indirects (raisons stratégiques).

Des analyses des données massives permettraient par exemple de relier les historiques d'achat. Cette utilisation commune de l'ensemble des données disponibles permet de faire émerger des relations jusqu'alors non apparentes, à l'image de la publicité ciblée d'Amazon qui propose des recommandations basées sur les historiques d'achats et de visites des clients.

Les données massives ont une dimension stratégique également au niveau de l'analyse des secteurs industriels et de leur évolution. Par exemple, elles vont changer la donne pour prédire les évolutions technologiques, le comportement des utilisateurs de services et des consommateurs de produits.

Il est certain que les données concernant les utilisateurs ont une très grande valeur et sont depuis longtemps au cœur des grandes questions sur les risques associés avec le développement des services en ligne (Wang, Lee, & Wang, 1998). Les utilisateurs sont souvent conscients de l'arbitrage qu'ils font lorsqu'ils donnent accès à leurs informations : ils comparent ce coût indirect d'accès à leur information avec le bénéfice rendu par le site Internet ou le service en ligne (Hann, Hui, Lee, & Png, 2003).

Mais au-delà des innovations de produits, il faut garder en tête que les données massives sont une innovation radicale de procédé. En effet, leur utilisation change la finesse des analyses et des procédés organisationnels. Ce nouvel outil permet d'orienter de façon plus efficace les procédés organisationnels (par définition) au sein même d'une entreprise, mais également à propos de la croissance externe des entreprises qui feront affaire avec de nouveaux partenaires qui auront été identifiés grâce aux conclusions issues de ces données.

Au-delà des innovations de produits, il faut garder en tête que les données massives sont une innovation radicale de procédé.

2.2 LE DOMAINE DE LA FINANCE

L'industrie financière repose grandement sur l'accès et l'analyse de l'information. Dans ce sens, les données massives permettent aux acteurs financiers d'accéder à plus de données, et après analyse, à plus d'information. Il y a des propriétés propres à chaque type de données et une utilisation qui est différente. Par exemple, les données financières sont utilisées et le resteront pour l'information contextuelle

qu'elles produisent. Cette information est critique pour les deux phases importantes en finance : la phase d'analyse et la phase de prise de décision. Les données massives peuvent être utilisées dans la phase d'analyse pour compléter les données financières et participer à améliorer l'information contextuelle sur les caractéristiques intrinsèques des produits financiers, mais elles seront surtout utilisées dans la phase de prise de décision pour informer les analystes des conclusions qu'ont tirées les autres analystes sur les marchés. Dans cet exemple, les données massives servent à mesurer l'interprétation des autres. Il s'agit donc d'une aide dans la mesure du résultat – l'interprétation des données financières – davantage que de la cause (l'information financière sur les caractéristiques intrinsèques).

Il est intéressant de noter que les nouveaux outils pour l'analyse des données massives peuvent aussi servir à assurer la rétro-compatibilité avec les théories financières qui utilisaient d'autres outils (Gao, Hongkong, & Chan, 2000). Comme nous l'avons dit en introduction, les informations disponibles sur les réseaux sociaux peuvent préciser ce que pensent les utilisateurs, leurs réactions ou interprétations après la publication d'un rapport annuel d'une compagnie ou ses données financières, etc. L'étude de la finance par le prisme du comportement des agents économiques a mené au développement de la finance comportementale. Cette discipline se base sur le fait que les décisions des investisseurs sont aussi influencées par leurs émotions, tant positives (optimisme, espoir, joie) que négatives (pessimisme, suspicion, conservatisme). La méthodologie utilisée en économie expérimentale repose sur des expériences en laboratoire où les conditions dans lesquelles les opérateurs de marché prennent leurs décisions sont contrôlées. Nofsinger propose d'observer l'évolution des cours de bourse comme le reflet de l'humeur générale, induisant une possibilité d'anticipation des performances boursières (Nofsinger, 2005). Cette notion d'avis ou d'humeur générale semble de plus en plus accessible avec le développement des applications sur Internet (Godbole, Srinivasaiah, & Skiena, 2007).

2.2.1 Le lien entre réseaux sociaux et cours de bourse

Avant la démocratisation des réseaux sociaux comme Twitter ou Facebook, la recherche scientifique s'est intéressée au rôle des forums de discussion et leur influence sur les mouvements boursiers (Ruiz, Hristidis, Castillo, Gionis, & Jaimes, 2012). Au lieu d'utiliser les informations issues d'une expérience en laboratoire, il est possible d'utiliser l'information issue des réseaux sociaux. L'analyse des

messages publiés sur les blogues spécialisés tels que Raging Bull ou HotCopper a démontré l'existence d'une relation entre les messages publiés et la volatilité des cours des actions (Antweiler & Frank, 2004), puis entre l'effet des rumeurs sur les volumes d'échange (Clarkson, Joyce, & Tutticci, 2006).

L'année 2005 marque la mise en ligne de Twitter, service Internet permettant de publier de courts messages de 140 caractères à partir de téléphones mobiles et d'ordinateurs. Des centaines de millions de messages s'échangent sur Twitter. Ce volume de messages permet de faire l'hypothèse qu'il est possible d'en extraire des informations pertinentes (Thomas & Sycara, 2000). Twitter se présente comme un réseau social « ouvert », facilitant l'accès aux messages publiés à travers son interface de programmation. Cette dernière caractéristique peut être considérée comme un facteur essentiel dans la production scientifique liée aux réseaux sociaux (à l'instar des travaux concernant l'encyclopédie collaborative Wikipédia). Ainsi, de nombreuses équipes de recherche se sont penchées sur la relation entre cotations boursières et messages publiés quotidiennement sur Twitter. Une majorité des travaux académiques se sont concentrés sur l'analyse du contenu des messages mis en ligne, aussi appelée « analyse de sentiment » (Pang & Lee, 2008). Les travaux de John Bollen et Hui Mao de l'Université d'Indiana apparaissent comme référence dans le traitement des messages de Twitter appliqué à la finance (Bollen, Pepe, & Mao, 2009). En développant un système d'analyse sémantique, ils assignent six types d'émotion à chaque message traduisant l'état d'esprit des utilisateurs : en contrôle/anxieux, confus/clair d'esprit, certain/incertain, fatigué/vif, agréable/hostile, joyeux/dépressif. Ils ont remarqué que la catégorie correspondant aux émotions du contrôle de soi (en contrôle ou anxieux) avait l'incidence la plus élevée sur les résultats boursiers des compagnies. Ainsi, ils réussissent à prédire l'évolution du Dow Jones Industrial Average avec 4 jours d'avance dans 86,7 % des cas (Bollen, Mao, & Zeng, 2011). Cette capacité de prédiction s'avère particulièrement précise dans les périodes de volatilité sur les marchés boursiers. D'autres auteurs mettent au point une méthodologie représentant les notions de peur et d'espoir collectifs, qui permet de corréler les indicateurs financiers (NASDAQ, S&P500, DJIA, VIX) avec les messages fortement « émotifs » (Zhang, Fuehres, & Gloor, 2011).

2.2.2 Le rôle des « influenceurs » sur les réseaux sociaux

Au-delà de la question du rôle de la masse, il y a aussi la question des influenceurs (Goldsmith & Desborde, 1991; Konrad, 2003). Certains auteurs se penchent sur

l'influence de certains utilisateurs au sein du réseau social (Brink, Rusinowska, & Steffen, 2011; Rijnsoever, 2008). Grâce au nombre de suiveurs qu'une personne possède, il est possible de connaître la taille et la composition de l'audience qui recevra les messages émis. Bar-Haim, Dinur, Feldman, Fresko, & Goldstein (2011) et Brown (2012) s'intéressent ainsi à la réputation des émetteurs de messages, en tentant d'identifier les acteurs influents.

Observant cette capacité d'anticiper avec justesse l'évolution des cours boursiers, plusieurs fonds d'investissement à caractère technologique ont été mis sur pied (Jordan, 2010). Derwent Absolut Return gère un fonds de 40 millions de dollars en se basant sur les travaux de Bollen et de Mao, mais aussi sur les messages échangés sur Facebook. MarketPsy Capital s'est intéressé au sentiment d'investisseurs concernant 6 000 compagnies et a obtenu un rendement de 45 % au cours de l'année 2008. Le fonds Flyberry Capital se spécialise dans le traitement des données massives et recoupe des informations provenant de multiples sources : météo, détection de séismes, sites gouvernementaux, blogues, médias sociaux. En seulement quatre mois, ils ont obtenu un rendement cumulatif de 30,6 % (Huang, 2012). La grande question est peut-être de savoir si l'on peut prévoir l'avenir avec les réseaux sociaux (Asur & Huberman, 2010).

2.3 LE DOMAINE DE LA SANTÉ

À la vue de la quantité et aussi de la diversité des données qu'une personne produit de façon quotidienne, l'utilisation des données massives s'impose comme une stratégie incontournable pour le secteur de la santé et plus précisément la santé publique.

2.3.1 Les réseaux sociaux et le suivi des maladies en temps réel

Les messages publiés sur Internet par les utilisateurs permettent d'avoir des informations concernant leur santé, leurs habitudes de consommation, par exemple un régime alimentaire spécifique à l'individu, leur pratique sportive, leur utilisation de services, leur exposition aux agents cancérigènes comme la pollution propre à chacun selon sa situation géographique, etc. Les avancées technologiques et cette diversité des données permettent de colliger des informations concernant par exemple la propagation et le suivi des maladies en temps réel.

Ginsberg et al. démontrent en 2009 que l'utilisation de messages instantanés comme ceux publiés sur Twitter permet d'améliorer le suivi en temps réel d'épidémies de grippe (Ginsberg et al., 2009). Alors que les structures

gouvernementales américaines telles que les CDC (*Centers for Disease Control and Prevention*) publient des communiqués à des intervalles allant jusqu'à deux semaines de l'évolution des maladies, les auteurs ont réussi à suivre en direct la propagation de la grippe sur neuf États américains. Ce décalage est confirmé à travers une autre étude estimant que Twitter s'avère adéquat afin de modéliser les éclosions de foyers de propagation de maladies (Culotta, 2010). Cette capacité à réagir rapidement s'explique par le fait que les données non structurées telles que les messages sur Twitter ou les requêtes de moteurs de recherche sont non seulement des données générées en temps réel, mais peuvent être aussi associées à une situation géographique précise (option de géolocalisation pour Twitter, adresse IP pour le cas des moteurs de recherche). Google illustre le potentiel de l'utilisation de ce type de données afin de cartographier la propagation en temps réel de la grippe et de la dengue en utilisant les requêtes générées par les utilisateurs du moteur de recherche.

2.3.2 L'utilisation des données massives et les applications médicales

La vision de Google des défis reliés à la santé est une vision en accord avec son cœur de métier, c'est-à-dire le traitement et l'archivage de données massives. Les initiatives gouvernementales ont pris de l'ampleur au cours des dernières années en mettant en place des structures visant à favoriser l'émergence de l'utilisation des données massives en santé (comme le séquençage des génomes, l'imagerie, etc.)

Le Big Data Institute de l'Université d'Oxford au Royaume-Uni, doté de 30 millions de livres sterling, se concentrera sur l'analyse de dossiers médicalisés de patients, d'essais cliniques et de séquences d'ADN grâce au travail de 600 chercheurs. Toujours au Royaume-Uni, le Wellcome Trust Sanger Institute, centre de recherche dédié au séquençage de l'ADN, a noué un partenariat avec la compagnie DataDirect Networks afin de traiter les 10 pentabytes de données générés quotidiennement par les séquenceurs d'ADN de l'Institut. Aux États-Unis, les National Institutes of Health ont mis sur pied un programme afin de favoriser l'accès aux nouvelles approches du Big Data pour des applications biomédicales (Big Data to Knowledge). Au Canada, Génome Canada a lancé à la fin 2013 un appel de candidature pour des projets reliés à l'utilisation des données massives dans la recherche génomique. Ces programmes gouvernementaux auront des impacts significatifs sur la génomique ou la médecine personnalisée.

2.4. LE DOMAINE POLITIQUE : LES DONNÉES MASSIVES ET LES ENJEUX INSTITUTIONNELS DU XXIE SIÈCLE

Les réseaux sociaux peuvent être d'une grande utilité pour les gouvernements. Ils sont le reflet de la société et des thématiques importantes pour les gens débattant sur les réseaux. Il est possible de mesurer les émotions et le sentiment public (Bollen et al., 2009).

2.4.1 Les réseaux sociaux, la dynamique démocratique et l'activisme mondial

Les réseaux sociaux sont aussi un formidable véhicule pour les organisations non gouvernementales qu'elles soient nationales ou internationales. La relation entre l'Internet, les réseaux sociaux et l'activisme mondial a commencé à être documentée (Bennett, 2003).

Les réseaux sociaux sont extrêmement importants à étudier pour estimer leurs impacts sur la dynamique démocratique. Par exemple, les élections américaines de 2008 furent considérées comme les premières élections 2.0. Twitter et Facebook devinrent les véhicules de communication des candidats, notamment pour atteindre une tranche de population plus connectée aux nouvelles technologies. Depuis, toute campagne d'élection ne se planifie sans une plateforme Internet et médias sociaux. Les données massives apportent transparence, participation et interaction citoyenne, tout en accentuant la responsabilité des candidats aux élections envers les électeurs, concepts pouvant être quantifiés par les techniques d'analyse de sentiment, d'analyse de réseaux et de fouille de données (Hsinchun, Chiang, & Storey, 2012). Plusieurs auteurs se sont donc penchés sur l'étude des médias sociaux dans un cadre d'anticipation de résultats politiques. Tumasjan, Sprenger, Sandner, & Welpe (2010) observent l'activité sur Twitter dans une période précédant les élections nationales allemandes de 2009. En collectant le nombre de messages publiés sur les six partis principaux, ils posent la question de savoir si Twitter ne peut être considéré que comme un moyen d'expression pour individu, ou bien plutôt comme un véhicule de conversation politique. La seconde option est celle vérifiée par les auteurs. En effet, en analysant et le nombre de messages et leur contenu (analyse de sentiment), ils démontrent que les marges d'incertitude sont légèrement supérieures à celles fournies par les sondages traditionnels (1,65 % pour Twitter contre de 0,80 % à 1,48 % pour les sondages traditionnels).

Un groupe de chercheurs utilise une méthodologie similaire basée sur l'analyse de sentiment de messages publiés sur Twitter. Bermingham & Smeaton (2011) s'intéressent aux élections irlandaises de 2011, et démontrent que l'utilisation de l'analyse de sentiment vient affiner les modèles prédictifs basés seulement sur le volume de messages mentionnant les candidats aux élections. O'Connor, Balasubramanyan, Routledge, et Smith (2010) étudient les élections américaines et trouvent une corrélation (73,1 %) entre les modèles prédictifs basés sur des messages de Twitter et les résultats aux sondages. Ils expliquent que Twitter apparait comme un outil de prédiction afin de connaitre les opinions des électeurs, notamment par le fait que les données soient disponibles et accessibles bien avant les sondages officiels et à une fréquence plus soutenue. Malgré ces études encourageant le développement de l'utilisation des données massives en politique, Gayo-Avello, Metaxas, & Mustafaraj (2011) tentent de nuancer ces résultats. En effet, leur étude concernant les élections du Sénat et du Congrès américains de 2010 révèle que la classification de messages en tant que messages à caractère positif, négatif ou neutre n'a pu permettre de prédire adéquatement les résultats de ces deux élections. Cette nuance permet donc de souligner deux facteurs permettant d'utiliser les données massives dans le cadre d'applications politiques : (1) la population présente sur Twitter devrait être représentative des pays dans lesquels se déroulent des élections et (2) l'utilisation des médias sociaux dans ces pays se révèle adéquat à véhiculer une conversation politique.

L'utilisation de l'analyse de sentiment vient affiner les modèles prédictifs basés seulement sur le volume de messages.

2.4.2 Les conversations sur Twitter et les élections provinciales au Québec

Le CIRANO a étudié les élections au Québec de 2012 et de 2014 en s'appuyant sur les conversations sur le réseau Twitter (Warin, Sanger, & Troadec, 2014). Les auteurs suivent le nombre de gazouillis pendant toute la durée de la campagne et aussi pendant des événements spéciaux comme les débats entre les chefs de parti. Ils ont noté des changements dans les thématiques et les mentions des chefs de parti bien avant que ces changements soient notés par les entreprises de sondage par exemple. La réactivité sur les réseaux sociaux est intéressante à visualiser et analyser en direct. Les partis politiques peuvent savoir en temps réel quelles sont les thématiques qui trouvent un écho auprès des utilisateurs des réseaux sociaux et quelles sont celles qui ne sont pas relayées. Il y a certes un danger de populisme, mais il y a aussi l'avantage de mieux comprendre les aspirations de la population.

Il est aussi intéressant de comprendre si une thématique - pourtant jugée importante et prioritaire pour un pays - est bien comprise par la population. Les

analyses permettent de mesurer la différence entre la perception et la réalité des priorités pour un futur gouvernement. Si la perception est aussi un élément de la démocratie, et les priorités objectives un élément de la république, alors les données massives et leur utilisation dans l'analyse politique illustrent en ce début de XXIe siècle cette différence entre démocratie et république. La plupart des sociétés occidentales sont un équilibre savant de forces démocratiques au sens d'Abraham Lincoln qui voit dans la démocratie une forme de gouvernement. La dynamique des réseaux sociaux crée une nouvelle pression sur les formes de gouvernement des sociétés en changeant la définition de la démocratie en l'éloignant un peu de la définition d'Abraham Lincoln et en la rapprochant de celle d'Alexis de Tocqueville (Tocqueville & Tocqueville, 2002) et de Karl Popper (Popper, 1949). Pour être plus spécifique, la dynamique des réseaux sociaux fait que les différences culturelles et les différents groupes d'une société civile ont maintenant une autre forme de représentation sur ces réseaux en plus de leur représentation, parfois partielle, au travers des institutions démocratiques de leur pays. Les réseaux sociaux augmentent donc l'importance de la composante organique des sociétés et soulèvent des questions sur les institutions des pays qui sont le fruit de l'héritage du XXe siècle. Les données massives générées par ces réseaux sociaux permettent de suivre de mieux en mieux ces évolutions organiques des sociétés. En ce sens, elles sont un outil important pour les gouvernements dans leur réflexion sur les réformes institutionnelles à considérer à l'aube de ce XXIe siècle.

Les données massives contribuent à la propagation du savoir en dévoilant de nouvelles possibilités jusqu'alors inexistantes.

2.5 Le domaine de l'éducation et des connaissances

Une des promesses d'Internet est l'accès universel à l'information. Les données massives contribuent à la propagation du savoir en dévoilant de nouvelles possibilités jusqu'alors inexistantes : suivre des cours universitaires poussés en ligne, rétroaction de dizaines de milliers d'utilisateurs, mise en lumière de zones d'exclusion de l'information et mise en valeur du contenu culturel font partie de la nouvelle réalité qu'offrent les données massives. Avec plus de 200 millions de blogues en ligne, 500 millions d'utilisateurs sur Twitter, 350 millions sur Sina Weibo, 1 milliard de comptes ouverts sur Facebook, l'information ou la connectivité entre les utilisateurs semble infinie. Un blogue ne peut à lui seul changer l'information au sein de cet océan de données, mais le tissu social bâti sur Internet permet de transmettre l'information plus rapidement et de manière plus globale.

2.5.1 La localisation géographique des utilisateurs produisant l'information sur les réseaux sociaux

Un problème est à soulever, celui de la représentation des utilisateurs produisant l'information. En effet, certaines zones géographiques ou tranches de la population ne participent pas à la création de contenu sur Internet (ou restent sous représentées, d'où la notion de « data shadows » (Graham, Hale, & Stephens, 2012). Cette disparité dans la présence numérique de certaines zones géographiques est révélée aussi à travers l'étude des articles de l'encyclopédie collaborative Wikipédia. En analysant la provenance des articles géolocalisés, on remarque que des zones sous représentées émergent, notamment l'Afrique du Nord, l'Afrique sub-saharienne et le Moyen-Orient. Cette différence par rapport à des régions plus présentes s'explique notamment par un réseau Internet moins fiable et moins efficace (Graham, Hogan, Straumann, & Medhat, 2014). Les données massives permettent ainsi de mettre en lumière et de cartographier ces zones d'ombre de données.

2.5.2 L'accès à la connaissance et le développement de cours ouverts et massifs

Un autre apport des données massives dans l'accès à la connaissance est le développement au cours des dernières années des cours en ligne ouverts et massifs, ou MOOCs (Massive Open Online Courses). Des professeurs universitaires proposent sur Internet de donner un séminaire ouvert à tout internaute désirant suivre ce cours, allant de l'histoire de l'art au traitement de données et à l'intelligence artificielle. Plusieurs plateformes existent afin d'avoir accès à un ensemble varié de plus de 500 cours, notamment Coursera, EdX (joint-venture de MIT et de Harvard) et Udacity (Belleflamme et Jacqmin, 2014). Le rôle des MOOC's est d'offrir à tous la possibilité de suivre des cours en ligne.

À l'automne 2012, l'Université Stanford a offert un cours en ligne sur l'intelligence artificielle où 104 000 étudiants se sont inscrits tandis que 120 000 étudiants se sont inscrits au cours conjointement donné par le MIT et par Harvard sur les circuits et l'électronique (McKenna, 2012). Cette interaction jusqu'alors sans précédent permet notamment une récupération de l'information des étudiants à travers des tests menés sur des cohortes de très grande taille. En 2012, ce sont plus de 1,7 million d'étudiants qui ont suivi en ligne les cours offerts par Coursera (Pappano, 2012). Toutefois en matière de mesure d'impact, ces constats sont à mitiger car beaucoup de ces étudiants ont déjà des diplômes et la plupart ne font pas les examens (et donc ne valident pas le cours).

En conclusion de cette section, nous comprenons bien qu'au-delà de l'innovation, les données massives représentent une vraie révolution. Qu'en est-il des opportunités et des risques?

TROISIÈME PARTIE

Les données massives : entre opportunités et risques

Les opportunités, enjeux, défis et risques sont multiples. Comme pour toute innovation radicale, les usages peuvent à la fois créer des opportunités immenses, mais accompagnées aussi par des risques importants. Pourtant, par le passé, rares ont été les innovations qui ont généré des risques mal gérés par les sociétés. Quelle serait la raison qui ferait que les données massives pourraient faire courir des risques encore plus importants?

En théorie, la seule raison serait que l'on franchisse la frontière entre risque et incertitude (Knight, 1929). Le risque se définit comme un ensemble fermé d'évènements probabilisables, alors que l'incertitude est un ensemble ouvert d'évènements par définition non calculables. L'ambigüité associée à des situations d'incertitude nous ferait entrer dans des zones de « risque » non calculable. Et il est vrai que la nature des données massives est d'être une innovation tellement radicale qu'il nous est difficile d'imaginer dès à présent toutes les opportunités et donc aussi tous les risques. Pour une revue des risques, nous vous invitons à la lecture du rapport Ouellet, Mondoux, Ménard, Bonenfant, & Richert (2013).

3.1 LES ENJEUX RELIÉS À LA PROPRIÉTÉ DES DONNÉES : LA CITOYENNETÉ JURIDIQUE

Comme nous l'avons vu précédemment, il est souvent compris que les données sont des éléments virtuels et même lorsqu'elles sont archivées, elles le sont dans un « nuage. » La plupart du temps ce nuage est aux États-Unis.

La réalité est évidemment plus complexe. Les données représentent certes un concept un peu flou à priori, mais elles ont une véritable existence matérielle qui leur est conférée par (1) le droit de la propriété intellectuelle (condition pas nécessaire, mais suffisante) et (2) la juridiction où se trouvent les sièges sociaux des entreprises ou institutions collectant ces données ou des juridictions des filiales (condition nécessaire et suffisante).

Dans ces conditions, nous comprenons que les données ont une citoyenneté juridique.

3.2 LA PROTECTION DE LA VIE PRIVÉE : LA CITOYENNETÉ MORALE

L'année 2013 a mis en évidence de nombreuses atteintes à la vie privée des utilisateurs d'Internet et plus globalement des technologies de l'information (la NSA et le programme PRISME aux États-Unis, les actions du groupe Anonymous, Wikileaks...).

La littérature sur la protection de la vie privée et les risques encourus par la propagation des données massives n'en est qu'à ses débuts. Les premiers articles ont pour objectif de mettre l'accent sur l'existence des risques et la nécessité d'auto-discipline. Certains auteurs ont commencé à faire une revue de la littérature existante, nous regarderons avec intérêt Tucker (2012). Dans ces conditions, nous comprenons que les données ont une citoyenneté en tant que personne morale différente de la personnalité physique qui génère ces données.

Les données ont une citoyenneté en tant que personne morale différente de la personnalité physique qui génère ces données.

De plus en plus d'études présentent les risques pour la protection de la vie privée. Par exemple, les contrats d'utilisateurs incluant des clauses de protection de la vie privée en ligne que les utilisateurs acceptent lors de l'enregistrement à un site Internet sont souvent discutables. Une étude de 2005 portant sur 500 entreprises en ligne présente les risques de l'auto-réglementation et ses limites (Ashrafi & Kuilboer, 2005). Il existe aussi de grandes différences à travers les pays. Une étude de 2011 fait par exemple une comparaison entre les USA et la Chine en matière de protection des données concluant de façon non surprenante à une meilleure protection aux USA, mais surtout pointant les différences culturelles intéressantes (Wu, Lau, Atkin, & Lin, 2011).

Il existe aussi toute une littérature sur la perception qu'ont les utilisateurs à propos des contrats électroniques signés par les utilisateurs. Certaines études concluent que les violations de la protection de la vie privée sont perçues de la même façon qu'une simple rupture de contrat (livraison d'un produit avec un défaut, etc.) (Flavián & Guinalíu, 2006).

À des fins de protection de la vie privée, il est intéressant de noter que certains auteurs proposent de créer un intermédiaire entre l'utilisateur et le site qui propose le service en ligne. Cet intermédiaire serait en charge de servir de protecteur des données privées tout en rendant le service auquel l'utilisateur s'attend (Taylor & Wagman, 2008).

3.3 LES ENJEUX STRATÉGIQUES POUR LA GOUVERNANCE PUBLIQUE : LA CITOYENNETÉ POLITIQUE

Les données massives ont donc une citoyenneté juridique. Cette citoyenneté juridique trouve un prolongement naturel dans la **question** et la **gestion** de la citoyenneté politique.

3.3.1 Les enjeux pour les gouvernements : la question de la citoyenneté politique

En ce qui concerne la question de la citoyenneté politique, les données massives posent à la fois des risques et représentent des enjeux importants. Comme nous l'avons déjà souligné, les risques sont d'abord ceux reliés à la protection de la vie privée dans le cadre des données générés par les gouvernements à propos de leurs citoyens ou gérés par les gouvernements. Il est de plus en plus fréquent que les gouvernements mettent en place des structures de diffusion des données ouvertes et de gestion de la conformité avec la protection de la vie privée de ces données ouvertes (McKinsey & Company, 2014).

Mais les risques ne concernent pas seulement la question de la protection de la vie privée pour les données gouvernementales. Il y a aussi les risques reliés à la perte de contrôle sur les données. Prenons l'exemple du domaine sensible de la santé publique. La part du système de santé publique dans le budget des pays est de plus en plus importante (près de 50 % des dépenses du gouvernement du Québec par exemple). Les prévisions démographiques laissent à penser que le vieillissement de la population s'accompagnera d'une hausse grandissante des dépenses de santé. Certains pays ne pourront faire face à l'augmentation de ces dépenses, et en conséquence des files d'attente peuvent se mettre en place quant à l'utilisation du système de santé. En plus de ces failles dans la fourniture du service public, nous assistons également à la naissance de formidables innovations technologiques reliées notamment aux données massives.

Jusqu'au début du XXIe siècle, les données sur le patient étaient sous contrôle et obéissaient à des cadres juridiques bien délimitées dont les principes pouvaient être protégés devant des cours de justice. Aujourd'hui, l'innovation technologique fait que des compagnies comme 23andMe, Whithings ou Nike peuvent créer des services ou des produits qui vont collecter des informations biométriques personnelles. Un pèse-personne mesure le poids et envoie l'information sur le nuage dans le compte de la personne. Grâce à ce compte, l'entreprise connaît l'âge de la personne, son genre, etc. Avec les produits de compagnies comme Nike ou

Adidas, nous pouvons en plus y associer les périodes d'entrainement sportif des utilisateurs. Si l'on croise toute cette information avec les données génétiques de 23andme, on comprend alors que cette information est extrêmement sensible.

En plus d'être sensible, cette information est aussi unique dans le sens où les lois sur la propriété privée et l'utilisation des données personnelles empêcheraient les gouvernements de faire ces regroupements de bases de données. En revanche, la dynamique des fusions-acquisitions dans le secteur privé fait que ces regroupements de bases de données apparaissent comme un résultat organique, propre à la dynamique stratégique des entreprises. Même si les entreprises ne sont pas en dehors des lois, le défi pour les gouvernements est de suivre toutes ces fusions-acquisitions et de vérifier la conformité de ces bases de données avec le cadre juridique du pays. Nous comprenons bien que le bénéfice du regroupement de ces bases de données est immense. Dans notre exemple, l'innovation technologique permet de pallier les inefficiences dans l'offre du service public. Mais les risques sont aussi très importants.

Un autre enjeu est celui portant sur la citoyenneté juridique. En effet, les entreprises qui offrent ces produits ou services qui permettront en conséquence de générer ces bases de données sensibles et uniques ont-elles une représentation juridique – et donc des responsabilités – dans la juridiction de l'utilisateur du service ou du consommateur du produit? Si la réponse est négative, alors les risques ne sont pas facilement gérables.

Cette matérialisation des données au sein d'une juridiction représente le même phénomène que la découverte d'un puits de pétrole.

3.3.2 Les politiques industrielles d'innovation : la gestion de la citoyenneté politique

Avec cette citoyenneté juridique viennent aussi de grands privilèges pour le ou les pays qui les hébergent dans sa juridiction. En effet, cette matérialisation des données au sein d'une juridiction représente le même phénomène que la découverte d'un puits de pétrole. À partir de ces données brutes (non structurées), nous allons pouvoir les raffiner pour faire un parallèle avec le pétrole – c'est-à-dire les structurer – et donc créer de la valeur ajoutée dans le pays qui possède les infrastructures, le capital financier et le capital humain. Cela va aussi générer des externalités positives pour toute la chaîne de valeur autour de ces données massives. La différence avec le pétrole est que la ressource est abondante et en expansion perpétuelle, voire de façon exponentielle avec toutes les interactions de variables possibles et imaginables. À l'avenir, c'est donc un enjeu capital pour un pays que de posséder les données massives et de savoir les raffiner. Il faut aussi

bien comprendre que les entreprises qui travaillent avec les données massives sont mobiles, contrairement aux ressources naturelles. Et en cas d'expatriation, avec elles partent les ressources, à la différence des ressources naturelles. Dans ces conditions, nous comprenons que les données ont une citoyenneté politique.

Avec cette citoyenneté politique viennent aussi de nouveaux besoins de politiques industrielles ou de révision des politiques industrielles existantes, notamment en matière d'accompagnement de l'innovation. Les politiques d'innovation peuvent être impactées dans plusieurs dimensions :

- la **collecte** de nouvelles données. Les données massives rendent plusieurs services : (1) elles complètent, sans être un substitut, les données déjà collectées par les différents organes statistiques d'un pays, (2) elles ajoutent une autre dimension aux analyses plus traditionnelles maintenant issues des données statistiques, (3) elles peuvent parfois remplacer les données traditionnelles. C'est le cas par exemple de l'évolution annuelle de certains indicateurs qui en raison de cet horizon temporel annuel sont en réalité peu pertinents. Avec les données massives, nous pouvons avoir des données en temps réel de l'évolution de ces indicateurs. À titre d'exemple, nous pourrions imaginer une mesure de l'indice des prix à la consommation en temps réel, ou une mesure des investissements en R&D des entreprises ou des gouvernements en temps réel et aussi dans quels secteurs ces investissements sont réalisés.
- Les nouvelles opportunités d'**évaluation d'impact** des politiques gouvernementales. En effet, avec les données massives, il est tout à fait envisageable de mettre en place une politique et de mesurer son impact sur les secteurs concernés à travers de nouveaux indicateurs en temps réel ou des indicateurs statistiques traditionnels adaptés aux données massives afin d'avoir une évaluation d'impact en temps réel. Au-delà de la question de l'efficacité mesurée par les protocoles d'évaluation d'impact des politiques, nous pouvons aussi imaginer que de nouveaux outils peuvent être développés pour mesurer la perception des citoyens par rapport aux impacts des nouvelles politiques mises en place. Ainsi, parmi les citoyens, nous pouvons mesurer le degré de satisfaction des employés du secteur concerné par la politique mise en place, le degré de satisfaction des entrepreneurs ou des utilisateurs des services publics s'il s'agit d'une politique purement à destination du service public. Le risque dans ce cas est la dispersion des agences qui collecteraient et analyseraient ces données. À

des fins d'efficacité, il faudrait une agence gouvernementale en charge de la collecte et de l'évaluation d'impact. Dans ce cas néanmoins, le risque est la réconciliation de la concentration - et par conséquent l'agrégation des données - avec le respect de la protection de la vie privée. C'est un vrai défi pour les gouvernements. Au Royaume-Uni, le gouvernement a mis en place en 1998 Nesta (National Endowment for Science, Technology and the Arts) afin de favoriser l'innovation. Aujourd'hui Nesta est une agence importante pour le financement et le soutien à l'innovation mais aussi pour assurer la sensibilisation sur les opportunités et les risques reliés aux données massives. Cette dernière fonction permet d'ajuster le cadre légal en fonction des situations et d'avoir la meilleure réponse gouvernementale possible (Nesta, 2014).

3.4 LES ENJEUX STRATÉGIQUES POUR LES ENTREPRISES : LA CITOYENNETÉ ÉCONOMIQUE

Les données massives sont à l'origine de la révolution industrielle du début du XXIe siècle.

3.4.1 Données massives et avantages concurrentiels

Les entreprises qui ont accès aux données et auront les capacités d'analyser et de faire ressortir les éléments stratégiques importants pour elles auront des avantages concurrentiels importants sur leurs concurrents. Associé aux notions vues précédemment sur les enjeux géostratégiques, il apparaît que les États auront un rôle encore plus important à jouer à l'avenir et que le concept d'avantage concurrentiel des nations sera encore plus proche de celui d'avantages concurrentiels des entreprises.

Cet avantage concurrentiel des entreprises bénéficiant de cette localisation et de toutes les externalités positives posera de sérieuses pressions et menaces sur les autres entreprises. Ces dernières devront revoir leurs procédés d'affaires, leurs implantations géographiques, etc. Il y aura des coûts d'ajustement importants pour les entreprises suiveuses. Mais globalement, il est certain que le principe de destruction créatrice sera très profitable aux économies (Schumpeter, 1939). Nous ne pouvons pas nous passer de cette révolution. En revanche, l'avantage compétitif sera amplifié par le fait qu'une entreprise soit l'entreprise meneuse, celle qui fait le premier pas. L'entreprise gagnante risque de rafler toute la mise. Dans ces conditions, nous comprenons que les données ont une citoyenneté économique.

Il y a aussi des enjeux importants pour le gouvernement dans le cadre de ces politiques industrielles. En effet, la dynamique concurrentielle fait que les entreprises sont sans cesse en changement, les parts de marché évoluent, les entreprises font de la croissance interne et aussi externe par fusions-acquisitions par exemple. Ces évolutions de marché sont analysées par secteur par les gouvernements afin d'assurer la plus juste concurrence. Au Canada, le Bureau de la concurrence explique : « Le fait que la concurrence soit profitable tant pour les entreprises que pour les consommateurs est la principale hypothèse opérationnelle sur laquelle se fonde le Bureau de la concurrence » (Gouvernement du Canada, 2005). Le Bureau de la concurrence enquête notamment sur les sujets suivants :

- dynamique concurrentielle faussée : Fixation des prix par des entreprises concurrentes, l'abus de position dominante, truquage d'offres, fusions et acquisitions
- création d'asymétries d'information : indications fausses ou trompeuses, l'exclusivité, les ventes liées et la limitation du marché, télémarketing trompeur, pratiques commerciales déloyales.

Sur chacun de ces sujets, les données massives peuvent être utilisées soit en complément des données traditionnelles, soit comme données principales lorsque les données traditionnelles sont difficiles ou impossibles à obtenir. Par exemple, un critère important est la définition du marché pertinent. Il est souvent difficile de procéder à la définition de ce marché. Les données massives portant sur les consommateurs des produits ou services peuvent aider à mesurer l'impact d'une fusion-acquisition sur la taille du marché pertinent et les conséquences pour les consommateurs.

Un autre intérêt dans le cadre de cette question sur la définition du marché pertinent est que cela permet de pousser les analyses au-delà des frontières politiques. Bien souvent en effet, les fusions-acquisitions concernent des multinationales. Dans ce cas de figure, la définition de la taille d'un marché qui permet de continuer à assurer une concurrence juste est difficile et les données statistiques sont alors peu utiles étant donné qu'elles sont souvent nationales. Les données massives n'ont pas de frontières et peuvent venir pallier quelques-unes de ces difficultés.

3.4.2 Données massives et réputation des entreprises

La valeur intangible de l'entreprise va devenir sa priorité, c'est une véritable source de création de valeur (Bromley, 1993; C. J. Fombrun & van Riel, 1997; Charles J. Fombrun, 1996). La question de la réputation est un vrai enjeu pour les entreprises (De Marcellis-Warin & Teodoresco, 2012). La question de la crédibilité est importante (Castillo, Mendoza, & Poblete, 2011), mais la viralité d'une rumeur, vraie ou fausse, est un risque important. Les réseaux sociaux peuvent faire et défaire des réputations (Jansen, Zhang, Sobel, & Chowdury, 2009). Les messages diffusés sur les réseaux sociaux peuvent devenir viraux et endommager pour très longtemps la réputation d'une entreprise (Jatin, Sunaina, & Anupama, n.d.). La viralité est un risque réel (Jones, Temperley, & Lima, 2009; Leskovec, Adamic, & Huberman, 2007) et les auteurs proposent que ces risques soient pris très au sérieux par la direction et par les conseils d'administration (Burke, Cooper, & Martin, 2011; Larcker, Larcker, & Tayan, 2012). On trouve aussi dans la littérature des recommandations de communication sur les réseaux sociaux à destination des entreprises (Brammer & Pavelin, 2004; Coombs, 2007).

Enfin, au-delà de la réputation, il y aura la question de l'influence. Avec les réseaux sociaux, une grande réputation amènera aussi une grande crédibilité quant aux positions normatives de ces entreprises. Elles auront donc un nouveau pouvoir d'influence (Leavitt, Burchard, Fisher, & Gilbert, 2009).

CONCLUSION

Au début de ce XXIe siècle, nous vivons une véritable révolution industrielle. Les données massives sont à l'origine de cette révolution. La convergence entre le matériel et le logiciel fait que nous vivons une vraie innovation radicale de procédé. Cette convergence permet de collecter, d'analyser et de former de nouvelles réponses aux problèmes faisant face aux différents acteurs (entreprises privées, gouvernements, etc.). L'intuition voudrait que les secteurs des services sont plus concernés par les données massives, mais comme nous l'avons vu, le comportement des consommateurs de produits est générateur de plus en plus de données. Les données massives sont donc aussi bien issues des utilisateurs de services que des consommateurs de produits.

Ce n'est pourtant qu'un début. Nous connaîtrons dans les années à venir de nouveaux développements, de nouvelles utilisations et de nouvelles intégrations qui ouvriront de nouvelles possibilités. Avec ces nouvelles opportunités, nous ferons face également à de nouveaux enjeux et de nouveaux risques. Même si cela semble alarmant, c'est la nature même de toutes les innovations que d'amener de nouvelles questions. Jusqu'à présent, l'être humain a plutôt bien profité de ces innovations. Cela ne devrait pas être autrement avec celle-ci.

Les données massives sont en voie de s'imposer comme une révolution dont nous ne pourrons pas nous passer. Elles deviendront une base de départ pour l'analyse de la compétitivité, du niveau de concurrence et de la croissance d'un pays et de ses entreprises.

Les données massives représentent une innovation radicale de procédé, et comme toutes les innovations radicales de procédé, elles sont un des vecteurs principaux des révolutions industrielles. Cela a été le cas pour les révolutions industrielles précédentes. À l'aube du XXIe siècle, nous sommes en train de vivre la naissance d'une nouvelle révolution.

À l'aube du XXIe siècle, nous sommes en train de vivre la naissance d'une nouvelle révolution

Bibliographie

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise ? The Information Content of Internet Stock Message Boards. *Journal of Finance*, 1259–1294.

Ashrafi, N., & Kuilboer, J.-P. (2005). Online Privacy Policies: An Empirical Perspective on Self-Regulatory Practices. *Journal of Electronic Commerce in Organizations (JECO)*, *3*(4), 61–74.

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01* (pp. 492–499). Washington, DC, USA: IEEE Computer Society. doi:10.1109/WI-IAT.2010.63

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1310–1319). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145569

Bennett, L. (2003). New Media Power: The Internet and Global Activism. In *Contesting Media Power: alternative media in a networked world* (pp. 17–37). Lanham: Rowman & Littlefield.

Bermingham, A., & Smeaton, A. F. (2011). *On Using Twitter to Monitor Political Sentiment and Predict Election Results*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=B453FD4CEE82B9D1BB0CC329A0E8E021?doi=10.1.1.222.2863&rep=rep1&type=pdf

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, Pages 1–8. doi:10.1016/j.jocs.2010.12.007

Bollen, J., Pepe, A., & Mao, H. (2009). *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena* (arXiv e-print No. 0911.1583). Retrieved from http://arxiv.org/abs/0911.1583

Brammer, S., & Pavelin, S. (2004). Building a Good Reputation. *European Management Journal*, *22*(6), 704–713. doi:10.1016/j.emj.2004.09.033

Brink, R. V. D., Rusinowska, A., & Steffen, F. (2011). *Measuring Power and Satisfaction in Societies with Opinion Leaders : An Axiomatization* (Université Paris1 Panthéon-Sorbonne (Post-Print and Working Papers) No. halshs-00587726). HAL. Retrieved from http://ideas.repec.org/p/hal/cesptp/halshs-00587726.html

Bromley, D. B. (1993). *Reputation, image and impression management* (Vol. viii). Oxford, England: John Wiley & Sons.

Brown, E. (2012). Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market. *SAIS 2012 Proceedings*. Retrieved from http://aisel.aisnet.org/sais2012/7

Burke, R. J., Cooper, C. L., & Martin, G. (2011). *Corporate Reputation: Managing Opportunities and Threats*. Gower Publishing, Ltd.

Carlton, D. W., & Perloff, J. M. (2005). *Modern industrial organization* (4th ed.). Boston: Addison Wesley.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675–684). New York, NY, USA: ACM. doi:10.1145/1963405.1963500

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*, 2–9. doi:10.1111/j.1475-4932.2012.00809.x

Clarkson, P., Joyce, D., & Tutticci, I. (2006). *Market Reaction to Takeover Rumour in Internet Discussion Sites* (SSRN Scholarly Paper No. ID 889785). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=889785

Coombs, W. T. (2007). Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate Reputation Review*, *10*(3), 163–176.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 115–122). New York, NY, USA: ACM. doi:10.1145/1964858.1964874

De Marcellis-Warin, N., & Teodoresco, S. (2012). Corporate Reputation: Is Your Most Strategic Asset at Risk? *CIRANO Burgundy Report*, *2012RB-02*.

Diebold, F. X. (2012). *A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline, Second Version* (PIER Working Paper Archive No. 13-003). Penn Institute for Economic Research, Department of Economics, University of Pennsylvania. Retrieved from http://ideas.repec.org/p/pen/papers/13-003.html

Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 71–80). New York, NY, USA: ACM. doi:10.1145/347090.347107

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. doi:10.1609/aimag.v17i3.1230

Flavián, C., & Guinalíu, M. (2006). Consumer trust, perceived security and privacy policy: Three basic elements of loyalty to a web site. *Industrial Management & Data Systems*, *106*(5), 601–620. doi:10.1108/02635570610666403

Fombrun, C. J., & van Riel, C. (1997). The Reputational Landscape. *Corporate Reputation Review*, *1*(2), 5–13. doi:10.1057/palgrave.crr.1540024

Fombrun, Charles J. (1996). *Reputation: realizing value from the corporate image*. Harvard Business Press.

Gama, J., Sebastião, R., & Rodrigues, P. P. (2009). Issues in evaluation of stream learning algorithms. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 329–338). New York, NY, USA: ACM. doi:10.1145/1557019.1557060

Gao, X., Hongkong, S., & Chan, L. (2000). An Algorithm for Trading and Portfolio Management Using Q-learning and Sharpe Ratio Maximization. In*: Proceedings of the International Conference on Neural Information Processing, 2000* (pp. 832–837).

Gayo-Avello, D., Metaxas, P., & Mustafaraj, E. (2011). Limits of Electoral Predictions Using Twitter. Presented at the International AAAI Conference on Weblogs and Social Media (posters).

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. doi:10.1038/nature07634

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. *ICWSM'07*.

Goldsmith, R. E., & Desborde, R. (1991). A validity study of a measure of opinion leadership. *Journal of Business Research*, *22*(1), 11–19.

Gouvernement du Canada, I. C. (2005, April 25). Bureau de la concurrence. page d'accueil; Pages d'accueil; Pages de renvoi; Page de navigation. Retrieved March 29, 2014, from http://www.bureaudelaconcurrence.gc.ca/eic/site/cb-bc.nsf/fra/Accueil

Graham, M., Hale, S., & Stephens, M. (2012). Featured graphic: Digital divide: the geography of Internet access. *Environment and Planning A*, *44*(5), 1009 – 1010. doi:10.1068/a44497

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). *Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty* (SSRN Scholarly Paper No. ID 2382617). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2382617

Hann, I.-H., Hui, K.-L., Lee, T. S., & Png, I. P. L. (2003). *The Value of Online Information Privacy: An Empirical Investigation* (Industrial Organization No. 0304001). EconWPA. Retrieved from http://ideas.repec.org/p/wpa/wuwpio/0304001.html

Hsinchun, C., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4). Retrieved from http://ai.arizona.edu/mis510/other/MISQ%20BI%20Special%20Issue%20Introduction%20Chen-Chiang-Storey%20December%202012.pdf

Huang, G. (2012, August 29). MIT Startup Flyberry Capital Emerges with Big-Data Hedge Fund. *Xconomy.com*. Retrieved March 3, 2014, from http://www.xconomy.com/boston/2012/08/29/mit-startup-flyberry-capital-emerges-with-big-data-hedge-fund/

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188. doi:10.1002/asi.21149

Jatin, P., Sunaina, R., & Anupama, S. (n.d.). A Study on Factors affecting the Exposure to Viral Marketing Messages. *International Journal of Management & Business Studies*, 2012.

Jones, B., Temperley, J., & Lima, A. (2009). Corporate reputation in the era of Web 2.0: the case of Primark. *Journal of Marketing Management*, *25*(9-10), 927–939. doi:10.1362/026725709X479309

Konrad, K. A. (2003). *Opinion leaders, influence activities and leadership rents* (Discussion Papers, Research Unit: Market Processes and Governance No. SP II 2003-29). Social Science Research Center Berlin (WZB). Retrieved from http://ideas.repec.org/p/zbw/wzbmpg/spii200329.html

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. META Group. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Larcker, D., Larcker, S., & Tayan, B. (2012). *Monitoring Risks Before They Go Viral: Is it Time for the Board to Embrace Social Media?* (SSRN Scholarly Paper No. ID 2035072). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2035072

Leavitt, A., Burchard, E., Fisher, D., & Gilbert, S. (2009). The Influentials: new approaches for analyzing influence on Twitter. Retrieved March 19, 2013, from http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Trans. Web*, *1*(1). doi:10.1145/1232722.1232727

McKenna, L. (2012). The big idea that can revolutionize higher education:"MOOC." *The Atlantic*, *5*. Retrieved from http://unm2020.unm.edu/knowledgebase/public-versus-private-universities-2020/4-the-big-idea-that-can-revolutionize-higher-education-the-atlantic-12-05-11.pdf

McKinsey & Company. (2014). Open data: Unlocking innovation and performance with liquid information |. Retrieved March 7, 2014, from http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

Nesta. (2014). Nesta. Retrieved March 29, 2014, from http://www.nesta.org.uk/

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance*, *6*(3), 144–160.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. Retrieved from http://cs.wellesley.edu/~cs315/Papers/From%20Tweets%20to%20Polls.pdf

OECD, & Eurostat. (2005). *Oslo Manual*. Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264013100-en

Ouellet, M., Mondoux, A., Ménard, M., Bonenfant, M., & Richert, F. (2013). *Big Data, Gouvernance et Surveillance* (p. 72). Montreal: Université du Québec à Montréal. Retrieved from http://www.cricis.uqam.ca/IMG/pdf/Big_Data-GRICIS_final.pdf

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1-2), 1–135. doi:10.1561/1500000011

Pappano, L. (2012). The Year of the MOOC. *The New York Times*, *2*(12), 2012.

Popper, K. R. (1949). *The open society and its enemies*. London,: Routledge & K. Paul.

Rijnsoever, F. van. (2008). *Opinion leaders in the domain of consumer electronics and their use of external search channels* (Innovation Studies Utrecht (ISU) working paper series No. 08-20). Utrecht University, Department of Innovation Studies. Retrieved from http://ideas.repec.org/p/uis/wpaper/0820.html

Rousseau, P. L. (2008). General Purpose Technologies. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan. Retrieved from http://www.dictionaryofeconomics.com/extract?id=pde2008_G000205

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 513–522). New York, NY, USA: ACM. doi:10.1145/2124295.2124358

Schumpeter, J. A. (1939). *Business cycles; a theoretical, historical, and statistical analysis of the capitalist process* (1st ed.). New York,: McGraw-Hill.

Shapiro, C., & Varian, H. R. (1999). *Economie de l'information - Guide Stratégique de l'économie des réseaux*. Bruxelles: De Boeck.

Taylor, C., & Wagman, L. (2008). *Who Benefits From Online Privacy?* (Working Paper No. 08-26). NET Institute. Retrieved from http://ideas.repec.org/p/net/wpaper/0826.html

Thomas, J., & Sycara, K. (2000). Integrating Genetic Algorithms and Text Learning for Financial Prediction. In *IN PROCEEDINGS OF THE GENETIC AND EVOLUTIONARY COMPUTING 2000 CONFERENCE WORKSHOP ON DATA MINING WITH EVOLUTIONARY ALGORITHMS, LAS VEGAS* (pp. 72–75).

Tocqueville, A. de, & Tocqueville, A. de. (2002). *De la démocratie en Amérique.* Chicoutimi: J.-M. Tremblay. Retrieved from http://classiques.uqac.ca/classiques/De_tocqueville_alexis/democratie_1/democratie_tome1.html

Tucker, C. E. (2012). The economics of advertising and privacy. *International Journal of Industrial Organization*, *30*(3), 326–329.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment (pp. 178–185). Presented at the Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Retrieved from http://scholar.google.de/scholar.bib?q=info:mc319eHjea8J:scholar.google.com/&output=citation&hl=de&as_sdt=0&ct=citation&cd=28

Wang, H., Lee, M. K. O., & Wang, C. (1998). Consumer Privacy Concerns About Internet Marketing. *Commun. ACM*, *41*(3), 63–70. doi:10.1145/272287.272299

Warin, T., & Sanger, W. (2014). Structuring Big Data : How Financial Models may Help. CIRANO.

Warin, T., Sanger, W., & Troadec, A. (2014). Élections au Québec sur Twitter. *CIRANO: Élections*. Retrieved from www.elections.cirano.qc.ca

Wu, Y., Lau, T., Atkin, D. J., & Lin, C. A. (2011). A comparative study of online privacy regulations in the U.S. and China. *Telecommunications Policy*, *35*(7), 603–616.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear.” *Procedia - Social and Behavioral Sciences*, *26*, 55–62. doi:10.1016/j.sbspro.2011.10.562

Quelques-uns des plus récents

Rapports bourgogne publiés par le CIRANO

Éducation et frais de scolarité
Rui Castro, Michel Poitevin

La reputation de votre enterprise ; est-ce que votre actif le plus stratégique est en danger?
Nathalie de Marcellis-Warin, Serban Teodoresco, avril 2012

Corporate Reputation: Is Your Most Strategic Asset at Risk?
Nathalie de Marcellis-Warin, Serban Teodoresco, avril 2012

Les marchés de la finance entrepreneuriale et du capital de risque
Jean-Marc Suret, janvier 2011

Venez voir de quel bois je me chauffe! Portrait d'une industrie en transformation basée sur une ressource renouvelable et écologique
Mathieu Laberge, Pierre Monahan , août 2009

The Euro at 10: Successes and Challenges
Thierry Warin, mai 2009

Extracting Value from Information Technologies and Productivity
Benoit A. Aubert, Blaize Horner Reich, février 2009

La crise financière vue par un banquier
Robert Amzallag, Michel Magnan, Bryan Campbell, janvier 2009

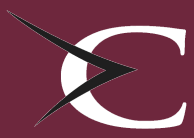
Are We Making a Dragon Out of a Dragonfly? Understanding China's Role in Global Production Networks
Ari Van Assche, janvier 2009

Experimental Economics: A Revolution in Understanding Behaviour
Jim Engle-Warnick, Sonia Laszlo, avril 2008

When and Why Does it Pay to be Green?
Paul Lanoie, Stefan Ambec, Iain Scott, novembre 2007

Des billets verts pour des entreprises vertes?
Paul Lanoie, Stefan Ambec, Iain Scott, novembre 2007

Ces publications sont disponibles sur le site **www.cirano.qc.ca**



1130, rue Sherbrooke ouest, bureau 1400, Montréal (Québec) H3A 2M8

Tél.: 514-985-4000 ● Téléc.: 514-985-4039

www.cirano.qc.ca ● info@cirano.qc.ca