Module 8
Sciences des données et éthiques

Sommaire

Dernière mise à jour le 19 septembre 2018

Introduction

La science des données continue à émerger avec la collecte massive des données dans tous les domaines. L'utilisation de ces données pose cependant beaucoup de défis à savoir les considérations éthiques qui doivent être respectées afin d'assurer la sécurité et la confidentialité des données. Dans ce texte, nous allons présenter les considérations éthiques liées à la science des données et les différentes mesures que nous pouvons prendre pour respecter ces considérations.
Nous sommes tous d'accord sur le fait que la science des données apporte beaucoup d'avantages aux organisations d'une manière générale dans tous les domaines. Cependant, la science des données nous expose aussi à beaucoup de défis liés aux données. L'utilisation inappropriée ou illégale des données pourraient nuire à des personnes et des organisations. À titre d'exemple, l'utilisation illégale des données des utilisateurs de Facebook par la compagnie Cambridge Analytica a entraîner la fermeture de cette grande compagnie en Mai 2018. Beaucoup de compagnies qui détiennent les données des utilisateurs ont connu des accès illégaux aux données, violation ou vol même des données. Par exemple, la compagnie Dropbox a connu un vol des données de 68 680 741 comptes en 2012. Pourquoi l'utilisation illégale des données pourrait nuire aux personnes et aux organisations :

- Les données sont généralement sur des individus (images, messages, vidéos, dossiers médicaux, etc.), donc peuvent causer des dommages.
- Les résultats des algorithmes pourraient être non explicables, ou par fois mal expliqués.
- Beaucoup de données sont collectées accidentellement dans la vie quotidienne des gens :
  - Réseaux sociaux
  - Caméras, capteurs, téléphones, suivi de déplacement, etc.
  - Traitement médical, etc.
- Les données sont maintenant utilisées dans des applications comme :

— Surveillance, militaire, fraude, etc.

— Assurance, crédit, admission scolaire, recrutement, publicité, etc.

— Diagnostique médical, prise de décisions, etc.

Tous ces aspects soulèvent des préoccupations éthiques qui peuvent être résumées dans les points suivants :

— Préserver l'intimité et la vie privée par l'utilisation des méthodes et techniques pour gérer les données sensibles.

— Éviter les biais par la sélection des données utilisables et pertinentes.

— Atténuer les attaques malveillantes par la mise en place de stratégies d'attaques intentionnelles afin d'améliorer les systèmes.

Nous allons commencer d'abord par définir quelques concepts tel que la sécurité, la confidentialité, et l'éthique, puis par la suite, nous aborderons les différentes mesures proposées pour assurer la mise en place de ces concepts dans le cadre des projets liés à la science des données.

8.1 Définitions

Il est important de définir les concepts liés à toutes les considérations éthiques. Ceux-ci incluent la notion de sécurité des données, la confidentialité des données, l'éthique d'une manière générale, l'intimité, etc.

Définition 1 (Sécurité des données). Selon l'office québécois de la langue française [1], la sécurité des données signifie "*l'assurance de la confidentialité et de l'intégrité des données durant leur traitement et leur conservation, grâce à un ensemble de mesures de sécurité. Il s'agit d'assurer la sécurité des données en les protégeant contre la divulgation, le transfert, les modifications ou la destruction non autorisés, que de telles actions soient accidentelles ou intentionnelles*".

1. http://www.oqlf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/Internet/fiches/8874248.html

Selon l'office québécois de la langue française,On confond parfois les termes sécurité des données et protection des données. La sécurité des données fait référence au résultat (le fait que les données soient protégées), tandis que la protection des données a trait aux mesures destinées à assurer cette sécurité.

Définition 2 (Confidentalité des données). Selon le dictionnaire de Cambridge[2], la confidentialité des données est définie comme : "*the practice of making sure that private information is kept secret*". C'est-à-dire, la pratique de s'assurer que l'information privée est gardée secrète. Selon la même source, le cryptage des données est un moyen de s'assurer que la confidentialité des données est maintenue.

Définition 3 (Éthique des données). Selon Luciano Floridi[3], l'éthique des données est défini comme : "*a new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values).*".

La définition de l'éthique des données peut être traduite comme suit : l'éthique des données est une nouvelle branche de l'éthique qui étudie et évalue des problèmes moraux liés au données (génération, enregistrement, conservation, traitement, diffusion, partage et utilisation), aux algorithmes (intelligence artificielle, agents artificiels, apprentissage automatique et robots) et aux pratiques correspondantes (y compris l'innovation responsable, la programmation, le piratage et les codes professionnels), afin de formuler et de soutenir des solutions moralement bonnes (par exemple, des bonnes conduites ou de bonnes valeurs).

2. https://dictionary.cambridge.org/dictionary/english/data-confidentiality

3. Floridi, L. and Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society, 374(2083), 1Ű4

La définition de Luciano Floridi comporte trois types d'éthique décrits comme suit :

— **Éthique des données** : se focalise sur les problèmes éthiques posés par la collecte et l'analyse de grands ensembles de données et sur des questions allant de l'utilisation des big data en recherche biomédicale et en sciences sociales au profilage, en publicité et en philanthropie des données ainsi que des données ouvertes. Dans ce contexte, les questions clés concernent la ré-identification possible des individus par le forage de données, la liaison des données, la fusion et la réutilisation d'ensembles de données volumineux, ainsi que les risques pour ce que l'on appelle la "vie privée de groupe", lorsque les individus, indépendamment de la dés-identification de chacun d'entre eux, peuvent conduire à des problèmes éthiques graves, de la discrimination de groupe (par exemple l'âgisme, l'ethnicisme, le sexisme) aux formes de violence ciblées sur le groupe. La confiance et la transparence sont également des sujets cruciaux dans l'éthique des données, en lien avec un manque de sensibilisation du public aux avantages, opportunités, risques et défis associés à la science des données.

— **Éthique des algorithmes** : aborde les problèmes posés par la complexité croissante et l'autonomie des algorithmes largement compris (par exemple, y compris l'intelligence artificielle et les agents artificiels tels que les robots Internet), en particulier dans le cas des applications d'apprentissage automatique. Dans ce cas, certains défis cruciaux incluent la responsabilité morale et la responsabilité des concepteurs et des scientifiques de données en ce qui concerne les conséquences imprévues et non désirées. Sans surprise, la conception et l'audit éthiques des exigences des algorithmes et l'évaluation des résultats indésirables potentiels (par exemple la discrimination ou la promotion de contenus antisociaux) suscitent de plus en plus de recherches et nécessitent d'être étudiées en profondeur.

— **Éthique des pratiques** : l'éthique des pratiques (incluant l'éthique professionnelle et de déontologie) aborde les questions urgentes concernant les responsabilités et les mandats des personnes et des organisations en charge des processus, stratégies et politiques de données, y compris les scientifiques de données, dans le but de définir

un cadre éthique pour façonner les codes professionnels sur l'innovation responsable, le développement et l'utilisation, qui peuvent assurer des pratiques éthiques favorisant à la fois le progrès de la science des données et la protection des droits des individus et des groupes. Trois questions sont centrales dans cette ligne d'analyse : le consentement, la vie privée de l'utilisateur et l'utilisation secondaire.

8.2 Méthodes et outils pour assurer l'éthique des données

Plusieurs méthodes ont été développées pour permettre d'assurer la sécurité et la confidentialité des données à savoir l'anonymisation et le cryptage. Dans cette section, nous allons présenter ces méthodes et les différents contextes dans lesquels ces méthodes peuvent être appliquées.

8.2.1 Anonymisation des données

Selon Investopedia, l'anonymisation des données est définie comme : "A data privacy technique that seeks to protect private or sensitive data by deleting or encrypting personally identifiable information from a database. Data anonymization is done for the purpose of protecting an individual's or company's private activities while maintaining the integrity of the data gathered and shared". Nous pouvons traduire cette définition comme, une technique de protection des données qui vise à protéger les données privées ou sensibles en supprimant ou en chiffrant les informations personnellement identifiables d'une base de données. L'anonymisation des données est faite dans le but de protéger les activités privées d'un individu ou d'une entreprise tout en maintenant l'intégrité des données recueillies et partagées.
L'anonymisation des données est appliquée dans plusieurs domaines avec différents types de données par exemple des noms et prénoms des individus, dates de naissance, adresses, codes postaux, revenus, etc. À titre d'exemple, si le revenu d'une personne est de 50 000

dollars, alors ce montant pourrait être écrit comme une catégorie de revenu $>$ 40 000 dollars. De cette façon, le revenu de l'individu ne pourra pas être identifié.
Cependant, l'anonymisation des données pourrait, dans certaine situations, ne pas fonctionner. Les données peuvent être fusionnées pour identifier les données anonymisées. Par exemple, 1997, Latanya Sweeney a pu identifier le dossier médical du gouverneur du Massachusetts aux États-Unis. Le gouvernement du Massachusetts a publié des dossiers médicaux anonymes d'un hôpital en supprimant les noms, adresses et numéro d'assurance sociale. Sweeney a utilisé les données de vote qui contiennent des noms, adresses, code postal, date de naissance, et le genre de chaque électeur, pour identifier le dossier médical du gouverneur en se basant sur les données de vote en particulier le code postal, la date de naissance et le genre.
Nous pouvons avoir des situations où les données ne peuvent pas être anonymisées. Par exemple, les séquences du génome sont intrinsèquement identifiables, et donc ne sont pas anonymisables.
Il existe beaucoup de méthodes d'anonymisation des données. Nous pouvons citer à titre d'exemple les trois méthodes les plus connues :

— K-anonymisation

— L-diversité

— Confidentialité différentielle

8.2.2 Cryptage des données

Le cryptage des données, appelé aussi chiffrement des données, consiste à les rendre illisibles. La clé de voûte dans la protection des données aujourd'hui repose sur l'ensemble des techniques de chiffrement de données, que l'on appelle aussi techniques de cryptographie. Donc, le chiffrement est l'opération qui consiste à transformer une donnée qui peut être lue par n'importe qui (donnée dite "claire") en une donnée qui ne peut être lue que par son créateur et son destinataire (donnée dite "chiffrée"). L'opération qui permet de

récupérer la donnée claire à partir de la donnée chiffrée s'appelle le déchiffrement [4].
Le chiffrement se fait généralement à l'aide d'une clé de chiffrement, le déchiffrement nécessite quant à lui aussi une clé de déchiffrement. On distingue deux types de clés :

— **Les clés symétriques** : il s'agit de clés utilisées en même temps pour le chiffrement et le déchiffrement. On parle alors de chiffrement symétrique ou de chiffrement à clé secrète.

— **Les clés asymétriques** : ici les clés utilisées pour le chiffrement et le déchiffrement sont différentes. On parle alors de chiffrement asymétrique ou de chiffrement à clé publique.

Il existe beaucoup d'outils pour le chiffrement des données à savoir :

— Des logiciels utilitaires multiplateformes qui permettent de créer des volumes chiffrés comme VeraCrypt [5].

— Des outils de chiffrement individuel de fichiers comme Axcrypt [6] (Windows), CCrypt [7] (Linux) et MEO [8] (Mac and Windows).

8.3 Partage des données

Le partage des données a une grande valeur scientifique. Ces données peuvent être utilisées par des chercheurs pour développer et tester des solutions aux services des gens. Par exemple, le partage des données des genomes et séquences a aidé grandement les progrès dans ce domaine. De la même façon, le partage des données cliniques dans les grandes bases de données comme PPMI http://www.ppmi-info.org/ sur la maladie du Parkinson, constitue une importante valeur ajoutée à la communauté scientifique. Malgré

4. https://openclassrooms.com/courses/protegez-l-ensemble-de-vos-donnees-sur-votre-ordinateur-1/introduction-a-la-cryptographie
5. https://www.veracrypt.fr/en/Downloads.html
6. https://www.axcrypt.net/fr/
7. http://ccrypt.sourceforge.net/
8. http://www.nchsoftware.com/encrypt/index.html

Le partage des données n'est pas restreint aux communautés scientifiques. Beaucoup d'organisations, de municipalités et gouvernements ont commencé à mettre leurs données disponibles au public. À titre d'exemple, la ville de Montréal a créé un portail pour les données ouvertes http://donnees.ville.montreal.qc.ca/. Ces données pourraient servir au développement de beaucoup d'applications dans différents domaines à savoir la santé, la sécurité, l'environnement, le transport, l'intelligence d'affaires, etc. Rappelons toujours que ces données sont anonymisées avant d'être mises à la disposition des gens.
Un des grands défis de partage des données et l'organisation des données d'une manière présentable, utile, et facile à utiliser. Par exemple, les données doivent toujours être accompagnées de fichiers de description des données pour permettre de comprendre les données, comment elles sont collectées, et comment elles sont organisées afin de faciliter l'utilisation et la réutilisation des données.
Beaucoup d'organismes gouvernementaux encouragent le partage des données mais requièrent que les données doivent respecter certaines normes de présentation comme l'Instituts de recherche en Santé du Canada (IRSC) qui finance même des initiatives qui permettent d'améliorer la qualité, l'accessibilité, le couplage, l'intégration et l'utilité des données[9]. Nonobstant l'encouragement pour le partage des données, certains domaines sont exposés à des contraintes liées à la confidentialité comme les dossiers médicaux des patients dans un hôpital. Les hôpitaux détiennent généralement le contrôle pour ces données et s'opposent à l'idée de partage. Donc, il reste beaucoup de travail à faire pour mettre en place des procédures et des normes pour faciliter le partage des données dans ces domaines. Dans cette perspective, le Canada a lancé un projet nommé Can-SHARE, une initiative canadienne sur le partage international des données http://www.p3g.org/news/can-share-new-initiatives-call-proposals. Selon le site de Can-SHARE, les objectifs de cette initiative sont :

— L'amélioration de la santé humaine par le partage efficace et responsable des données médicales et/ou génomiques ;

— Le développement de lignes directrices ou de cadres de pratique pour le partage des

9. http://www.cihr-irsc.gc.ca/f/50182.html

données cliniques et/ou génomiques ;

— Le développement de normes ou de protocoles pour le partage, la conservation et l'analyse sécuritaire des données génomiques internationales ;

— Le développement de politiques et de directives d'harmonisation afin d'assurer le respect de la vie privée et la confidentialité des données médicales et génomiques.

Un autre aspect important lors de l'utilisation des données partagées est la citation de la source de ces données. Dans la plupart des cas, les organismes qui détiennent les données obligent les utilisateurs de citer la source ou les auteurs qui ont collecté ces données. C'est une façon de reconnaître à la fois la propriété des données, à qui elles appartiennent, et l'organisme responsable pour leur stockage, leur maintenance et leur gestion.

8.4 Science des données et discrimination

La science des données suscite beaucoup de questions sur la discrimination et l'utilisation inappropriée des données. Il est illégal de faire des choix basés sur la race, le genre, la religion, l'origine, etc. Notez bien que les données ne font pas la discrimination, mais la façon dont ces données sont utilisées pourrait créer et renforcer la discrimination. Ce genre de problèmes pourrait survenir lorsque des modèles sont utilisés pour faire des prédictions dans des domaines tels que l'assurance, les prêts financiers, et la sécurité. Par exemple, si les membres d'une certaine communauté raciale ont historiquement été plus susceptibles de faire défaut sur leurs prêts, ou ont été plus susceptibles d'être reconnus coupables d'un crime, alors le modèle peut juger ces personnes plus risquées. Cela ne signifie pas nécessairement que ces membres adoptent un comportement criminel ou ont plus mal à gérer leur argent. Ils peuvent simplement être ciblés de manière disproportionnée par des organismes financiers (assurances, banques, etc) ou des organismes de sécurité comme la police par exemple.
Pour éviter que de telles situations se produisent, une façon plus simple de procéder est d'omettre les attributs sensibles qui peuvent identifier ces personnes comme la race, la

religion, le genre, etc. De cette façon, nous pouvons écarter toute possibilité de discrimination lors de l'analyse des données.
Un point important que les scientifiques de données doivent prendre en compte est l'internet ouvert. Par exemple, Microsoft a développé un agent conversationnel (*chatbot*) appelé **Tay** pour interagir avec les utilisateurs sur Twitter. Tay est développé en utilisant des données publiques pertinentes, en combinaison avec des données provenant de l'équipe d'édition. L'objectif était que Tay apprenne et s'améliore au fur et à mesure qu'il interagit avec les utilisateurs. Dans les 24 heures suivant son dévoilement, Tay a fini par faire beaucoup de discrimination, de racisme, d'insultes, etc. Cela indique les types d'interactions que Tay a eu avec certains utilisateurs. Les algorithmes d'apprentissage de Tay ne faisaient malheureusement pas de distinctions dans les propos inappropriés des utilisateurs.

8.5 Règles simples pour une science des données responsable

Nous concluons ce texte par quelques règles simples pour une science des données responsable. Ces règles sont inspirées de l'article de Matthew Zook et ses collègues publié dans la revue PLoS Computer Biology [10]. Ces règles peuvent être résumées dans les points suivants (traduction française des règles) :

1. Reconnaître que les données représentent des informations sur des personnes et peuvent faire du mal ;
2. Reconnaître que la vie privée est plus qu'une valeur binaire ;
3. Protégez-vous contre la réidentification de vos données ;
4. Pratiquer l'éthique de partage des données ;

10. Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3) : e1005399. https ://-doi.org/10.1371/journal.pcbi.1005399

5. Tenir compte des forces et des limites de vos données ; massives (*big*) ne signifient pas automatiquement mieux ;
6. Débattre des choix difficiles et éthiques ;
7. Développer un code de conduite pour votre organisation, votre communauté de recherche ou votre industrie ;
8. Concevoir vos données et vos systèmes pour la vérification ;
9. S'engager avec les conséquences plus larges des données et des pratiques d'analyse ;
10. Savoir quand casser ces règles.