Toward understanding the COVID-19 impact on Data Science Innovation in Canada

Michelle Alexopoulos
m.alexopoulos@utoronto.ca
University of Toronto
Canada

Kelly Lyons
kelly.lyons@utoronto.ca
University of Toronto
Canada

ABSTRACT

Data science provides methods to understand and solve problems in an evidence-based manner by combining data and experience with scientific methods. When included with advances in robotic technologies, telecommunication technologies and internet coverage (digitally enabling infrastructure), and computer hardware, data science creates a host of "digital technologies" expected to bring value to business, government, and individuals. Recent adoption and development of these innovations have been affected by the pandemic, even as COVID-19 sped up the move to online activities from shopping to working at home. In this paper, we analyze trends and changes to understand the levels of disruption from COVID-19 on the data science innovation ecosystem. We consider three phases of innovation: early-stage research and development; late-stage research and development; and commercialization and diffusion. We show that the COVID-19 pandemic has had a negative impact on innovation in data science across the phases of innovation to varying degrees; however, our analysis suggests a return to previous growth may be expected in the short term should the economy continue to improve. These findings suggest that researchers and practitioners should be prepared to take advantage of this return in growth to invest in early stage R&D, to build research programs in data science, and to find ways to commercialize and adopt data science innovations. Our research also identifies the need for coordinated efforts to make current and up-to-date data for tracking innovation impacts available.

CCS CONCEPTS

• Social and professional topics → Economic impact; **• Computing methodologies → Artificial intelligence**.

KEYWORDS

data science, innovation, covid-19

ACM Reference Format:
Michelle Alexopoulos and Kelly Lyons. 2021. Toward understanding the COVID-19 impact on Data Science Innovation in Canada. In *CASCON '21: Annual International Conference on Computer Science and Software Engineering, November 2021, Toronto, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

CASCON '21, November 2021, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The potential for data science to bring value to business, government, and individuals has been rising, in part, as a result of access to growing amounts of data from business applications, sensors, and individuals and the availability of numerous and diverse off-the-shelf artificial intelligence (AI) and data analytics tools, many accessible through virtualized cloud-based computational resources. In 2017, companies' data asset volume was found to be growing an average of 40% per year [24]. Recent studies [14, 16] confirm that mastering big data affords strategic advantages to corporate users, and that early adopters of the most advanced analytics capabilities outperform their competitors. As technologies around digital platforms evolve and mature, the opportunities for local, national and international economies will increase. In fact, some pre-pandemic predictions suggest AI could contribute up to $15.7T to the global economy in 2030 [22]. This is partly because AI is creating new industries and lowering barriers to participation and access [20].

Preliminary data from Statistics Canada shows that, between 2010 and 2017, nominal Gross Domestic Product (GDP) for the digital economy grew more than 40% in Canada (twice the rate of the total economy), accounted for 5.5% of Canada's GDP in 2017, and supported 886,114 jobs (4.7% of the workforce) [25]. However, despite the clear benefits from its adoption, many organizations have faced challenges, be that legal, organisational, or due to existing business practices when seeking to integrate data science within their business frameworks [1, 2]. Most recently they have also faced unprecedented challenges caused by the pandemic. Even before COVID-19, Canadian organizations lagged behind those in other jurisdictions in readiness to adopt data science innovations [26, 28].

There is ample reason to believe that COVID-19 has altered the incentives to invest in and adopt these innovations [5]. For example, COVID-19 has created short-run challenges moving forward for certain applications because of factors such as supply-chain interruptions for necessary hardware; lack of funds due to plunging profits and share prices (e.g., [18] and [30]) and access to skilled workers due to travel restrictions (e.g., the inability to have skilled labour immigrate and poor/blocked internet access from other jurisdictions). For some organizations, it would appear COVID-19 has created significant opportunities. Legislative lock-downs have driven sharp spikes in demand for products and services allowing for remote work (e.g., [15]) and online purchases (e.g., [6]). Review of privacy issues and development related to contact-tracing apps, health care advances, remote work/studies and collection and use of data has been accelerated (e.g., [17]). Soaring share prices and record profits for some big tech companies have also created opportunities for these companies to invest more in their research and development (R&D) activities or purchase smaller start-ups. In short, this shifting landscape will likely have both short- and long-term impacts on the path that digital technology adoption will take and these impacts will vary across firms and industries.

This paper presents analyses in an attempt to quantify the extent of the general level of disruption on data science adoption and innovation in Canada. We use common metrics from the literature to analyze innovation trends and the effects of COVID-19 across three phases of innovation: early-stage R&D; late-stage R&D; and commercialization and diffusion. Specifically, we review research funding and investment data to understand impacts on early-stage research investment, consider patent filings and publications to understand impacts on late-stage R&D, and look at interest in machine learning (ML) development packages, trends in job titles, numbers of newspaper articles, and venture capital (VC) investments to understand impacts on commercialization & diffusion. Our results show that the pandemic has negatively impacted innovation in data science across the three innovation phases but our analysis also reveals some very recent hopeful trends that suggest the downturn may be short lived. In Section 2 we describe the innovation phases in more detail and identify measures of innovation in each of the phases. In Sections 3, 4, and 5, we discuss our assessment of the pandemic's impact on early stage R&D, late stage R&D, and commercialization and diffusion, respectively, and in Section 6 we conclude and discuss implications of the findings.

2 METHODOLOGY

We view trends through the lens of three phases of innovation to understand the levels of disruption from COVID-19 on the data science innovation ecosystem. The phases of innovation are broken into three parts: the early-stage R&D phase, the late-stage R&D phase, and the commercialization & diffusion phase. As shown in Figure 1, each phase blends into the other with the standard S-shape diffusion [23] of the technologies occurring as the late stage research and development leads to commercialization and adoption by individuals and businesses in the economy.

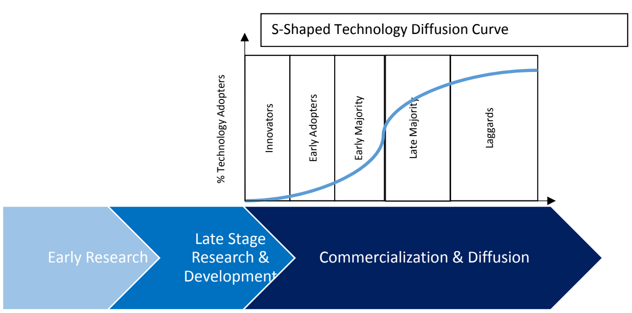

Figure 1: Three phases of innovation.

In order to discuss the impacts of COVID-19 on the three phases of innovation outlined in Figure 1, we need to find appropriate measures we can utilize to characterize potential disruption for each. We begin with the knowledge production function first proposed by Griliches [13] which captures the notion that research and development expenditures influence the production of knowledge since R&D is a critical input in the production process. As knowledge evolves, this in turn leads to changes in tangible and intangible outputs - some of which can be observed through patent filings and the introduction of new products to market. Bibliometric measures based on publications have been used by Alexopoulos [3] and others as an alternative measure of changes in knowledge. This, along with the fact that funding is a major input into research and development is reflected in the augmented knowledge production function displayed in Figure 2.

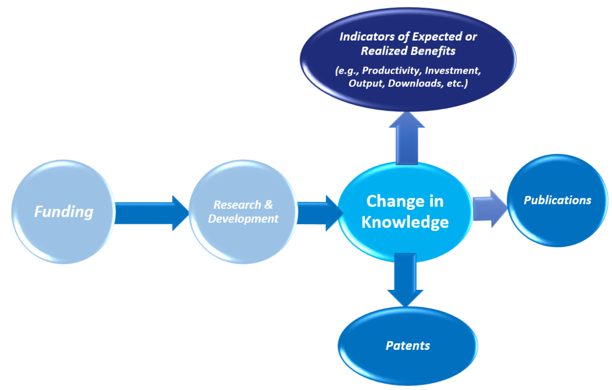

Figure 2: An augmented version of the knowledge production function first proposed by Griliches [13].

Utilizing this process as a basis for our investigation, we collect data on research funding and investment to understand impacts on early-stage R&D investment. We then gather statistics on publication trends in journals and conferences to understand impacts on late-stage R&D. Detailed patent filing data is only made available after a considerable lag that renders it problematic for an analysis on data science related patents filed during the pandemic. For example, the United States Patent and Trademark Office (USPTO) and the Canadian Intellectual Property Office (CIPO) report that applications generally become public 18 months after filing. As a result, we only consider aggregate patent data and total number of filings in our analysis below. Finally, we assemble data on interest in machine learning (ML) development packages, trends in job titles related to data science and AI, coverage of data science and AI related newspaper articles, and VC investments over time to understand the pandemic's potential impacts on commercialization and diffusion.

3 EVALUATING THE IMPACT ON EARLY STAGE R&D

To measure early stage R&D, we focused on investment data and analyzed research funding data over time. R&D expenditure data was gathered from Statistics Canada. Government funding and application statistics were extracted from the Natural Sciences and Engineering Research Council (NSERC) and Compute Canada.

Statistics Canada conducts yearly surveys to determine the levels of Canadian R&D. These data are broken down by type of funder of the R&D activities as well as by performing sector. While this official data does not provide details that would allow us to uniquely identify R&D committed to data science projects, it helps provide some insights into pre-COVID-19 trends in support for natural science related R&D as well as revealing what the funding intentions were before the pandemic occurred for 2020.[1]

As Figure 3 highlights, growth in inflation-adjusted R&D expenditures in Canada have been trending upwards over the 2010-2018 period, and while 2019 saw a decline in expenditures, the reported planned spending for 2020 (in nominal dollars) would have seen inflation adjusted values slightly increase over the 2019 levels. Figures 4 and 5 display the anticipated 2020 pre-pandemic distributions for funding sources and performers of R&D activities, respectively.

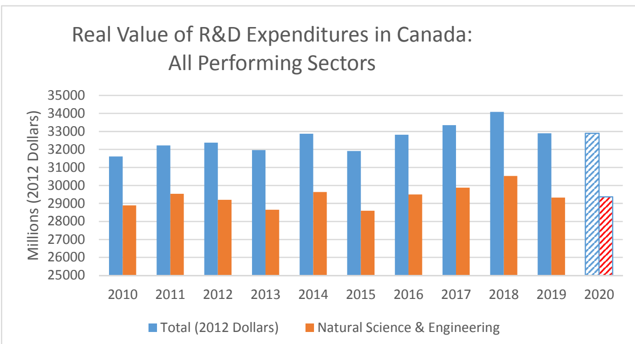

Figure 3: R&D expenditures in Canada.

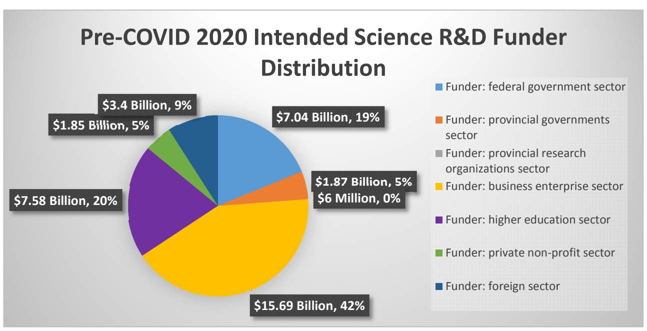

Figure 4: Pre-COVID-19 intended natural science and engineering R&D funder distribution.

Two things are evident from the statistics in these figures. First, the business enterprise sector is the largest funder and performer of natural science and engineering related R&D activities in Canada. Second, the higher education sector is the second largest performer of natural science and engineering R&D. The distribution across groups is consistent with the previous year's patterns. Further as Figure 6 highlights, the amount of R&D related to information and communication technology (ICT) and computer and electronic manufacturing in business enterprise in-house R&D has been growing over time as the economy has increasingly transitioned to a more digital economy. Together these two sectors now account for about 46% of total business enterprise in-house R&D [9] and in-house

[1]Spending intentions for 2020 were collected before the onset of COVID-19.

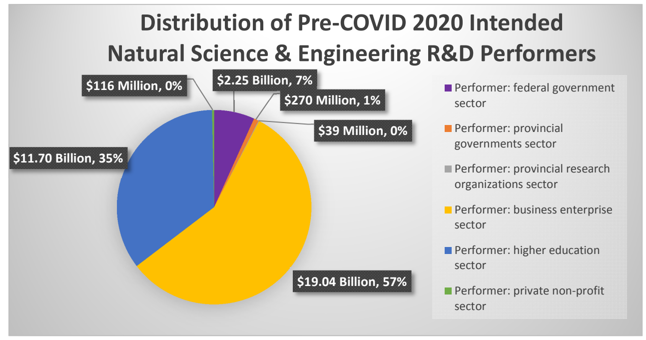

Figure 5: Pre-COVID-19 intended natural science and engineering R&D performer distribution.

R&D spending normally represents about two thirds of business sector R&D funding.[2] Similar to the pattern seen at the aggregate level, data on 2020 spending intentions in these areas collected pre-COVID-19 were expected to increase over the 2019 levels.

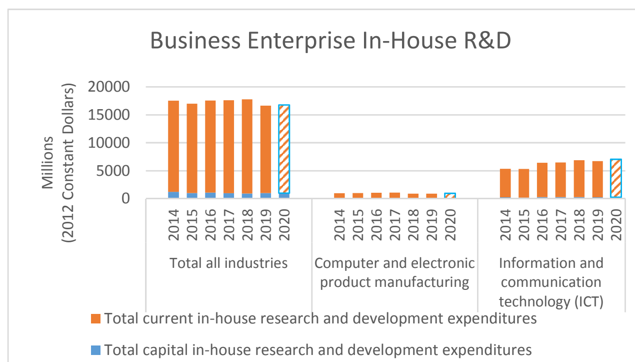

Figure 6: Business enterprise in-house R&D.

Unfortunately, GDP data indicates that COVID-19 has affected the growth in these sectors of the economy. Prior to the pandemic, the average annual growth rates in real GDP in the computer and electronic product manufacturing and the ICT sectors over the 2015-2019 period were approximately 3.2% and 4.64% respectively. In contrast, their annual growth rates during 2020 were -11.7% for the computer and electronic product manufacturing sector and only 2.8% for the ICT sector [10]. Given that R&D is an investment and, at the time that intended expenditure information was collected, organizations would have likely forecast higher growth rates for their sectors than what was realized, the reduction in expected profits and revenues will likely impact their actual investment levels in a negative way and delay late stage R&D-related commercialization.

The next largest funders of Canadian R&D are the federal government and higher education sectors (see Figure 4). Figure 7 displays the time series for actual (2010-2019) and intended 2020 spending for the federal government's support for the higher education sector's natural sciences and engineering R&D. It confirms, along

[2]https://www150.statcan.gc.ca/n1/daily-quotidien/201209/dq201209b-eng.htm

with announcements that the federal government increased funding for science and technology and R&D expenditures during the pandemic, that available research funding for the sector was less affected than that of the private sector [8]. For example, Canada's Natural Sciences and Engineering Research Council's (NSERC's) 2020/2021 budget increased by $10 million during the 2021/2021 fiscal year [7]. However, as we see from other metrics (see Sections 3.1 and 3.2), it would seem that COVID-19 likely disrupted the researchers' planned and future projects in data science and AI related areas in spite of the funding being more stable than that in the private sector.

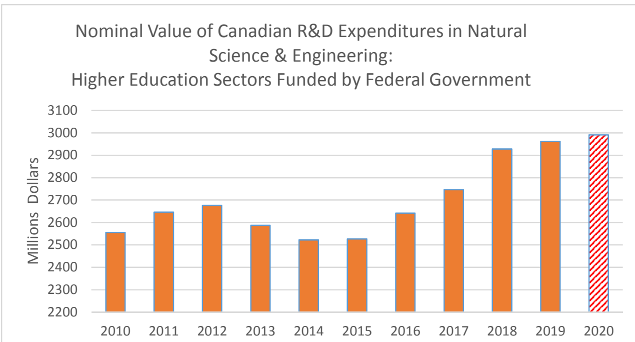

Figure 7: Canadian R&D natural science and engineering funding from federal government to higher education.

3.1 Requests for Data Science related Infrastructure: Compute Canada

Many of the new advances in AI and ML require substantial hardware support. Therefore, given the lack of direct measures of disruption for research in the area, we examine requests for data science related infrastructure as a proxy. Specifically, we collect data about the change in requests by application year through Compute Canada's annual resource allocation competition from 2012 to 2020. We separately identify the trends in requests for CPU and GPU resources for research support. Figure 8 shows an increase in the number of computer hardware requests for CPU and GPU resources to support research over the 2012-2019 period. However, the series also indicates a negative COVID-19 impact on academic research related to AI and data science given the significant decrease in GPU allocation requests and a slowdown in the asks for CPU support in the 2020 application year (which took place in the fall of 2020). These decreases point to a lower level of planned R&D activity in the grant period which runs from April 2021 through March 2022, potentially further dampening the pace of Canadian R&D production at least in the short run.

3.2 Funding Requests: Natural Sciences and Engineering Research Council (NSERC)

We complement our data from Compute Canada with available data from NSERC. Most Canadian Researchers working in the natural sciences and engineering maintain support for their ongoing research through an NSERC Discovery Grant (DG); therefore, DGs

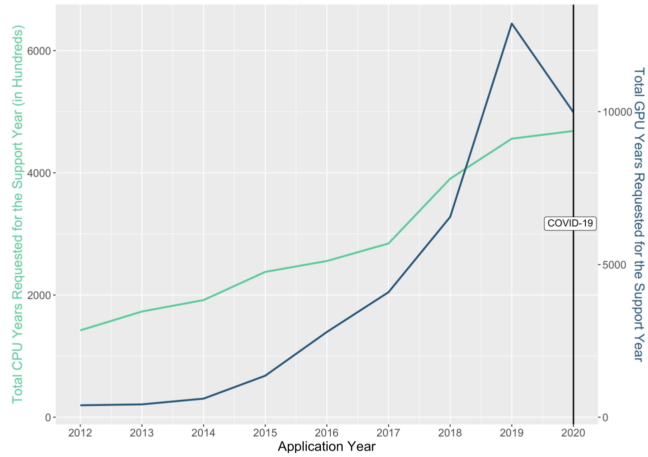

Figure 8: CPU and GPU hardware resource requests.

are a good proxy for the quantity of academic activity in these areas. Given our focus in data science related research, we collected data from the Computer Science (CS), Mathematics and Statistics, and Electrical and Computer Engineering DG Committees on the number of grants awarded, success rates and values of awards (see Table 1).

There are a few key takeaways from Table 1. First, it is clear that there has been a large drop in awarded grants in the 2020-2021 period. Electrical and Computer Engineering appeared to have experienced the largest decline, but even the least impacted of the three groups (CS) had a drop of approximately 41%. Second, the success rates by committee have typically fluctuated from 60%-87% for the 4 years prior to the pandemic. Even though the success rate statistics by committee for the 2020-2021 competition are not yet available (indicated by n/a in Table 1), if the success rates are similar to those in the recent past, the fall in awards again points to a large decrease in applications to support new projects. The magnitude of the drop is at least partly explained by the fact that, in recognition of the extraordinary impact of the pandemic and lockdowns on the ability to work in offices or in labs, NSERC announced on April 8, 2020 that, "To lessen the impact due to COVID-19 ..., all active Discovery Grants can elect to receive a one-year extension with funds at their current funding level ... Grantees due to apply in competition 2021 can elect not to apply for funding this summer/fall during these tumultuous times." [3] However, it is important to recognize that this additional funding and extension of existing awards was a response to the unprecedented negative impact of the pandemic on research in the higher education sector, and, while the situation in Canada improved somewhat in the first few months following the NSERC decision, the second wave of the pandemic that forced the higher education sector online again in the fall was worse. Overall, the evidence available would point to a substantial disruption to Canadian early stage R&D in the area for both the business and higher education sectors.

[3] https://www.nserc-crsng.gc.ca/Media-Media/NewsDetail-DetailNouvelles_eng.asp?ID=1144

Table 1: NSERC Discovery Grant Statistics, 2016-2021.

	Grants Awarded			Success Rate		
Competition Year	Computer Science	Math and Statistics	Electrical & Computer Engineering	Computer Science	Math and Statistics	Electrical & Computer Engineering
2016-2017	206	185	182	62.4	87.0	63.9
2017-2018	231	193	149	67.2	86.5	64.7
2018-2019	222	204	147	67.0	79.4	63.1
2019-2020	245	175	173	66.7	72.8	60.5
2020-2021	144	89	80	n/a	n/a	n/a

4 EVALUATING THE IMPACT ON LATE STAGE R&D

While some of the COVID-19-related funding changes discussed above would also likely have an impact on late stage R&D due to the negative effect on on-going projects, here we discuss the metrics tied to the output of the R&D activities on knowledge creation, shown in Figure 2. This will further enhance our understanding of COVID-19 on late stage of the innovation life cycle. Here we consider both patents and publications as late-stage innovation measures since they each require disclosing innovation outputs.

4.1 Canadian Patent Data

Data available from the USPTO and CIPO does indicate that there has been a decrease in the overall number of patents filed during the 2020 year. Figure 9 displays the CIPO's patent filings as reported by their monthly production statistics. It shows that surges in the number of Canadian COVID-19 cases are correlated with application filings. Patent filings dropped from 38,825 in 2019 to 36,173 in 2020 (a 6.8% decrease) with the largest downturns occurring during the initial onset of the first wave of COVID-19 cases and with surges in cases associated with the second wave. The USPTO also has seen a dropoff in applications with 597,175 filings in 2020, a decrease of 24,278 filings from the prior year. [4] This 3.9% decline in activity was smaller than seen in Canada, and could be partly explained by the differential lockdowns and restrictions imposed by the two countries.

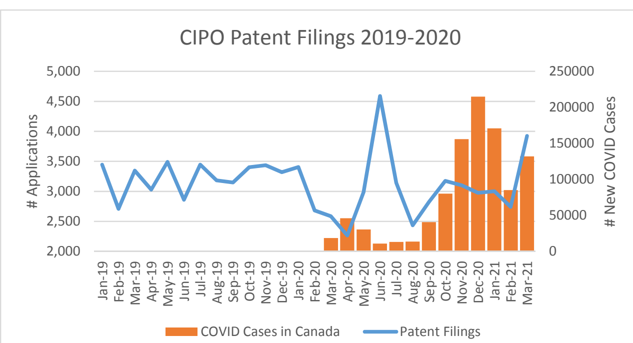

Figure 9: CIPO patent filings and Canada's COVID-19 case count over time.

[4] https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm

4.2 Journal and Conference Publication Data

While the aggregate patent statistics give some support to the hypothesis that COVID-19 had a negative effect on late stage R&D across the board, given the fact that the USPTO and CIPO do not make patent applications publicly searchable until months after their filing dates, we are unable to currently use this source to identify the pandemic's effects on areas related to AI and data science more precisely. Thus, we collected statistics on the other output measures related to late stage R&D, counts of journal and conference publications related to AI and data science. For our purpose, we primarily focused on items indexed in Engineering Village (EV)[5] which provides access to twelve engineering document databases that include journals, conference proceedings, trade publications, patents, and government reports, and confirmed the trends with additional searches on the material indexed in the Web of Science (WoS) Science Citation Index Expanded[6] and Conference Proceedings Citation Index[7] databases. The text of the EV and WoS basic searches are provided in Appendix A.1 and A.2, respectively.

Figure 10 shows an upwards growth over the time period 2010-2019 that clearly indicates an explosion of research output in these areas. Moreover, there is a noticeable break in this trend dated from the onset of the pandemic where the annual growth rate went from 22.1% in 2019, to -3.7% in 2020. This pandemic-related impact on AI and data science publications could be due to the pandemic's effect on researchers directly, as well as its impact on the speed of peer-review, and delays in knowledge transfer due to the postponement and cancellations of conference events.

Given that these statistics include papers authored by researchers in any region, and COVID-19 had differential impacts on countries due to different experiences with lockdowns and infection rates, we further broke down the statistics by region using information on the location of each author's affiliation to determine if there were differences in the publication patterns across countries such as Canada, China, the US, and the UK that experienced significant differences in the severity of the pandemic locally. For example, the number of cases per 100,000 individuals in Canada in June 2021 is 3723, compared to 10,000, 6761, and 8 for the US, UK, and China, respectively.

Due to the fact that a paper can be authored by researchers in multiple jurisdictions, we created two sets of regional breakdowns. In the first case (identified with the label *1+ Authors in...* in Figures

[5] https://www.engineeringvillage.com/home.url
[6] https://clarivate.com/webofsciencegroup/solutions/webofscience-scie/
[7] https://clarivate.com/webofsciencegroup/solutions/webofscience-cpci/

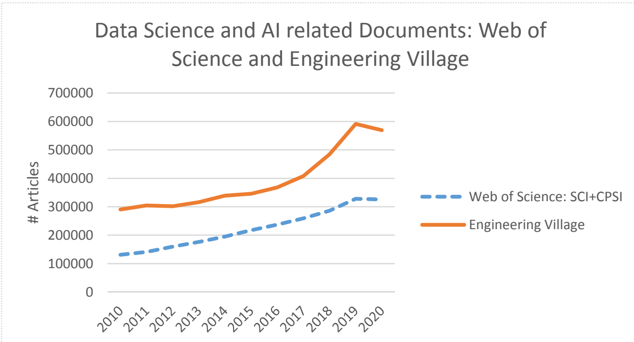

Figure 10: Data science and AI related documents from WoS and EV.

11 and 12) a paper was included in the country's count if there existed at least one author identified as being at an institution in that country. In the second case, a paper was included in a country's count only if all authors were identified as being in an institution in that country (labelled as the *Authors all in...* in the two figures). The rationale for examining these two cases relates to the notion that having some co-authors outside your country may provide a type of insurance on disruption since members of the team in other jurisdictions may be able to pick up slack if one author is unable to work temporarily due to restrictions, family care responsibilities associated with school closures, illness, etc.

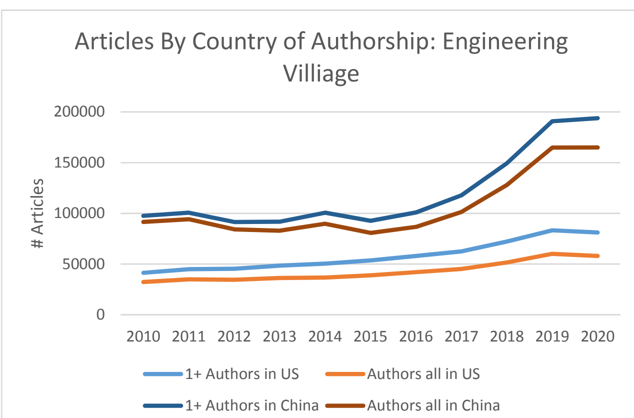

Figure 11: EV articles by authors in the US and China.

Figure 11 shows a robust growth in the number of papers in AI and data science for both the US- and China-based authors from 2010-2019. However, the authors from the US (the country with the highest infection rate of the four countries), experienced a decline in overall publications regardless of the regional measure employed. Specifically, the growth rates for the *1+ Authors in US* group and the *Authors all in US* group went from 15.2% and 16.5% in 2019 to -2.7% and -3.6% in 2020 respectively. China, on the other hand, experienced the lowest infection rate of the four countries we examined, and groups that had an author in China experienced a slightly positive growth in the overall number of their publications. Although

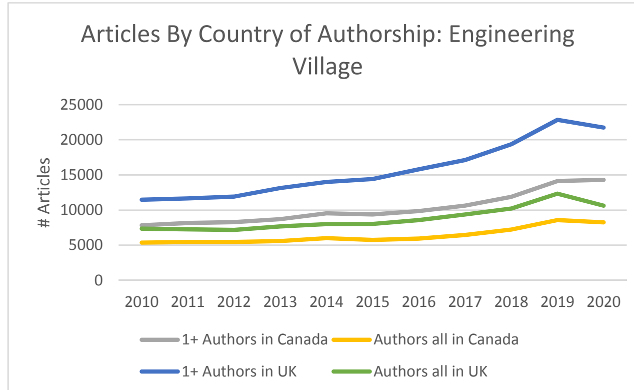

Figure 12: EV articles by authors in Canada and the UK.

this increase fell far below what would have been predicted based on recent years' levels. In particular, the growth rates for the *1+ Authors in China* and *Authors all in China* groups went from 27.9% and 28.7% in 2019 to 1.5% and 0.1% in 2020 respectively.

Figure 12 shows that similar patterns emerge for the UK and Canada. As is the case with the US and China authored papers seen in Figure 11, there is a break in publication trends for UK- and Canada-affiliated papers that coincided with the pandemic year. Moreover, it is clear that the impact was larger for the UK groups which had a higher infection rate. The growth rates for the *1+ Authors in UK* group and the *Authors all in UK* group went from 17.9% and 20.8% in 2019 to -4.9% and -13.8% in 2020 respectively. The growth rates for the *1+ Authors in Canada* and *Authors all in Canada* groups, in comparison, fell from 19.0% and 18.8% in 2019 to 1.3% and -3.7% in 2020, respectively.

Overall, the data supports a case for a COVID-19-related disruption in AI and data science related late-stage research. Moreover, it would appear that, while in normal times there may be a slight advantage to having teams of researchers working within the same geographical area, during the pandemic, there was likely some productivity gains to have diversification of team members across regions due to regional variations in lockdown measures and infection rates.

5 EVALUATING THE IMPACT ON COMMERCIALIZATION & DIFFUSION

Next we explore whether there is evidence that COVID-19 affected the last phase of the innovation life-cycle - commercialization and diffusion. Once again, the type of current disaggregated data at a monthly or quarterly level we would optimally like to examine to answer this question is unavailable from official sources; therefore, we have turned to a series of alternative statistics. Specifically, to examine commercialization and diffusion patterns, we first collected data on the volume of Google and Bing searches for ML Python package names and download rates of ML Python packages over time. This search behaviour is used as a proxy for diffusion since it reflects individuals' desires to learn about specific ML Python packages over time which should be directly related to development, training, and implementation of these tools and their related methods. Second, we analyzed the trends in data science job titles over time since they signal changes in the number of individuals that are serving in roles where data science adoption is (or has been) taking place. We also considered patterns in the number of newspaper articles that discuss data science related terms since coverage of the innovations are typically referenced in relation to topics such as firm productivity, new investment opportunities and the effect of adoption on employment, and, as such, can serve as a proxy for commercialization and diffusion of innovation over time.

5.1 Trends in ML Python Package Use

While some papers in the past have utilized Google Trend data to capture interest in a topic or track the spread of disease (e.g., [11] and [29]), we use a close variant in our work based on data from Keywords Everywhere.[8] This tool combines data from Google Trends with some of Google's other database information such as Keyword Planner to retrieve raw search volumes. This makes it easier to compare across justifications over time since the data is not normalized by other searches.[9]

For our investigation we collect monthly Google and Bing search volume data (worldwide and Canada specific) for popular ML Python package names [12, 19, 21] from April from 2017 to April 2021 (Appendix A.4 lists the search terms considered; the data was downloaded between April 10-15, 2021). Figure 13 shows the trend over time of the number of searches by package name worldwide and Figure 14 shows the same for Canada only. We again see signs of a potential disruption in activities due to COVID-19. The volume of Google and Bing searches for ML Python package names decreased during the months of the pandemic. Global search data shows a steady decline in the volume of searches starting in April 2020 whereas, in Canada, the volume of searches starts to decline in March 2020 during the peak of the first wave of the pandemic in this country. In both cases, there is a small increase in search volume after the annual drop in December 2020 but overall searches remain below the pre-pandemic volume. Given that search behaviour changes across dates is expected to be correlated with changes in interest in the use of these data science related packages, it would appear that these numbers signal a potential disruption in data science system development which may indicate a delay in commercialization and diffusion of data science innovations.

While there are multiple sources where one can download ML Python libraries, we explored statistics from Anaconda's condastats.[10] We chose this source since Anaconda is touted as the world's most popular data science platform with over 25 million users across 235 regions currently reported, and over 2.4 billion package downloads in 2019.[11] The statistics on package downloads from condatstats for the tensorflow, pytorch and keras packages, displayed in Figure 15 over the period Jan 2019-December 2021 appear to follow the same downward trend following the COVID-19 surge in cases in early 2020 (Note: this data was downloaded on March 9, 2021).

[8]https://keywordseverywhere.com/
[9]Google trends data is normalized according to the following procedure https://support.google.com/trends/answer/4365533?hl=en
[10]https://www.anaconda.com/blog/get-python-package-download-statistics-with-condastats
[11]https://www.anaconda.com/about-us

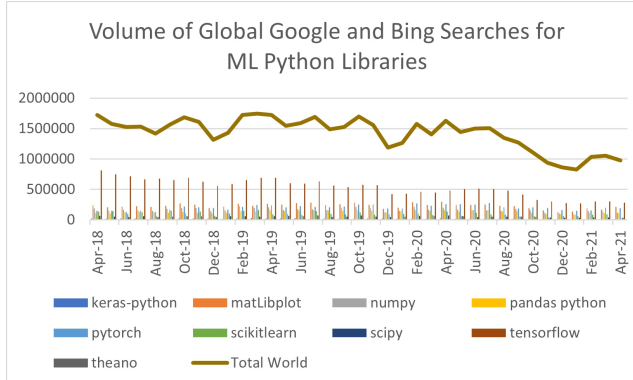

Figure 13: Searches for ML Python packages over time worldwide.

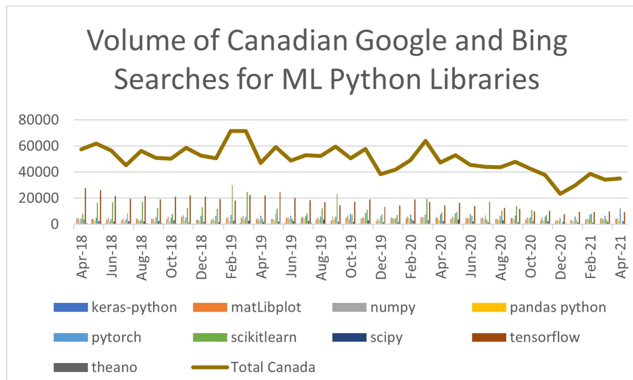

Figure 14: Searches for ML Python packages over time in Canada.

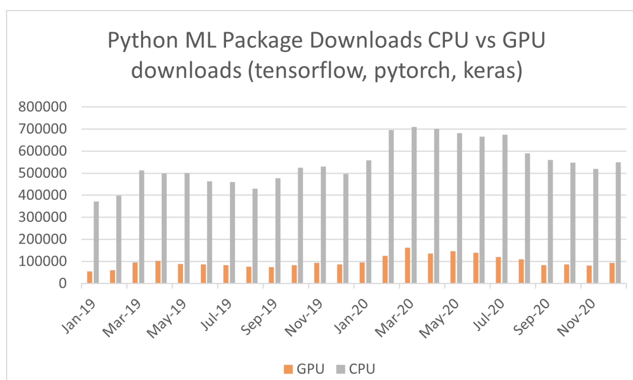

Figure 15: Python ML package downloads CPU vs GPU for tensorflow, pytorch, and keras.

This would seem to confirm our hypothesis that the decline in search behaviour is correlated with downloading and installing certain popular ML packages during COVID-19. Moreover, these

	Global	US	Canada
June 2017-June 2018	3.83%	4.29%	n/a
June 2018-Oct 2019	2.81%	2.33%	1.02%
Oct 2019-Nov 2020	1.24%	0.81%	0.68%
Nov 2020-April 2021	4.79%	1.39%	2.40%

Table 2: Estimated Monthly Growth in Data Scientist Job Titles on LinkedIn.

effects likely stem from pandemic-related lockdowns and restrictions on business, workers' health and wellness, and supply chain disruptions that have led to a shortage for and significant price increase for GPUs and other high performance chips.[12]

5.2 Trends in LinkedIn Job Titles

Building on the idea that workers using ML and data science tools are a good measure of commercialization and diffusion, we merged existing data on the number of people with the title "data scientist" on LinkedIn at discrete points in time from June 2017-April 2021 to give some indication on the trend's trajectory. The trends are based on the occurrence of the term data scientist appearing in the current job title on the platform (similar to the method used in [27]). The statistics are reported in Table 2. Consistent with the pattern seen from our other metrics, it would appear that there was a noticeable decline in the growth of data science jobs during the period October 2019 to November 2020. However, since the economies have begun to reopen following the vaccine rollouts, we appear to be returning to a more robust growth environment. If the most recent trends continue, it would suggest that at least the downturn we see evidence of during the pandemic should have a short run impact on diffusion.

5.3 Trends in Data Science Topics in the News

While trends in academic publications can help us examine trends in late stage R&D activities, examining counts in newspaper articles tied to AI and data science can be equally useful in discerning implementation of these tools and techniques in industries and businesses [4]. The rationale behind these metrics is similar to that for other bibliometric measures. First, given that newspapers make money by selling copies and/or generating funds through selling advertisements, they seek to cover events that are of interest to their readers and which will maintain or expand the number of people reading their articles. Second, implementation of data science and AI related innovations have important implications for employment opportunities, job loss, training required for future jobs, as well as firm profitability, international competitiveness and government policies. Based on these two observations, it follows that more news items appear when more firms are adopting these innovations over time. Figure 16 presents coverage of AI and data science related topics in English language newspapers in Factiva's database[13] and shows data confirming the same trends exist when we consider only US and Canadian newspapers (search terms are given in Appendix A.3). Whether or not we restrict the articles to be both on the topic of AI / data science AND tagged to be related to the region the newspaper is published in, or just focus on any articles referencing AI / data science related terms, it would appear that coverage grew rapidly until 2018, slowed during 2019, and then fell further during the pandemic. This decline in the number of newspaper articles that discuss data science related terms again may suggest that there were fewer innovations and investment opportunities on which to report since the start of the pandemic. Figure 17 shows the coverage of data science topics in Canadian newspapers monthly from January 2020 to May 2021. These more recent trends clearly indicate a rebound consistent with the pattern seen in the LinkedIn Data.

To provide insights into areas where implementation is taking place or is likely to take place in the near future, we examined patterns of frequently-mentioned industries in the newspaper data using the industry tags assigned by Factiva. The findings reveal, unsurprisingly, that the sector discussed most often is the technology industry itself, with the next most frequently mentioned sectors being the financial services industry, business and consumer services, retail and wholesale trade, media/entertainment, the automotive sector and the industrial goods sectors. The patterns of mentions are not uniform across the sectors. For example, during the pandemic when the overall number of articles related to AI and data science fell, articles related to these technologies and healthcare rose as the innovations in the area were utilized to help track the pandemic and aid in the search for treatments and diagnostic tests. Moreover, as the economy reopens, it is clear that the discussion on the innovations in relation to other sectors are again front and center. By June 2021, the number of articles related to the technical change in the retail and wholesale trade, agricultural and industrial goods sectors have surpassed the totals seen in 2020. Moreover, should the current trends in the average rates of publication by industry continue for the remainder of the year, the only sectors that would not have the number of articles surpass their 2020 totals are the leisure and hospitality, real estate and construction, and transportation related industries, although the totals for industrial goods, retail and wholesale trade, health care/life sciences, business and consumer services, agriculture, utilities and the technology sectors itself would surpass their 2019 pre-pandemic levels. Currently, this wide-spread rebound would suggest adoption of commercialized technologies are returning towards more normal levels.

Overall these patterns are also consistent with those seen in the VC investment. Specifically, statistics from CB insights highlight a decrease in overall VC during COVID-19, as well as a dip in investment directed to AI VC globally and a subsequent resurgence in funding activities during the first two quarters of 2021.[14] Investments in healthcare related AI companies, however, does not appear to have experienced the same headwinds and disruptions during the pandemic as seen in other sectors.[15]

[12] https://www.pcmag.com/news/prices-for-nvidia-rtx-3000-graphics-cards-are-getting-insane-on-ebay
[13] https://professional.dowjones.com/factiva/
[14] e.g., https://www.newswire.ca/news-releases/canadian-vc-reported-3-59b-in-first-9-months-2020-continuing-downward-trend-826170816.html
https://techcrunch.com/2021/07/22/canadas-startup-market-booms-alongside-hot-global-vc-investment/
https://www.cbinsights.com/research/report/ai-in-numbers-q1-2020/
[15] https://www.cbinsights.com/research/report/ai-trends-healthcare

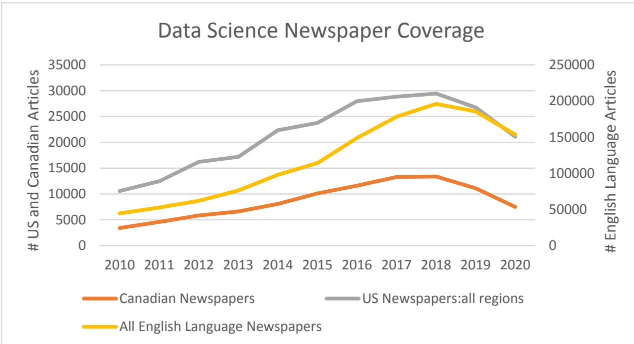

Figure 16: Coverage of data science topics in English-language newspapers.

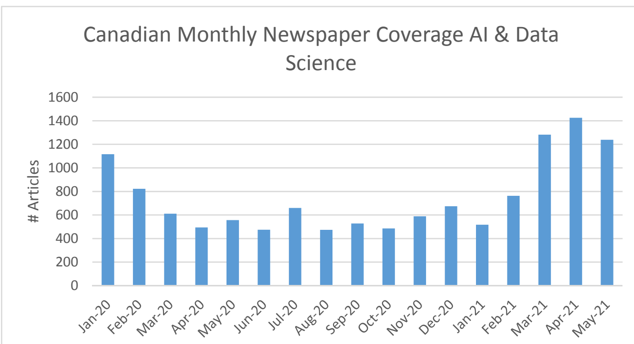

Figure 17: Monthly coverage of data science topics in Canadian newspapers.

6 CONCLUSIONS AND IMPLICATIONS

Overall, our analysis indicates that there was widespread disruption across all three phases of the innovation lifecycle. There were declines in research funding requests, patent filings, number of articles published, searches for AI and ML packages, newspaper coverage of data science and AI topics, venture capital funding and a drop in the growth rate of data science related jobs. While some of our evidence suggests that the slowdown in Canada began pre-pandemic in 2019, R&D statistics previously suggested there would have been a partial turnaround in 2020 without the pandemic. It would appear that the supply chain disruptions, the lockdown of business, universities and colleges, and health considerations have culminated in a negative impact on innovation in AI and data science. The good news is that the most recent few months of data show some hopeful trends suggesting that the downturn will be short lived. Overall patent filings in Canada are beginning to rise again, the number of news articles related to AI and data science are starting to increase along with VC funding activity. Search volume for ML python libraries has a recent uptick, and data from LinkedIn shows that positions related to AI and data science that had stalled during the pandemic may be on the rise again. These findings have implications for researchers and practitioners of data science. While opportunities for data scientists and data science researchers have generally diminished as a result of the pandemic, these changes appear to be temporary and researchers and practitioners should be prepared to take advantage of the opportunity to invest in early stage R&D, develop research programs in data science, and to find ways to commercialize and adopt data science innovations.

Another important implication of our research points to the lack of data on issues of innovation. We were able to consider measures of innovation that have been used in past work by collecting data from unconventional sources to track the effects of the pandemic. Future research is needed to understand the long-term impacts of the pandemic on data science innovation, adoption, and diffusion in Canada but this will require that more disaggregated and up-to-date data be collected and made available. While some economic data is being made available,[16] a coordinated effort is needed in the same way that the availability of COVID-19-related data around the world was fast-tracked early during the pandemic.[17]

ACKNOWLEDGEMENTS

The authors wish to thank P. Layarda, H. Pandya, K. Mahetaji, K. Chiu, A. Yang, and M. Diaz for their contributions.[18] This work was co-funded by the Social Sciences and Humanities Research Council and the Government of Canada's Future Skills program.

REFERENCES

[1] Rohan Alexander and Kelly Lyons. 2020. Barriers to Service Innovation Using Data Science. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 57–62.

[2] Rohan Alexander, Kelly Lyons, Michelle Alexopoulos, and Lisa Austin. 2019. Workshop on barriers to data science adoption: why existing frameworks aren't working. In *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*. 384–385.

[3] Michelle Alexopoulos. 2011. Read all about it!! What happens following a technology shock? *American Economic Review* 101, 4 (2011), 1144–79.

[4] Michelle Alexopoulos and Jon Cohen. 2019. Will the new technologies turn the page on US productivity growth? *Economics Letters* 175 (2019), 19–23.

[5] Michelle Alexopoulos, Kelly Lyons, Rohan Alexander, Aije Egwaikhide, and R Blair Frost. 2020. How has COVID-19 changed the development and adoption of data science across firms and industries?. In *Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering*. 260–261.

[6] Jason Aston, Owen Vipond, Kyle Virgin, and Omar Youssouf. 2020. Retail e-commerce and COVID-19: How online shopping opened doors while many were closing. Available online at: https://www150.statcan.gc.ca/n1/pub/45-28-0001/2020001/article/00064-eng.htm, Last accessed 22 June 2021.

[7] Statistics Canada. 2021. Table 27-10-0005-01 Federal expenditures on science and technology in current and constant dollars (x 1,000,000). Available online at: https://doi.org/10.25318/2710000501-eng, Last accessed 22 June 2021.

[8] Statistics Canada. 2021. Table 27-10-0026-01 Federal expenditures on science and technology, by major departments and agencies - Intentions. Available online at: https://doi.org/10.25318/2710002601-eng, Last accessed 22 June 2021.

[9] Statistics Canada. 2021. Table 27-10-0333-01 Business enterprise in-house research and development expenditures, by industry group based on the North American Industry Classification System (NAICS), country of control and expenditure types (x 1,000,000). Available online at: https://doi.org/10.25318/2710033301-eng, Last accessed 22 June 2021.

[10] Statistics Canada. 2021. Table 36-10-0434-06 Gross domestic product (GDP) at basic prices, by industry, annual average, industry detail (x 1,000,000). Available online at: https://doi.org/10.25318/3610043401-eng, Last accessed 22 June 2021.

[11] Hyunyoung Choi and Hal Varian. 2012. Predicting the present with Google Trends. *Economic record* 88 (2012), 2–9.

[12] Claire D. Costa. 2020. Best Python Libraries for Machine Learning and Deep Learning. Available online at: https://towardsdatascience.com/best-python-libraries-formachine-learning-and-deep-learning-b0bd40c7e8c, Last accessed 22 June 2021.

[13] Zvi Griliches. 2007. *13. Patent Statistics as Economic Indicators: A Survey*. University of Chicago Press.

[16] e.g., https://www.statcan.gc.ca/eng/covid19
[17] e.g., https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/epidemiological-economic-research-data.html
[18] https://futurejobscanada.economics.utoronto.ca/

[14] Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. 2016. The age of analytics: Competing in a data-driven world. Available online at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world, Last accessed 22 June 2021.

[15] Mansoor Iqbal. 2021. Zoom Revenue and Usage Statistics (2021). Available online at: https://www.businessofapps.com/data/zoom-statistics/, Last accessed 22 June 2021.

[16] Ian MacGregor. 2018. Big Data: The Canadian Opportunity. Centre for International Governance Innovation. Available online at: https://www.cigionline.org/articles/big-data-canadian-opportunity, Last accessed 22 June 2021.

[17] Fabrício Martins Mendonça and Mário Antônio Ribeiro Dantas. 2020. Covid-19: Where is the Digital Transformation, Big Data, Artificial Intelligence and Data Analytics? (2020).

[18] BBC News. 2020. Coronavirus: Uber customer activity falls sharply. Available online at: https://www.bbc.com/news/business-53687422, Last accessed 22 June 2021.

[19] Hacker Noon. 2021. Top 8 Python Libraries for Machine Learning & Artificial Intelligence. Available online at: https://hackernoon.com/top-8-python-libraries-for-machine-learning-and-artificial-intelligence-y08id3031, Last accessed 22 June 2021.

[20] G7 Academies of Science. 2018. The 2018 G7 Academies' statement on Realizing Our Digital Future and Shaping its Impact on Knowledge, Industry, and the Workforce. Available online at: https://rsc-src.ca/sites/default/files/G7Statement-Digital.Final.pdf, Last accessed 22 June 2021.

[21] Irina Popova. 2021. Libraries for Machine Learning, LIGHT IT, Light of the Future. Available online at: https://light-it.net/blog/top-10-python-libraries-for-machine-learning/, Last accessed 22 June 2021.

[22] Anand Rao and Gerard Verweij. 2017. Sizing the prize: PwC's Global AI Study—Exploiting the AI Revolution. Available online at: https://www.pwc.ch/en/publications/2017/pwc_global_ai_study_2017_en.pdf, Last accessed 22 June 2021.

[23] Everett M. Rogers. 1962. *Diffusion of innovations (1st ed.).* Free Press of Glencoe.

[24] James Short and Steve Todd. 2017. What's Your Data Worth? *MIT Sloan Management Review* 58, 3 (2017), 17.

[25] Amanda Sinclair. 2019. Measuring digital economic activities in Canada: Initial estimates, Statistics Canada. Available online at: https://www150.statcan.gc.ca/n1/en/pub/13-605-x/2019001/article/00002-eng.pdf, Last accessed 22 June 2021.

[26] Richard Stirling, Hannah Miller, and Emma Martinho-Truswell. 2017. Government AI Readiness Index 2017. Available online at: https://www.oxfordinsights.com/government-ai-readiness-index/, Last accessed 22 June 2021.

[27] Stitch. 2015. The State of Data Science. Available online at: https://www.stitchdata.com/resources/the-state-of-data-science/, Last accessed 22 June 2021.

[28] Nigel Wallis. 2012. Big Data in Canada: Challenging Complacency for Competitive Advantage. Available online at: https://silo.tips/download/w-h-i-t-e-p-a-p-e-r-b-i-g-d-a-t-a-i-n-c-a-n-a-d-a-c-h-a-l-l-e-n-g-i-n-g-c-o-m-p, Last accessed 22 June 2021.

[29] Lynn Wu and Erik Brynjolfsson. 2015. The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy*. University of Chicago Press, 89–118.

[30] Thorsten Wuest, Andrew Kusiak, Tinglong Dai, and Sridhar R Tayur. 2020. Impact of COVID-19 on manufacturing and supply networks—The case for AI-inspired digital transformation. *Available at SSRN 3593540* (2020).

A KEYWORDS AND SEARCH TEXTS USED

A.1 Search text used with EV

(((((learning (artificial intelligence) or cloud computing or artificial intelligence or data mining or big data or machine learning or predictive analytics or Internet of Things or data analysis or deep learning or neutral nets or feature extraction or neural networks or learning algorithms or support vector machines or automation or computer vision or remote sensing or robotics or data analytics or heuristic algorithms or intelligent computing or recommender systems or data science or clustering algorithms or data acquisition or advanced analytics or convolutional neural networks or mobile computing or intelligent systems or mobile robots or logistic regression or multilayer neural networks or virtual reality or deep neural networks or natural language processing systems or natural language processing or computational linguistics or speech recognition or text analysis or pattern classification or learning systems or text mining or text processing or sentiment analysis or speech processing or pattern clustering or speech synthesis or ontologies (artificial intelligence or recurrent neural nets or industrial robots or robots or robot applications or computer aided manufacturing or intelligent robots or robots, industrial or agricultural robots or medical robotics or service robots)) WN ALL)))

A.2 Search text used with WoS

AB=(learning (artificial intelligence) OR cloud computing OR artificial intelligence OR data mining OR big data OR machine learning OR predictive analytics OR Internet of Things OR data analysis OR deep learning OR neutral nets OR feature extraction OR neural networks OR learning algorithms OR support vector machines OR automation OR computer vision OR remote sensing OR robotics OR data analytics OR heuristic algorithms OR intelligent computing OR recommender systems OR data science OR clustering algorithms OR data acquisition OR advanced analytics OR convolutional neural networks OR mobile computing OR intelligent systems OR mobile robots OR logistic regression OR multilayer neural networks OR virtual reality OR deep neural networks OR natural language processing systems OR natural language processing OR computational linguistics OR speech recognition OR text analysis OR pattern classification OR learning systems OR text mining OR text processing OR sentiment analysis OR speech processing OR pattern clustering OR speech synthesis OR ontologies (artificial intelligence OR recurrent neural nets OR industrial robots OR robots OR robot applications OR computer aided manufacturing OR intelligent robots OR robots, industrial OR agricultural robots OR medical robotics OR service robots))

A.3 Search text used with Factiva

(“cloud computing" OR "artificial intelligence" OR "data mining" OR "big data" OR "machine learning" OR "predictive analytics" OR "Internet of Things" OR "data analysis" OR "deep learning" OR "neutral nets" OR "feature extraction" OR "neural networks" OR "learning algorithms" OR "support vector machines" OR "automation" OR "computer vision" OR "remote sensing" OR "data analytics" OR "heuristic algorithms" OR "intelligent computing" OR "recommender systems" OR "data science" OR "clustering algorithms" OR "data acquisition" OR "advanced analytics" OR "mobile computing" OR "intelligent systems" OR "logistic regression" OR "virtual reality" OR "natural language processing systems" OR "natural language processing" OR "computational linguistics" OR "speech recognition" OR "text analysis" OR "pattern classification" OR "learning systems" OR "text mining" OR "text processing" OR "sentiment analysis" OR "speech processing" OR "pattern clustering" OR "speech synthesis" OR "recurrent neural nets" OR robot* OR "computer aided manufacturing")

A.4 Search terms used with Keywords Everywhere

(each term was searched separately) tensorflow, pytorch, keras-python, theano, numpy, scipy, scikit-learn, pandas-python, and matplotlib