Moderniser la formation en sciences sociales :

Sciences des données et sciences sociales

Etienne Gagnon

etienne.gagnon4@mail.mcgill.ca

McGill University Political Science

Les départements canadiens de sciences sociales ne parviennent pas à fournir à leurs étudiants les compétences requises pour réussir sur le marché du travail universitaire et professionnel actuel. Les diplômés en sciences sociales peuvent jouer un rôle clé dans l'économie, la vie politique et la fonction publique canadiennes. Les vastes connaissances et les compétences analytiques des diplômés en sciences sociales leur permettent de rendre un hommage positif à presque toutes les organisations du secteur public ou privé. En même temps, ils sont souvent confrontés à un manque de compétences pratiques en matière d'analyse quantitative des données, ce qui les empêche d'appliquer pleinement leurs connaissances. Plus de 25 % des titulaires d'un baccalauréat en sciences sociales sont considérés comme surqualifiés pour leur poste [Statistique Canada, 2017a] et leurs gains sont inférieurs à la médiane canadienne des diplômés universitaires [Statistique Canada, 2017b].

La formation en sciences des données offre une solution parfaite à ce problème. La science des données est un domaine interdisciplinaire qui s'articule autour de la collecte et de l'analyse appropriée des données, en utilisant des techniques statistiques et informatiques. Il peut donner aux étudiants en sciences sociales un ensemble de compétences pratiques très pertinentes à la fois pour leur bourse d'études et pour des postes futurs dans la fonction publique ou dans le secteur privé. Inspiré d'une initiative japonaise similaire, cet essai développe une proposition politique pour développer une stratégie nationale cohérente visant à améliorer la formation en sciences des données dans les départements de sciences sociales.

1 Sciences des données et sciences sociales

Les compétences en sciences des données sont déjà largement utilisées dans les départements de sciences sociales à travers le pays. L'utilisation des

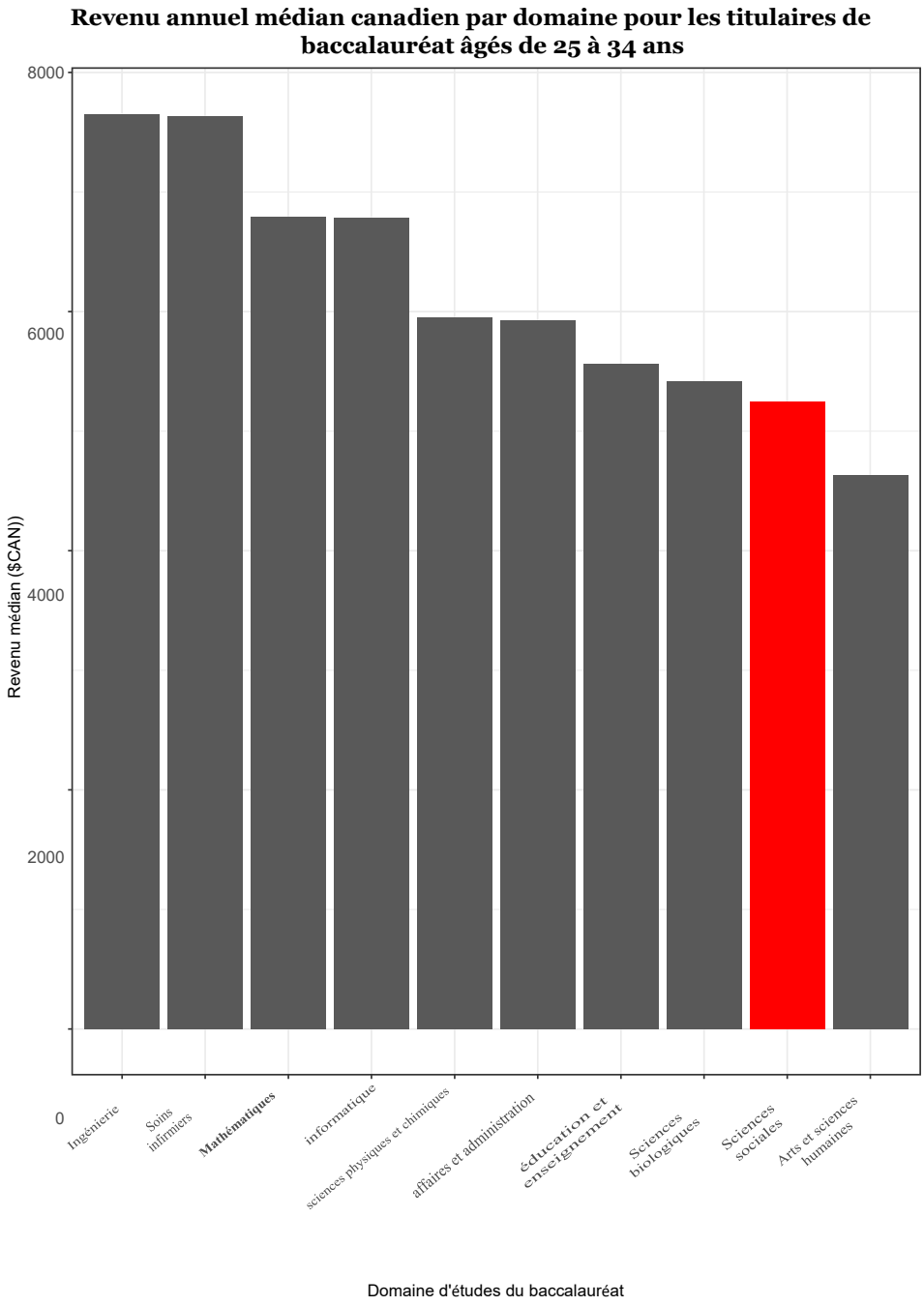

Figure 1: Source: Social Science graduates' income is less than most other field's graduates [Statistics Canada, 2017b]

Taux de surqualification par domaine pour les titulaires de baccalauréat âgés de 25 à 34 ans

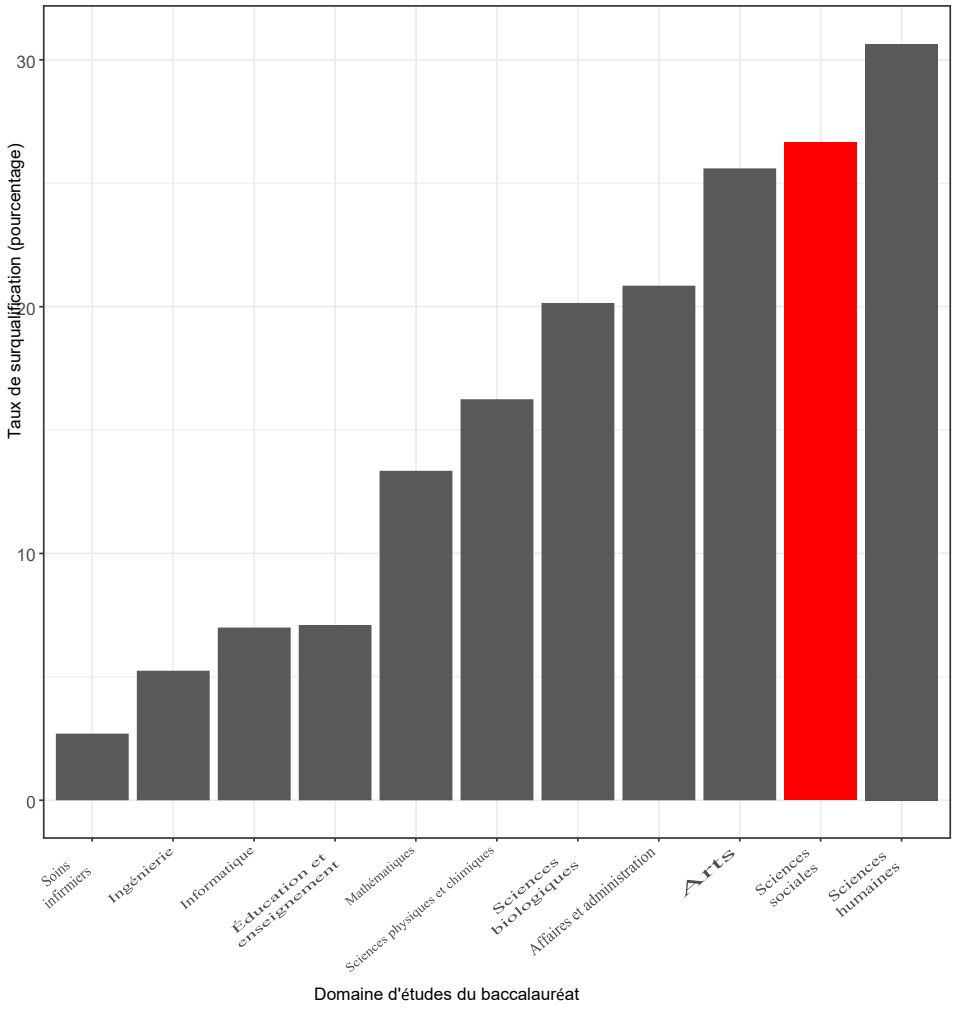

Figure 2: Social Science graduates' over qualification rate is higher than other fields' [Statistics Canada, 2017a]

techniques de la science des données en sciences sociales offre des récompenses considérables aux chercheurs. L'analyse de grandes données permet aux chercheurs d'accéder à des sources de données de plus en plus granulaires et d'étudier des questions jusque-là impossibles à étudier [Shaked, 2015]. Un exemple de ceci est l'analyse des données des médias sociaux pour comprendre les comportements sociaux, comme la polarisation politique [Gruzd, 2014]. Malgré cela, l'acquisition de ces compétences est extrêmement difficile dans presque tous les départements de sciences sociales du pays. L'enseignement est souvent dispensé de manière non systématique, à l'aide de logiciels obsolètes, et racle rarement la surface de sujets complexes. Le manque de connaissances mathématiques des étudiants conduit les professeurs à éviter de discuter de sujets complexes, ce qui conduit à des idées fausses persistantes sur les statistiques [Wasserstein et Lazar, 2016]. Ceux qui souhaitent utiliser les méthodes de la science des données doivent s'appuyer sur l'auto-apprentissage, suivre des cours d'été d'un coût prohibitif dans les universités américaines ou suivre des cours dispensés par des départements d'informatique ou de mathématiques, où la courbe d'apprentissage est extrêmement rapide et le contenu non destiné à des applications en sciences sociales.

2 Proposition de politique

Il est clair qu'il est nécessaire d'avoir une stratégie nationale cohérente pour améliorer l'enseignement des sciences des données pour les chercheurs en sciences sociales. Cette proposition politique s'inspire d'une initiative similaire mise en place par le ministère japonais de l'Éducation en 2015, et étendue en 2019 grâce à son succès remarquable [ministère japonais de l'Éducation et de la Technologie, 2019]. Étant donné que l'éducation est une compétence provinciale, les politiques nationales peuvent être difficiles à mettre en œuvre

au Canada. L'administration fédérale peut toutefois créer un institut de recherche national et octroyer un financement aux universités. Cette proposition de politique propose la création d'un institut de recherche interuniversitaire, le Consortium des sciences des données sociales, où les écoles désignées pourront se joindre et recevoir des fonds pour offrir une formation en sciences des données destinée aux étudiants en sciences sociales sur une base interministérielle.

2.1 Le Consortium des sciences des données sociales

Six universités doivent être choisies pour recevoir des fonds du gouvernement canadien et former le "Social Data Science Consortium". Le choix de ces universités doit tenir compte des ressources immédiatement disponibles en IA et en sciences des données, de la nécessité d'offrir cette formation dans toutes les grandes régions canadiennes et de la nécessité de l'offrir dans les deux langues officielles. Sur cette base, l'Université de Toronto, l'Université McGill, l'Université de la Colombie-Britannique, l'Université de l'Alberta, l'Université de Montréal et l'Université Dalhousie sont proposées comme candidats pour former le Consortium.

Les écoles du Consortium recevront des fonds pour élaborer une série de cours de sciences des données offerts sur une base interministérielle aux étudiants en sciences sociales. Les écoles membres du consortium doivent modifier la structure de leurs programmes de sciences sociales afin d'exiger qu'au moins 6 crédits soient suivis dans le programme de sciences sociales par les étudiants en dernière année, 12 crédits pour les étudiants en honneur et jusqu'à 24 crédits par les étudiants les plus enthousiastes. Bien que chaque université doit disposer d'une marge de manœuvre suffisante pour élaborer les cours qui répondent le mieux aux besoins de ses étudiants, certains préceptes généraux doivent être adoptés. L'un d'entre eux est que le contenu et les problèmes des cours doivent être enseignés d'une manière qui soit en rapport étroit avec les applications des sciences sociales de la matière et qui leur

rappelle constamment. Il est important de maintenir un haut niveau d'engagement de la part des étudiants en sciences sociales, qui souvent ne sont pas intrinsèquement intéressés par des sujets tels que la programmation, les statistiques et les mathématiques. L'idée derrière cette proposition n'est pas de dévaloriser l'importance de la formation en sciences sociales par opposition aux compétences en sciences des données, mais plutôt de souligner la synergie entre les deux. Les compétences de base suivantes en sciences des données sociales doivent être offertes :

- Programmation statistique : Les étudiants doivent être initiés à la programmation statistique à l'aide des langages de recherche actuels et des langages pertinents pour l'industrie, tels que Python ou R, qui offrent la flexibilité nécessaire pour appliquer de grandes données ou des méthodes IA. Ils ne doivent explicitement pas être enseignés à l'aide de logiciels statistiques tels que STATA ou SPSS, qui sont plus faciles à apprendre mais de plus en plus obsédés par des langages de programmation plus flexibles et puissants. Une partie du cours devrait porter sur l'élaboration de bonnes pratiques de programmation en matière de reproductibilité, un problème commun à la recherche en sciences sociales [Freese et Peterson, 2017]. Les techniques d'optimisation de code nécessaires à l'analyse de grandes données doivent également être enseignées.

- Principes de base des mathématiques : Les étudiants en sciences sociales apprennent généralement à utiliser les statistiques de manière appliquée, sans avoir une bonne compréhension des opérations mathématiques qui sous-tendent les techniques statistiques. Ceci provoque souvent des interprétations erronées très fondamentales des modèles statistiques et les rend incapables d'appliquer correctement des méthodes statistiques plus avancées [Good et Hardin, 2012]. Les cours de réintroduction aux concepts mathématiques clés de la science des données, comme l'algèbre linéaire ou le calcul, et la façon dont ils s'appliquent aux modèles statistiques communs,

doivent faire partie du programme. Il faut également transmettre aux étudiants de solides connaissances fondamentales en théorie des probabilités et en statistique.

- Apprentissage machine appliqué : L'apprentissage automatique permet d'ouvrir de nouveaux domaines de données qui, autrement, ne seraient pas disponibles pour les chercheurs en sciences sociales. Des techniques comme le traitement du langage naturel (TLN) donnent aux spécialistes des sciences sociales les outils nécessaires pour analyser d'énormes corpus de données médiatiques, de discours politiques ou d'autres objets textuels [Grimmer, 2014]. La capacité de tirer parti de ces techniques est un élément clé de l'ensemble des compétences des spécialistes des données sociales.

- Visualisation des données : Les chercheurs en sciences sociales doivent maîtriser des techniques efficaces de visualisation des données, que ce soit dans la présentation de leurs travaux universitaires ou lors de la présentation de données à des clients hors du milieu universitaire [Manovich, 2011]. Il convient d'enseigner la théorie de la visualisation efficace des données ainsi que les connaissances en programmation nécessaires à la production de graphiques efficaces.

Les écoles du consortium devraient permettre aux étudiants des écoles non membres du consortium de leur région de s'inscrire dans ces classes dans la mesure du possible, afin de s'assurer que ces connaissances sont mises à la disposition de presque tous les étudiants en sciences sociales du pays, quelle que soit leur université.

Des laboratoires interdisciplinaires dédiés aux méthodes de recherche en sciences des données sociales seront également mis en place et encourageront les étudiants diplômés à appliquer ces méthodes à leur recherche. Grâce à cette initiative, les universités membres du consortium pourront former des étudiants diplômés hautement qualifiés et dotés des outils nécessaires pour réussir sur un marché du travail universitaire de plus en plus concurrentiel [Schillebeeckx et al., 2013].

L'objectif principal est de concevoir un programme d'études et une façon d'enseigner les sciences des données aux étudiants en sciences sociales qui leur donne ces compétences tout en les maintenant engagés dans le contenu.

3 Conclusion

Malgré les préoccupations de longue date en matière d'employabilité des diplômés en sciences sociales, le Canada a été remarquablement lent à adopter la grande révolution des données comme moyen de doter ses spécialistes en sciences sociales de compétences pratiques pertinentes à leur recherche. Le marché du travail universitaire étant sursaturé d'étudiants au doctorat [Nature Editorial, 2017], il est important de donner aux étudiants en sciences sociales des compétences qui leur permettront également de réussir dans le secteur privé s'ils ne décrochent pas un emploi universitaire. En mettant l'accent sur l'interaction et la synergie entre les méthodes des sciences des données et la recherche en sciences sociales, cette initiative stratégique propose un plan pour renverser la situation actuelle et créer un bassin d'étudiants en sciences sociales bien formés en sciences sociales et en sciences des données, en quelques années.

Bibliographie

[Freese and Peterson, 2017] Freese, J. and Peterson, D. (2017). Replication in Social Science. *Annual review of Sociology*.

[Good and Hardin, 2012] Good, P. I. and Hardin, J. W. (2012). *Common Errors in Statistics (and How to Avoid Them)*. John Wiley & Sons.

[Grimmer, 2014] Grimmer, J. (2014). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS - Political Science and Politics*, 48(1):80–83.

[Gruzd, 2014] Gruzd, A. (2014). Investigating Political Polarization on Twitter: A Canadian Perspective. *Policy and Internet*, 6(1):28–45.

[Japanese Ministry of Education and Technology, 2019] Japanese Ministry of Education, Culture, S. S. and Technology (2019). daigakuniokerusu-urridetasaiensukyouikunozenkokutenkainokyouryokukounosenteinitsuite. Regarding the selection of schools cooperating with the national expansion of Data Science and Mathematics. Technical report.

[Manovich, 2011] Manovich, L. (2011). Trending: The promises and the chal- lenges of big social data. In *Debates in the digital humanities 2*, pages 460–475.

[Nature Editorial, 2017] Nature Editorial (2017). Many junior scientists need to take a hard look at their job prospects.

[Schillebeeckx et al., 2013] Schillebeeckx, M., Maricque, B., and Lewis, C. (2013). The missing piece to changing the university culture. *Nature Biotechnology*, 31(10):938–941.

[Shaked, 2015] Shaked, N. (2015). Social Science in the Era of Big Data. *Social Technology Magazine*, 20(3):6.

[Statistics Canada, 2017a] Statistics Canada (2017a). Are young bachelor's de- gree holders finding jobs that match their studies? Technical report.

[Statistics Canada, 2017b] Statistics Canada (2017b). Is field of study a factor in the earnings of young bachelor's degree holders? Technical report.

[Wasserstein and Lazar, 2016] Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133.