4. L'éthique de la science des données

La nécessité de l'éthique

Dans la plupart des disciplines empiriques, l'**éthique** est introduite tôt dans le processus éducatif et finit par jouer un rôle crucial dans les activités des chercheurs.

Les scientifiques des données qui arrivent dans le domaine par le biais des mathématiques, des statistiques, de l'informatique, de l'économie, ou de l'ingénierie sont toutefois moins susceptibles d'avoir rencontré des comités de recherche éthique ou une **formation formelle en éthique**.

Les discussions sur les questions d'éthique sont souvent **mises de côté** au profit de considérations techniques ou administratives urgentes lorsque les délais sont serrés.

Mais cette échéance est remplacée par une autre échéance, puis par une autre, et ainsi de suite, le résultat final étant que la conversation **peut ne jamais avoir lieu**.

La nécessité de l'éthique

Lorsque la collecte de données à grande échelle devient possible, elle est accompagné d'une mentalité "Far West" : **tout est permis tant que faisable**.

La science des données moderne a des **codes de conduite professionnels**

- décrivant des façons **responsables** de pratiquer la science des données
- légitime plutôt que frauduleuse, éthique plutôt que contraire à l'éthique

Cela confère une **responsabilité supplémentaire** aux scientifiques des données, mais offre une **protection** contre les clients/employeurs qui veulent qu'ils effectuent des analyses de manière douteuse.

La nécessité de l'éthique

L'accent mis sur l'éthique des données récemment ne semble pourtant pas avoir ralenti les brèches :

- Volkswagen
- Whole Foods Markets
- General Motors
- Cambridge Analytica
- Amazon
- Ashley Madison

Qu'est-ce que l'éthique ?

L'éthique fait référence à l'étude et à la définition des **bonnes** et des **mauvaises** conduites :

- en général
- appliqué dans des circonstances spécifiques

L'éthique n'est pas (nécessairement) la même chose que :

- convention sociale
- convictions religieuses
- lois

Qu'est-ce que l'éthique ?

En Occident, les théories éthiques sont utilisées pour encadrer les débats autour des questions éthiques :

- **règle d'or :** faites aux autres ce que vous voudriez qu'ils vous fassent ;
- **conséquentialisme :** la fin justifie les moyens ;
- **utilitarisme :** agir de manière à maximiser l'effet positif ;
- **droits moraux :** agir pour maintenir et protéger les droits et privilèges fondamentaux des personnes affectées par les actions ;
- **justice :** répartir les avantages et les préjudices entre les parties prenantes de manière juste, équitable et impartiale.

Qu'est-ce que l'éthique ?

Il y a une grande variété de codes/cultures éthiques, notamment :

- Confucianisme
- Taoïsme
- Bouddhisme
- Ubuntu
- Te Ara Tika (Maori)
- etc.

Il est facile d'imaginer des contextes dans lesquels l'un de ces éléments serait mieux adapté à la tâche à accomplir – **renseignez-vous**.

L’éthique et science des données

Comment ces théories éthiques peuvent-elles s'appliquer à l'analyse des données ?

- qui, le cas échéant, est **propriétaire des données** ?
- y a-t-il des **limites** à l'utilisation des données ?
- certaines analyses comportent-elles des **biais de valeur** ?
- y a-t-il des catégories qui ne devraient jamais être utilisées dans l'**analyse des données personnelles** ?
- les données doivent-elles être accessibles **publiquement** ?

Les réponses dépendent d'un certain nombre de facteurs. Pour vous donner une idée de certaines des complexités, posons la première question : *qui, le cas échéant, est propriétaire des données ?*

L’éthique et science des données

Est-ce que ce sont les **analystes de données** qui transforment le potentiel des données en informations exploitables ?

Est-ce que ce sont les **collecteurs de données** qui ont une copie et rendent le travail possible ?

Sont-ce les **commenditaires** ou les **employeurs** qui ont rendu le processus viable ?

Dans certains cas, la **loi** peut également intervenir.

Il n'est pas facile de répondre à cette question simple ; il faut s’y prendre au cas par cas.

Vérité cachée : **l'analyse des données ne se limite pas** à l'***analyse* des données**.

L’éthique et science des données

Défi similaire pour les **données ouvertes** (les "pro" et les "anti" ont de solides arguments).

Principe général de l'analyse des données : éviter l'**anecdotique** pour le **general** (se concentrer sur des observations spécifiques peut masquer la vue d'ensemble).

Mais les données **ne sont pas seulement** des marques sur le papier ou des octets sur le “cloud”. Les décisions prises sur la base de la science des données peuvent **affecter des gens/la planète de manière négative**. On ne peut ignorer que les individus périphériques et les groupes minoritaires souffrent souvent de manière disproportionnée aux mains des décisions dites “fondées sur l’evidence”.

Principes de **PCAP** (propriété, contrôle, accès, possession) des Premières Nations.

Les meilleures pratiques

"Ne faites pas de tort" : les données recueillies auprès d'un individu **ne doivent pas être utilisées pour lui nuire**.

Consentement éclairé :

- les individus doivent **accepter la collecte et l'utilisation** de leurs données
- les individus doivent avoir une **réelle compréhension de ce à quoi ils consentent**, et des **conséquences possibles** pour eux et pour les autres.

Respecter la "vie privée" : excessivement difficile à maintenir à l'ère du “scraping” constant de l'Internet pour recueillir des données personnelles.

Meilleures pratiques

Les données doivent être gardées **publiques** (toutes ? la plupart ?).

Opt-In/Opt-Out : le consentement éclairé exige la possibilité de **se désengager**

Anonymiser les données : suppression des champs d'identification des données avant l'analyse.

"Laissez parler les données" :

- pas de sélection à la carte
- l'importance de la validation
- corrélation vs. causalité
- répétabilité

Le bon, la brute, et le truand

Les projets de données pourraient être classés de façon fantaisiste comme **bons**, **mauvais,** ou encore **laids**, soit d'un point de vue technique, soit d'un point de vue éthique (ou les deux).

- les **bons** projets accroissent les connaissances, peuvent aider à découvrir des liens cachés, etc., de la manière la plus inoffensive possible
- les **mauvais** projets peuvent conduire à de mauvaises décisions, qui peuvent à leur tour diminuer la confiance du public et potentiellement nuire à certains individus
- les projets **moches** sont, carrément, des applications peu recommandables ; ils sont mal exécutés d'un point de vue technique, ou mettent beaucoup de personnes en danger ; ces projets (et les approches/études similaires) doivent être évités **à tout prix** !

Le bon, la brute, et le truand

Bons projets (?) :

- P. A. B. Bien Nicholas AND Rajpurkar, “Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet,” *PLOS Medicine*, vol. 15, no. 11, pp. 1–19, 2018, doi: 10.1371/journal.pmed.1002699.
- BeauHD, “Google AI claims 99 percent accuracy in metastatic breast cancer detection,” *Slashdot.com*, Oct. 2018.
- Columbia University Irving Medical Center, “Data scientists find connections between birth month and health,” *Newswire.com*, Jun. 2015.

Le bon, la brute, et le truand

Mauvais projets (?) :

- Indiana University, "Scientists use Instagram data to forecast top models at New York Fashion Week," *Science Daily*, Sep. 2015.
- D. Wakabayashi, "Firm led by Google veterans uses A.I. to 'nudge' workers toward happiness," *New York Times*, Dec. 2018.
- N. Cohn, "How one 19-year-old illinois man is distorting national polling averages," *The Upshot*, 2016.

Le bon, la brute et le truand

Projets moches (?) :

- J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 2018.
- I. Johnston, "AI robots learning racism, sexism and other prejudices from humans, study finds," *The Independent*, Apr. 2017.
- M. Judge, "Facial-recognition technology affects African-Americans more often," *The Root*, 2016.
- M. Kosinski and Y. Wang, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of Personality and Social Psychology*, vol. 114, no. 2, pp. 246–257, Feb. 2018.

Lectures suggérées

L’ éthique de la science des données

Data Understanding, Data Analysis, Data Science
Data Science Basics

Ethics in the Data Science Context

- The Need for Ethics
- What Is/Are Ethics?
- Ethics and Data Science
- Guiding Principles

Exercices

L’ éthique de la science des données

1. Faites une recherche sur les récents scandales d'éthique des données impliquant Volkswagen, Amazon, Whole Foods Markets, Cambridge Analytica, Ashley Madison, General Motors ou toute autre organisation. Que s'est-il passé ? Qui a été affecté ? Quelles ont été les conséquences pour le grand public, l'organisation, la communauté des données ? Comment cela aurait-il pu être évité ?

2. Établissez une déclaration d'éthique pour votre travail sur les données. Y a-t-il des domaines sur lesquels vous n’acceptez pas de travailler ?