Modernizing Social Science Training:

Data Science and the Social Sciences

Etienne Gagnon

etienne.gagnon4@mail.mcgill.ca

McGill University Political Science

Word Count: 1480 Words

Canadian social science departments are failing to provide their students with the required skills to succeed in the current academic and professional job market. Graduates from Social Science fields have the potential to play a key role in the Canadian economy, political life and public service. Social science graduates' vast knowledge and analytical skills give them the potential to contribute positively to almost any organization, in the public or private sector. At the same time, they are often plagued by a lack of practical skills in quantitative data analysis that makes it hard for them to apply their knowledge to its' full potential. Over 25% of social science bachelors' degree holders are considered to be overqualified for their position[1] [Statistics Canada, 2017a] and their earnings are under the Canadian median for University graduates [Statistics Canada, 2017b].

Data Science training offers a perfect solution to this problem. Data science is an interdisciplinary field that centers around the gathering and proper analysis of data, using techniques from statistics and computer science. It can give social science students a practical skillset highly relevant to both their scholarship and to future positions in public service or the private sector. Inspired by a similar Japanese initiative, this essay develops a policy proposition to develop a coherent national strategy aiming to enhance data science formation in social science departments.

1 Data Science and Social Science

Data science skills are already widely used in social science departments across the country. The use of data science techniques in social science offers considerable rewards for researchers. Big data analysis allows researchers to access

[1]Meaning that they occupy a position that requires a high school diploma as opposed to their university level diploma

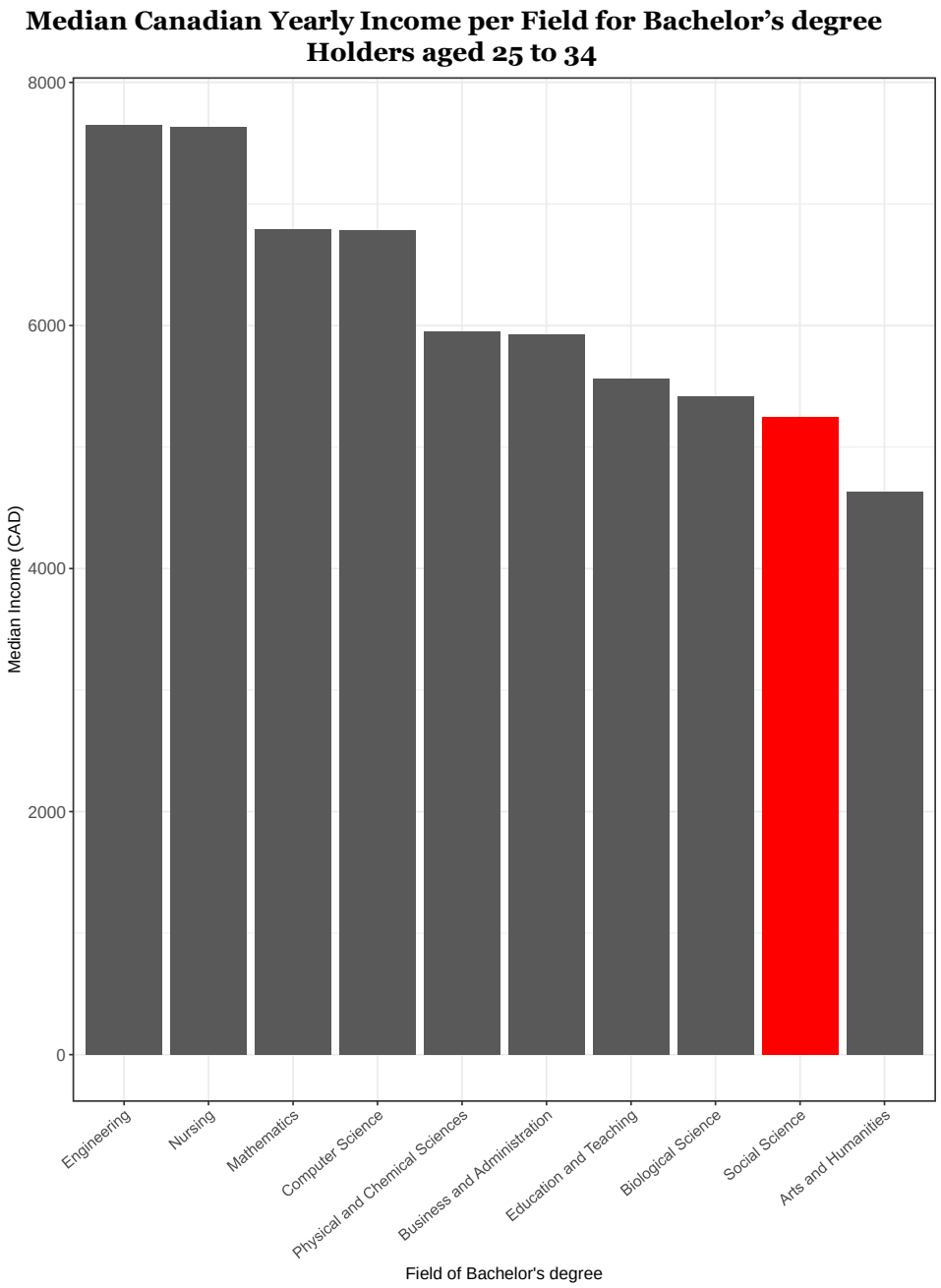

Figure 1: Source: Social Science graduates' income is less than most other field's graduates[Statistics Canada, 2017b]

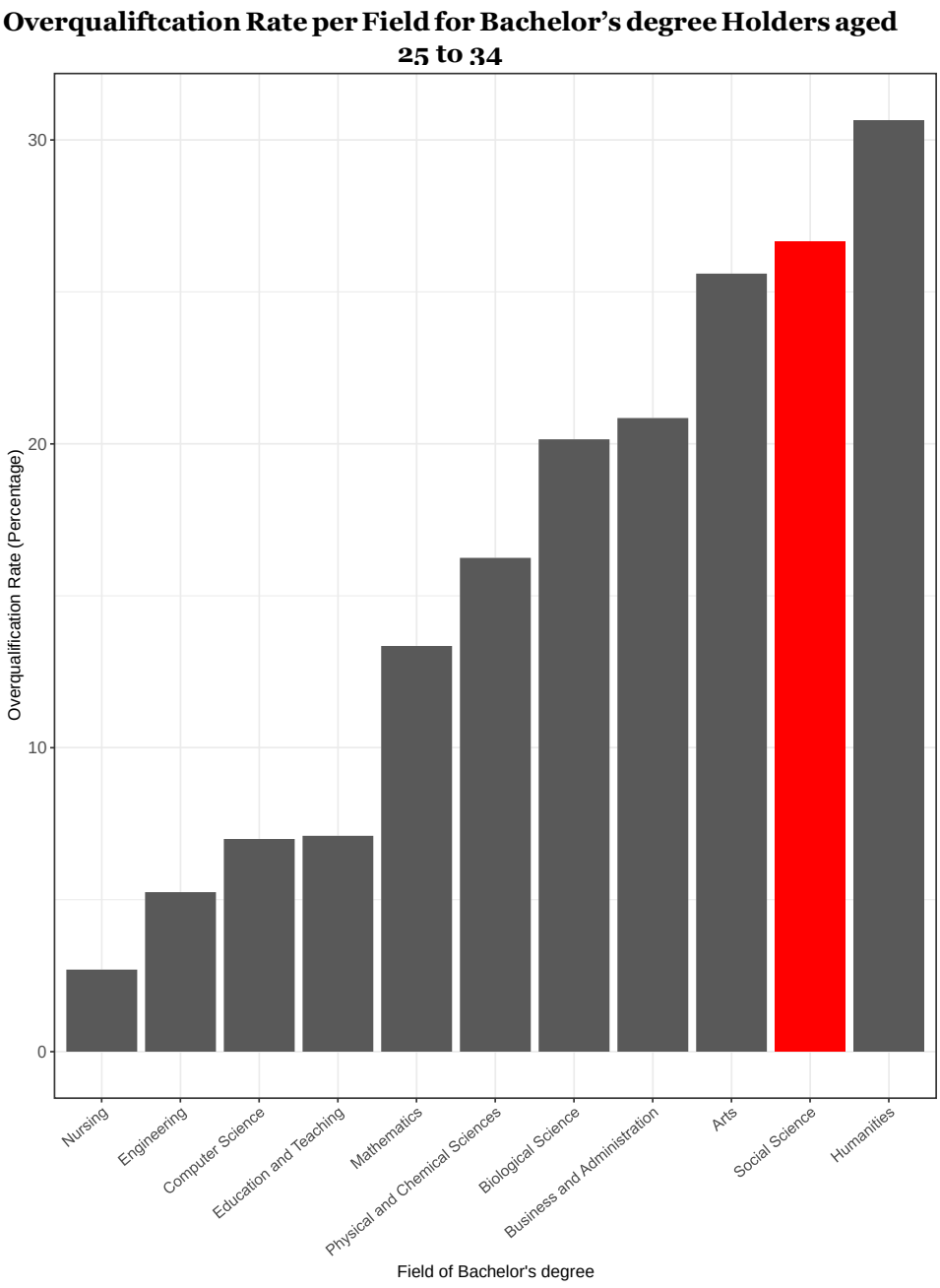

Figure 2: Social Science graduates' overqualification rate is higher than other fields' [Statistics Canada, 2017a]

ever more granular sources of data and to investigate previously impossible to research questions [Shaked, 2015]. An example of this is the analysis of social media data to understand social behaviors, such as political polarization [Gruzd, 2014]. In spite of this, acquiring these skills is extremely difficult in almost every social science department across the country. Teaching is often carried out unsystematically, through the use of obsolete software, and rarely scrapes past the surface of complex topics. Students' lack of mathematical knowledge leads to Professors avoiding to discuss complex topics, which leads to long lasting misconceptions regarding statistics [Wasserstein and Lazar, 2016]. Those wishing to pursue data science methods have to rely on self-studying, attending prohibitively expensive summer schools in American universities, or on taking classes given by computer or mathematics departments, where the learning curve is extremely steep and the content not intended for social science applications.

2 Policy Proposal

There is a clear need to have a coherent national strategy to improve data science teaching for social scientists. This policy proposal takes inspiration from a similar initiative put in place by the Japanese ministry of Education in 2015, and expanded in 2019 thanks to its' remarkable success [Japanese Ministry of Education and Technology, 2019]. Due to education being a provincial jurisdiction, nationwide policies can be hard to implement in Canada. The federal administration can however establish a nationwide research institute and provide funding to Universities. This policy proposal suggests the formation of an inter-university research institute, the Social Data Science Consortium, where designated schools will be allowed to join and receive funding to offer data science training intended for social science students on an inter-departmental basis.

2.1 The Social Data Science Consortium

Six universities are to be chosen to receive funding from the Canadian government and form the ”Social Data Science Consortium”. The choice of these Universities needs to take into consideration their immediately available AI and data science resources, the necessity to offer this training in all major Canadian regions, and the need to offer it in both official languages. On that basis, the University of Toronto, McGill University, The University of British Columbia, The University of Alberta, l'Université de Montréal and Dalhousie University are suggested as candidates to form the Consortium.

The Consortium schools will receive funding to develop a series of data science courses offered on an inter-departmental basis to social science students. Consortium schools are to modify the structure of their social science programs to require at least 6 credits to be taken in the social data science program by students in their senior year, 12 credits for honor students, and allow up to 24 credits to be taken by the most eager students. While every University should be given enough leeway to develop coursework that best fits the needs of its' students, some general precepts are to be espoused. One of them is that the courses' content and problems need to be taught in a way that relates closely to, and constantly reminds of, the social science applications of the material. This is important to keep a high level of engagement from social science students, who are often not intrinsically interested in topics such as programming, statistics and mathematics. The idea behind this proposal is not to devalue the importance of social science training as opposed to data science skills but rather to emphasize the synergy between both. The following core skills to social data science need to be offered:

- **Statistical Programming:** Students are to be introduced to statistical programming using current research and industry relevant languages, such

as Python or R, that offer the necessary flexibility to apply big data or AI methods. They are explicitly not to be taught using statistical packages such as STATA or SPSS, which are easier to learn but increasingly obsoleted by more flexible and powerful programming languages. Part of the coursework should focus on developing good programming practices with regards to reproducibility, a common problem for social science research [Freese and Peterson, 2017]. Code optimization techniques necessary to big data analysis must also be taught.

- **Mathematics Fundamentals:** Students in Social Science are generally taught to use statistics in an applied manner, without being given a proper understanding of the mathematical operations behind statistical techniques. This often causes very fundamental misinterpretations of statistical models and makes them unable to apply more advanced statistical methods properly [Good and Hardin, 2012]. Classes re-introducing students to mathematical concepts key to data science, such as linear algebra or calculus, and how they apply in common statistical models, need to be part of the curriculum. Strong fundamentals in probability theory and statistics also need to be transmitted to students.

- **Applied Machine Learning:** Machine learning skills allow to open up new realms of data otherwise unavailable for social scientists. Techniques like Natural Language Processing (NLP) give social scientists the tools to analyze huge corpuses of media data, political speeches or other textual objects [Grimmer, 2014]. Being able to leverage these techniques is a key part of the skillset of social data scientists.

- **Data Visualization:** Social scientists need to master effective data visualization techniques, both in the presentation of their academic papers or when presenting data to clients out of academia [Manovich, 2011]. Theo-

ries of effective data visualization alongside the programming knowledge required for producing effective plots should be taught.

Consortium schools should allow students from non-consortium schools in their region to enroll in these classes as much as possible, so as to make sure that this knowledge is made available to almost every social science students in the country, regardless of their University.

Cross-disciplinary laboratories dedicated to social data science research methods will also be put into place and encourage graduate students to apply these methods to their research. With this initiative, consortium universities will be able to graduate highly trained graduate students equipped with the tools to succeed on an increasingly competitive academic job market [Schillebeeckx et al., 2013].

The main objective is to conceive a curriculum and a way of teaching data science for social science students that gives them these skills while keeping them engaged in the content.

3 Conclusion

In spite of long standing employability concerns for social science graduates, Canada has been remarkably slow to embrace the big data revolution as a way to equip its' social scientists with practical skills relevant to their scholarship. With the academic job market being over-saturated with PhD students [Nature Editorial, 2017], it is important to give social science students skills that also allow them to succeed in the private sector if they fail to land an academic job. By empathizing the interplay and synergy between data science methods and social science research, this policy initiative proposes a plan to turn the current situation around and create a pipeline of social science students well trained in both data and social science, within a few years.

References

[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). daigakuniokerusuurridetasaiensukyouikunozenkokutenkainokyouryokukounosenteinitsuite. 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 challenges 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 degree 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.