Canada's Innovation Landscape

Challenges, Opportunities, and the Future of Canadian Innovation

March 31st, 2024

Chris Fitzgibbon
Neil Sanjiv Punjani
Pat Dhanespramoj
Ziying Xu
Layth Al-Nemri

Executive Summary

Canada's innovation landscape is embodied with historical successes and world-changing inventions. But in 2023, the World Intellectual Property Organization ranked Canada just 15th globally, behind rich-world peers such as Japan, Denmark, the UK, and the US. This report explores how Canada can improve its innovation in a data-driven approach. Using an ensemble of regression models, the underlying indices of the WIPO ranking were reverse engineered to establish their relative weights in influencing the WIPO innovation rank. The most influential indices were then filtered to isolate the input indicators, i.e. indices which can impact innovation and are not merely a result of already good innovation or a strong economy. Eight highly influential input indicators were identified, and Canada's performance was compared directly to that of other OECD countries. The findings show that Canada lags significantly behind in two critical indices: GERD (i.e. R&D) financed by business (as opposed to public institutions), and national spending on education. Both indices were below the OECD average – revealing opportunities for Canada to drastically improve.

Looking to other countries for inspiration on improvement across these two metrics, a sentimental analysis of news postings across wealthy countries revealed that while Canada does not show a relative distaste for business R&D, media focus on this topic lags substantially behind that of education in STEM fields. Further, public-facing national innovation strategy documents reveal the efforts of highly innovative countries are largely related to the roll-out of highly integrated digital technologies such as smart cities, and investment in sustainable infrastructure such as renewable energy.

Using this data-driven approach to identify Canada's weaknesses compared to its peers, key recommendations are to create a more business-friendly environment by removing unnecessary regulatory barriers, invest more in education and specifically into the emerging field of sustainability, and provide funding and tax incentives to businesses developing clean technologies.

Regression modeling of significant input indicators show that increasing Canada's performance in just one metric is not sufficient to improve innovation ranking, but rather it takes a well-rounded approach to propel Canada forward. A linear regression model suggests that a 36% increase across all input indicators will place Canada 5th globally. While this is ambitious, there is a knock-on effect of putting money and effort at the core of innovation: good education, and an environment where businesses flourish.

Table of Contents

Executive Summary 2
Table of Contents 3
Introduction 4
Canada's Current Global Status 4
Canada's Innovation Policy 5
What Influences Innovation Performance? 6
Gaps in Canada's Innovation Strategy 8
What are other countries doing? 9
How can Canada Improve its Global Standing? 11
Conclusions 12
Definitions & Acronyms 13
References 14

Introduction

Canadian innovations – what do we think of? Basketball, the igloo, perhaps Nanaimo bars. What about the *world changing* inventions which credits Canada as their motherland? Therapeutic insulin, after being isolated in a University of Toronto basement laboratory in 1922, saved millions of lives of people around the world living with the then-crippling disease known as diabetes (University of Toronto, 2024). The telephone, a revolutionary remote conversation in March 1876 in Brantford, Ontario which paved the way for modern communications (Canadian Encyclopedia, 2024). The neural network, conceived by Geoffrey Hinton and his graduate students at the University of Toronto opened the door to the seemingly limitless potential for generative artificial intelligence (TIME Magazine, 2023). These inventions mark parts of history for which Canada defined the global path forward. But where does Canada stand today, and what is Canada's future in the world's innovation landscape?

This report will identify Canada's current standing amongst countries in the world and will highlight key challenges and opportunities. By utilizing the powerful capabilities of data science, statistical models, and automated text processing, recommendations for improvements to Canada's innovation-shaping policy will be made, fortifying Canada's innovation space and placing Canada amongst the world's top innovating countries.

Frederick Banting and Charles Best pose on the rooftop of their University of Toronto laboratory with a beagle which helped them discover and isolate therapeutic insulin in 1922.

Canada's Current Global Status

The World Intellectual Property Organization publishes annually a snapshot of the globe's innovation space by ranking countries based on their "Global Innovation Index" (GII), a compound index comprised of 80 individual indices including creative outputs, infrastructure, regulatory environments, and business sophistication, among others. With the top rank awarded to Switzerland, Canada ranks 15th in the world (WIPO, 2023). While this puts Canada ahead of 89% of the world and 69% ahead of the OECD countries, plenty of room for improvement remains. Where does Canada excel, and where does it struggle? We first take a look at Canada's current innovation policy, which sheds light on where our nation's policymakers have focused their attention and efforts to improve our global standing in innovation.

Canada's Innovation Policy

Canada's innovation policy is published by Innovation, Science, and Economic Development Canada (ISEDC) in their 2019 document *Building a Nation of Innovators* (ISEDC, 2019). Using PDF text extraction, natural language processing, and generative AI, Canada's policy as outlined in the ISEDC document is summarized and key interventions extracted (refer to Appendix 1 for details on language processing methods). The key take-aways from the ISEDC policy are as follows:

Canada's Existing Programs and Expected Impacts:

- Encourage 1 million youth per year to pursue STEM as part of the **PromoScience** program
- Enroll 1 million students and recruit 53,000 teachers to the **CanCode** course
- Provide 11,500 work-integrated learning opportunities through the **Student Work Placement** program
- Increase work-integrated learning placements to 8,000 through the **Mitacs** program
- Provide 600 internships to students through the **Digital Skills for Youth** program
- Enhance rural internet connections in 900 communities through the **Connect to Innovate** program
- Create 40,833 high-skilled jobs aimed at immigrants through the **Global Skills Strategy** program
- Create 50,000 jobs and grow Canada's economy by $50B over the next 10 years through the **Innovation Superclusters** initiative

Canada's Key Interventions for 2019:

- Allocate $4B to support researchers in Canada as part of the **2018 Federal Budget**
- Allocate $2.3B in funding for clean technology as part of various cleantech initiatives, such as **Export Development Canada**, **Cleantech Scale Up Initiative**, and **Sustainable Development Technology Canada**
- Expand the **Strategic Innovation Fund** by $1.15B to provide late-stage funding for growth firms
- Create **a Women Entrepreneurship Strategy** and allocate $2B to reduce systemic barriers that women face in growing their careers
- Simplify business support programs by improving access and consolidation of programs through the **Innovation Skills Plan**

Canada's Highlighted Strengths:

- High competencies in science, math, reading for Canadian students
- Well-educated workforce with 56.7% of the working-age population holding a tertiary degree
- Home of world-class post-secondary institutions
- Productive innovation partnerships and technology hubs
- Strong foundation of economic fundamentals and GDP growth

Canada's Highlighted Weaknesses:

- Gaps in STEM, business, creative, and digital skills within the workforce
- Lack of overall diversity in research community
- Decrease in investment in business R&D
- Limited late-stage capital, hampering scale-up of small firms
- Difficult regulatory environment with firms struggling to navigate regulations and interact with government bodies

What Influences Innovation Performance?

To understand the driving factors of what shapes a country's success in innovation, we look to other countries to contrast their policies and economies with that of Canada. Using statistical modeling tools, the macroeconomic metrics which are most responsible for a country's innovation success were identified. To achieve this, the indices of the GII 2023 innovation rankings were extracted using a custom PDF-scraping program and cleaned into a readable dataset (refer to Appendix 2 for data cleaning methods and Appendix 3 for regression models). Taking the innovation score as a target variable, feature correlation and two supervised machine learning algorithms identified the indices which most strongly correlate with high innovation scores as per the GII ranking. These indices are listed in Figure 1 for each model.

Correlation Matrix (correlation > 0.70)	**Random Forest Regression** (top 15)	**Linear Regression** (top 15)
• High-tech exports	• **Patent families**	• Logistics performance
• Mobile app creation	• **ICT use**	• **ISO 14001 environment**
• Cost of redundancy dismissal	• **ICT access**	• **ICT use**
• **Graduates in science and engineering***	• **GERD fiannced by business***	• **ICT access**
• **Global corporate R&D investors**	• **Regulatory quality***	• **Government effectiveness**
• QS university ranking	• **Government effectiveness**	• **Regulatory quality***
• Electricity output	• **Government online service***	• **GitHub comments**
• **Gross capital formation**	• PCT patents by origin	• GERD financed by abroad
• Firms offering formal training	• Citable documents H-index	• **Patent families**
• Patents by origin	• GitHub comments	• Unicorn valuation
• **Policies for doing business***	• Gross expenditure on R&D*	• **E-participation***
• PISA scales in reading	• High-tech manufacturing	• **Government online service***
• **ICT use**	• **ISO 14001 environment**	• **Expenditure on education***
• **Governments online services***	• Mobile app creation	• **Graduates in science and engineering***
• **E-participation***	• Tertiary enrollment	• **Finance for startups and scaleups***
• **Finance for startups and scaleups***		
• **GERD financed by business***		

Figure 1: Macroeconomic indices which show high correlation or importance to GII innovation score. Bold indicates index was found within the top features for two or more methods, * indicates input indicators.

It was noted that many highly influential indices are merely a ***result*** of an already productive and innovative economy (known as "output indicators"). For example, the macroeconomic index *gross capital formation*, a measure of the change in infrastructure and inventory assets as a proportion of total GDP, is more likely to be an indicator of good economic performance rather than a lever which can be manipulated by policy. As the purpose of this report is to identify adjustable factors which can shape Canada's innovation performance, the focus will be on indices which, when manipulated through policy, can ***change*** Canada's innovation landscape (known as "input indicators"). The top 8 most influential input indicators are shown in Figure 2. These indices are broadly categorized as either financial, regulatory, knowledge, or online presence metrics. Figure 3 (next page) displays each country's relative performance on these 8 key input indicators.

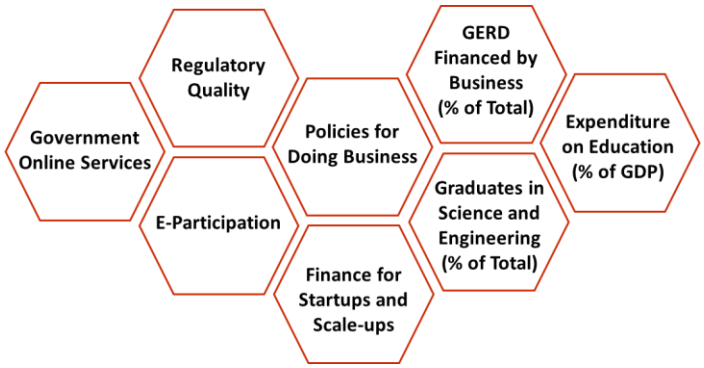

Figure 2: Influential input indicators which shape innovation across the globe.

Innovation-shaping metrics across the globe

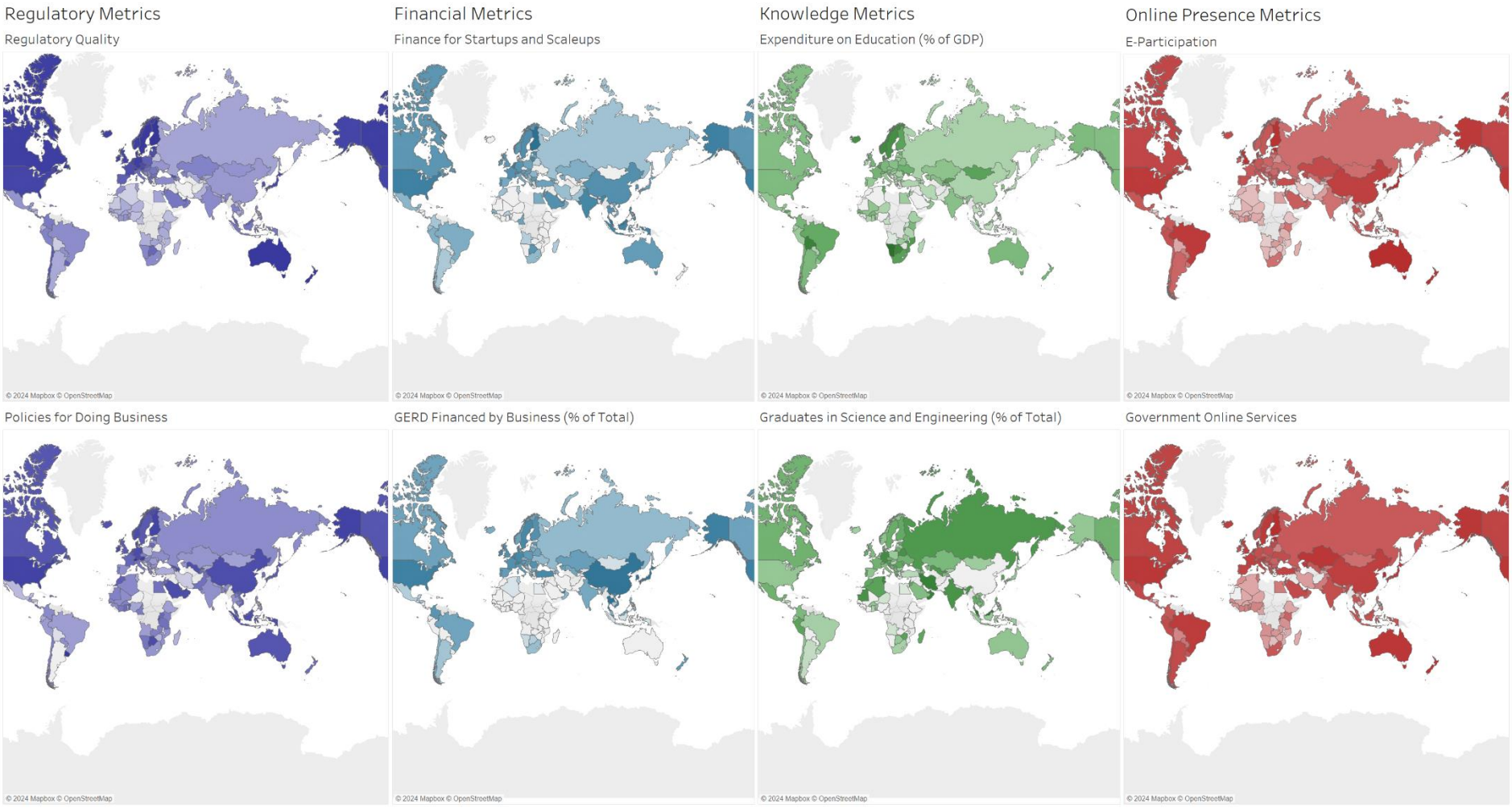

Figure 3: Relative global scores of 8 macroeconomic indices which shape innovation. Darker colours indicate better performance relative to global peers.

Gaps in Canada's Innovation Strategy

When compared on a global scale, Canada's performance on the 8 key innovation indices shown in Figure 2 highlight where Canada excels and where Canada lags behind its global peers. Using the visual in Figure 3, where Canada shows a lighter colour than other rich nations in North America, Oceania, and Europe, opportunities are revealed where Canada can improve. Notably, Canada shows visually lower scores on the financial metrics ***finance for startups and scaleups*** and ***GERD financed by business*** and the regulatory metric ***policies for doing business***.

Finance for Startups and Scaleups
A survey-based index representing the availability of capital for small and mid-siezd businesses in their growth phase. This index attests to the financial strength of domestic and international debt and equity investing and includes capital from banks, venture capital firms, and governments.

GERD Financed by Business
A numerical index measuring the total amount of money businesses spent on R&D activities as a proportion of total national R&D expenditures. GERD (gross domestic expenditures on research and development) is a measure of R&D spending.

While visual tools for highlighting Canada's performance globally is useful to identify gaps, it is important to compare Canada directly to its rich-world peers to narrow down indices which require the most attention for improvement. Figure 4 compares Canada against the OECD countries for the same 8 input indicators, revealing where Canada performs above or below the OECD average. Canada is shown to have outstanding performance in ***regulatory quality*** and ***E-participation***. However, Canada lags below the 75th percentile in six metrics, with two below the OECD average. The poorest performing index is ***GERD financed by business***, placing in the bottom 17% of OECD countries. As a result, it can be concluded that improving performance in input indicators ***policies for doing business***, ***finance for startups and scaleups***, ***expenditure on education***, ***graduates in science and engineering***, ***government online services***, and ***GERD financed by business*** can provide the most impact in improving innovation in Canada.

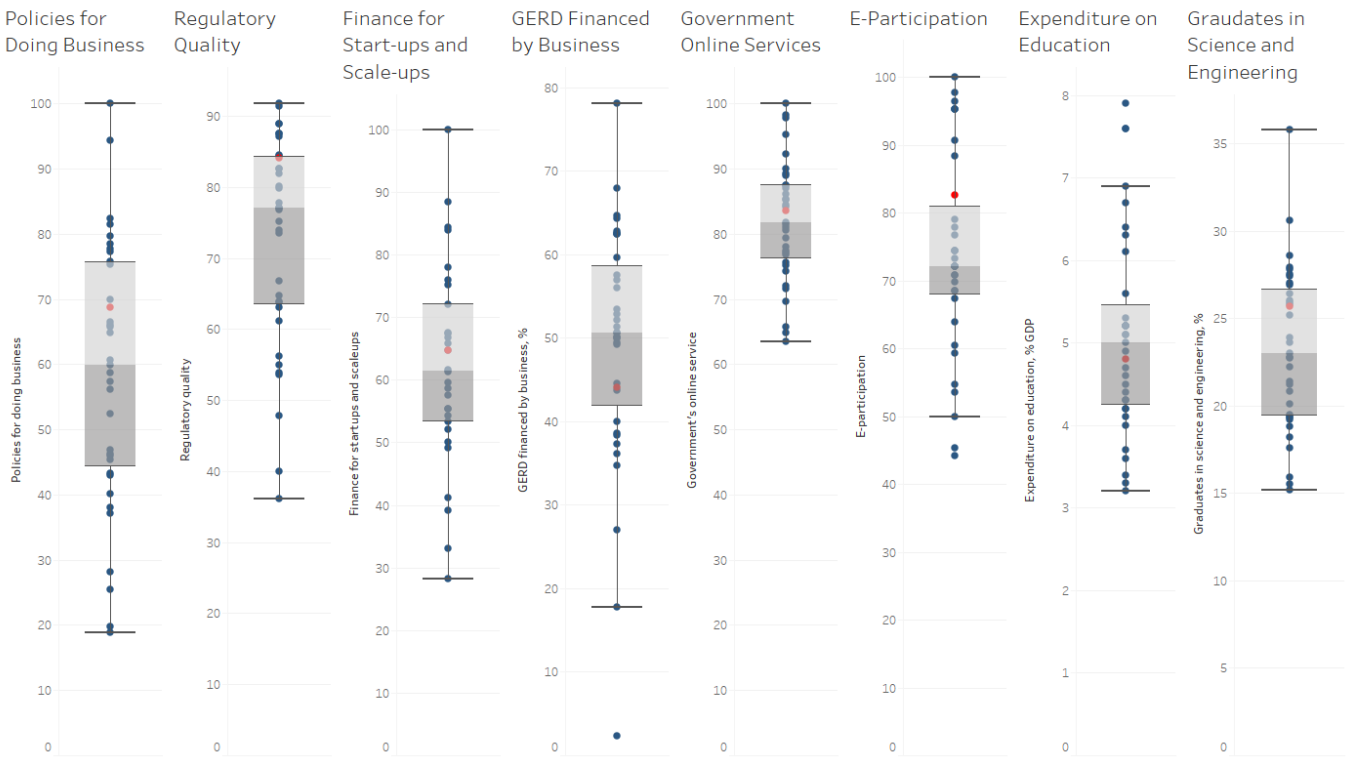

Figure 4: Box and whisker plot for 8 input indicators for OECD countries. Canada is highlighted in red.

Canada's poorest performing indices relative to the OECD occurs in two spending categories: educational, and R&D. As the expenditure on education metric is limited to public-sector education spending, this input indicator is the direct result of the federal government's education budget allotment for that year. Therefore, altering this metric is straightforward and is a changed solely by policymaking at the federal level.

By contrast, GERD financed by business is a metric which by definition is an output of the private sector only and is therefore much more difficult to modify with public policy.

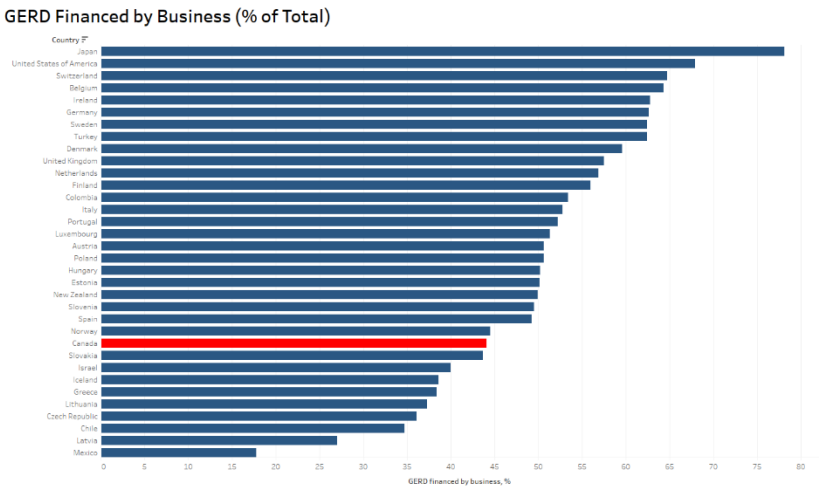

Figure 5: GERD financed by business for OECD countries. Canada is highlighted in red.

Government Online Services
A compound index representing the quality and availability of online services such as passport, visa processing, and health insurance services.

Expenditure on Education
A numerical index representing the percentage of a country's GDP which is spent on education, including primary, secondary, and post-secondary education.

Policies for Doing Business
A compound index representing the relative ease of doing business. This may consider regulatory streamlines, tax incentives, or other business-friendly policies.

Graduates in Science and Engineering
A numerical index representing the proportion of students graduating post-secondary with engineering or science degrees.

What are other countries doing?

Turning to the innovation strategies of other nations is helpful in ideating where Canada can be doing better based on what its successful rich-world peers are implementing and talking about. Two data-mining approaches were undertaken: news scraping combined with a sentimental analysis, and frequency analysis for public-facing strategy documents originating from consultancies, think-tanks, and foreign governments.

Sentimental analysis of foreign and domestic news

By analyzing the underlying sentiment of news headlines across several wealthy for the keywords "education in STEM" and "business R&D", Canada's relative apathy or sentiment towards its two ailing innovation metrics can be compared amongst its successful peers. Google News stories were analyzed for 8 high-performing countries, with the bulk of stories appearing between 2014 and 2024. Results of this analysis are visualized in Figure 6 (refer to Appendix 4 for news scraping and sentiment analysis methods).

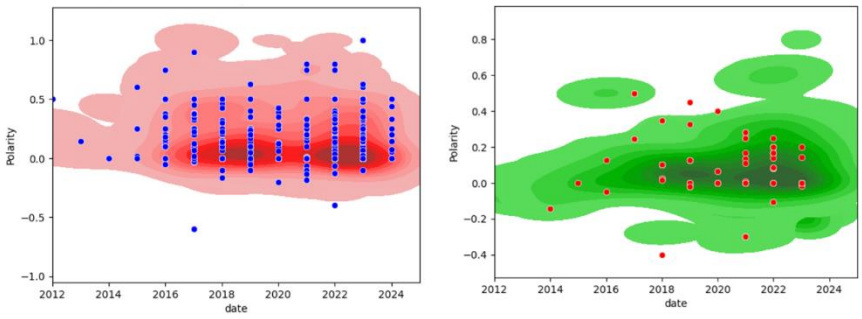

Figure 6: Sentimental analysis of Canada (dots) compared to other highly innovative nations (heat cloud) of news stores in Google News for keywords "education in STEM" (left) and "business R&D" (right). Higher values represent "positive" or "favorable" sentiment.

While there appears to be no significant difference amongst Canada's media sentiment when compared across the rich world, news headlines are undoubtedly discussing graduating students in STEM fields more so than GERD financed by business. This again seems on-par with other countries but highlights a need for increased attention to business innovation which is revealed as an overlooked but critical input to a country's innovation performance.

Frequency analysis of foreign innovation strategies

Both governments and international consultancies publish public-facing documents highlighting national strategies (or recommendations for strategies, in the case of consultancies) for improving innovation. By extracting keyword frequency from these documents, a concise view of international innovation-boosting strategies can be generated. Several countries or regions were considered: Dubai, the EU, Japan, Korea, Singapore, Switzerland, UK, and USA.

Strategies extracted from the national documents were contrasted with those outlined in the ISEDC document for Canada's strategy. Only strategies which were not found in Canada's ISEDC were retained and are presented in Figure 7 as a showcase of the unique strategies taken by other countries. Note that four of the strategies refer to green energy or sustainability.

- Advancing cybersecurity for small businesses and charities
- **Focusing on ESG marketplaces for sustainable investments**
- **Developing smart sustainable cities for efficiency and sustainability**
- **Harnessing renewable energy from various sources**
- Exploring social robots and AI for social services
- Utilizing technology for mental health care
- Increase connectivity and interdependence
- **Transition to a low-carbon, resource-efficient economy**
- Communicate the value-add of R&I to the public
- Planning of national data centers

Figure 7: Innovation-boosting strategies planned by other countries which Canada is not considering as part of its 2019 ISEDC publication.

In addition to identifying unique strategies, a frequency analysis was done on all mentioned strategies and were categorized into five distinct categories to understand which fields are at the forefront of other countries' efforts to boost innovation.

Several keywords appeared repeatedly, particularly pertaining to roll-out of digital technology, investment in sustainable energy and cleantech, development of strong governance and standards, and international cooperation. Results of this analysis are visualized in Figure 8 (refer to Appendix 5 for frequency analysis methods).

The interpretation of this analysis is that Canada should be looking more towards the use of technology and investing development of clean, sustainable technologies to place it among other strong performing countries.

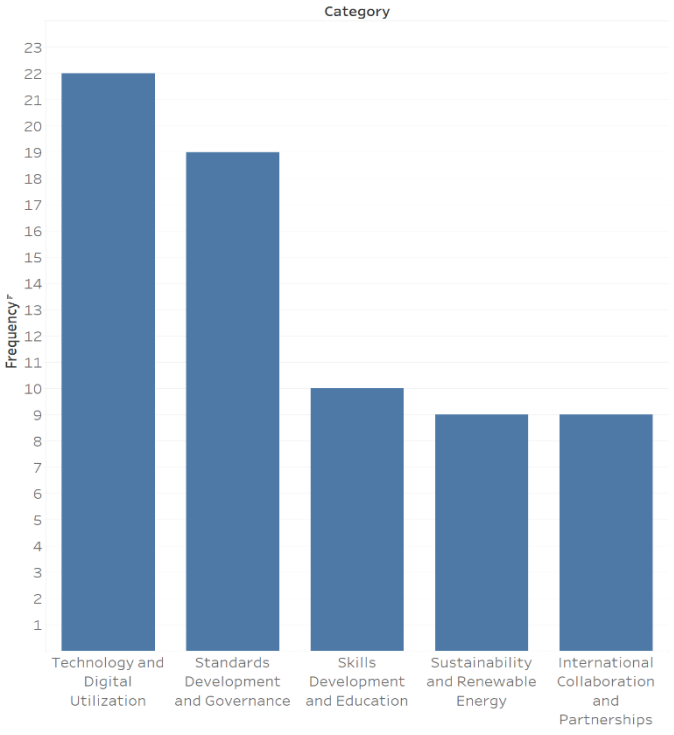

Figure 8: Frequency of different strategy categories based on relative appearance in national innovation strategy documents.

How can Canada Improve its Global Standing?

Canada's current innovation policy as outlined by the ISEDC's 2019 publication recognizes the weaknesses in financing startups, regulatory complexity, and low business R&D spending. This analysis however reveals an alternative approach may be required to propel Canada forward in its global innovation standing, summarized by the following recommendations.

Foster a business-friendly environment by breaking down regulatory barriers.

Canada is uniquely mired by unnecessary complexity, and several examples are seen in the engineering profession alone. Canada has twelve regional engineering licensure bodies, creating unnecessary bureaucracy and cost for engineers looking to work inter-provincially (The Economist, 2016). Similarly, excessive control over the engineering designation has prevented software engineers from working in Alberta thanks to a recent ruling by APEGA, Alberta's professional engineering watchdog, that these experts are unable to refer to their profession as "engineering" due to their lack of licensure (APEGA, 2023). Interprovincial barriers have historically handicapped Canada's ability to do business, so Canada should invest in removing red tape for businesses, ultimately improving Canada's future R&D outlook.

Enhance funding for businesses developing sustainable technologies.

Canada provides a 35% tax credit up to $3M (15% after $3M) to businesses performing R&D under the SR&ED federal tax credit (INNOTAX, 2022). Additional programs exist such as the Strategic Innovation Fund (SIF) provide funding on a fixed-amount grant basis (Government of Canada, 2023). Although these programs have positive impacts across all sectors, attaining net-zero goals while boosting domestic innovation can only be possible if Canada directs more of its efforts towards businesses working specifically on clean technologies or roll-out of renewable energy generation. The Canadian government should take the working principles of the SR&ED incentives and the SIF but apply their funding and focus specifically to Canadian-based cleantech startups such as General Fusion (which is currently funded largely by the SIF already (General Fusion, 2023)). This redirection of innovation funds to sustainability and climate change mitigation efforts will pay dividends in abated climate risk.

Expand financing of education in sustainability fields.

Canada's innovation policy puts a strong emphasis on funding STEM and computer sciences education, such as the PromoScience program (Government of Canada, 2016) and the CanCode course (Government of Canada, 2022), with considerable funding continuing since the ISEDC's publication in 2019. As a relatively new standalone field, it is unsurprising that sustainability is left out of education funding policies. However, with the emergence of clean technologies prevailing not just our infrastructure but the economy as well, the importance of teaching and promoting sustainability education is understated in Canada's current policy. Canada should not only increase its overall spending on education but direct the focus of new spending into education of clean energy, circular economy, and other sustainability initiatives.

Improve innovation with a well-rounded approach.

Regression modeling of Canada's input indicators reveals that simply increasing one metric cannot alone improve Canada's global innovation rank as evidenced by the infeasible increase in single-score performance required to put Canada in the top 5 (refer to Table 1 in Appendix 3). This reveals that a well-rounded approach is required. Linear regression modelling indicates that increasing numerous input indicators (refer to Table 2 in Appendix 3) by 36% each result in a GII ranking increasing to the top 5.

Conclusions

While Canada performs on-par with OECD countries in its global innovation ranking, as a historical innovation powerhouse Canada should strive to improve its current innovation landscape to land itself among the world's top innovating countries today. The significant input indicators which influence innovation scoring were identified using an ensemble of regression modeling, which then compared Canada directly to other OECD countries. Two indices in particular, expenditure on education and GERD financed by business, stood out as poor performers putting Canada well behind the OECD average. With a focus on these two indices, other country's strategies and innovation sentiment were solicited to understand where the world's top innovators are putting their efforts. Sentiment analysis of news and keyword frequency of national strategy documents reveal that many nations are focused on the roll-out of digital technology and sustainable infrastructure.

Given Canada's current state among the OECD and the strategies of its rich-world peers, Canada can improve its innovation score by removing unnecessary regulatory burden to improve its business environment, investing in education in sustainability fields, and providing funding and tax incentives to businesses looking to commercialize clean technologies.

Regression modelling shows that increasing input indicators by 36% each can propel Canada into the top 5 of global innovators. While ambitious, the plan outlined in this report focuses on the core of innovation which includes strong education and business-first policies. The cascading effects of investing in the core input indicators can substantially improve Canada's global standing and re-assert itself as the world's best innovator.

Definitions & Acronyms

Acronym	Definition
APEGA	Alberta Professional Engineers and Geophysicist Association
B	Billion
GDP	Gross Domestic Product
GERD	Gross Domestic Expenditures on Research and Development
GII	Global Innovation Index
ISEDC	Innovation, Science, and Economic Development Canada
NRCC	National Research Council Canada
NSERC	Natural Sciences and Engineering Research Council of Canada
R&D	Research and Development
SIF	Strategic Innovation Fund
SR&ED	Scientific Research & Experimental Development
STEM	Science, Technology, Engineering, and Math
OECD	Organization for Economic Co-operation and Development
WIPO	World Intellectual Property Organization

References

APEGA, 2023. APEGA Appealing Court Decision on Engineer Title Usage [WWW Document]. APEGA. URL https://www.apega.ca/news/2023/12/11/apega-appealing-court-decision-on-engineer-title-usage (accessed 3.19.24).

Canadian Encyclopedia, 2024. Alexander Graham Bell and the Invention of the Telephone [WWW Document]. URL https://www.thecanadianencyclopedia.ca/en/article/whats-better-than-bells-telephone-feature (accessed 3.19.24).

General Fusion, 2023. Canadian Government Awards New Funding to General Fusion [WWW Document]. Gen. Fusion. URL https://generalfusion.com/post/canadian-government-awards-new-funding-to-general-fusion/ (accessed 3.22.24).

Government of Canada, I., 2023. Strategic Innovation Fund - Home [WWW Document]. URL https://ised-isde.canada.ca/site/strategic-innovation-fund/en/ (accessed 3.22.24).

Government of Canada, I., 2022. CanCode - Home [WWW Document]. URL https://ised-isde.canada.ca/site/cancode/en/cancode (accessed 3.20.24).

Government of Canada, N.S. and E.R.C. of C., 2016. NSERC - Science Promoters - PromoScience [WWW Document]. Nat. Sci. Eng. Res. Counc. Can. URL https://www.nserc-crsng.gc.ca/Promoter-Promotion/PromoScience-PromoScience/Index_eng.asp (accessed 3.20.24).

INNOTAX, 2022. Scientific research and experimental development (SR&ED) tax credit details - Canada | INNOTAX Portal [WWW Document]. URL https://stip.oecd.org/innotax/incentives/CAN1 (accessed 3.22.24).

ISEDC, 2019. Building a Nation of Innovators. Government of Canada.

The Economist, 2016. The great provincial obstacle course, The great provincial obstacle course. The Economist.

TIME Magazine, 2023. TIME100 AI 2023: Geoffrey Hinton [WWW Document]. Time. URL https://time.com/collection/time100-ai/6309026/geoffrey-hinton/ (accessed 3.19.24).

University of Toronto, 2024. Insulin 100: University of Toronto [WWW Document]. Insul. 100. URL https://insulin100.utoronto.ca/ (accessed 3.19.24).

WIPO, 2023. Global Innovation Index 2023. [object Object]. https://doi.org/10.34667/TIND.48220

Appendix 1: Language Processing for Summarization of ISEDC Document

Step	Method
Extracting PDF information and summarizing main points using ChatGPT API	
PDF text extraction	Utilizes `PyPDF2` to extract text from specified PDF pages, combining it for analysis. This method targets particular document sections efficiently.
Text cleaning	After converting the text to lowercase for uniformity, it's tokenized into words using nltk.word_tokenize, stripping out stopwords via nltk for clarity. Tokenization splits the text into manageable pieces, essential for precise analysis and removing irrelevant words to highlight the text's core content.
GPT model configuration	Sets parameters for GPT-3.5-turbo, including creativity (`temperature`), output length (`max_tokens`), and variance and output control (`frequency_penalty`, `presence_penalty`), to tailor the summary output.
API summary request	Sends cleaned text to OpenAI's API, leveraging GPT to summarize and highlight key data. Parameter adjustments influence summary style and detail.
Performing keyword analysis for each relevant section of the report	
Custom stopwords for contextual relevance	Defined custom stopwords for different sections (e.g., `cust_exec`, `cust_disrup`) to remove frequent but non-informative words specific to the domain, such as "innovation" and "government." This improves keyword relevance by excluding common terms that do not contribute to distinguishing themes or topics within each section.
Part-of-speech tagging for enhanced lemmatization	Implemented `get_wordnet_pos` to map POS tags from the Penn Treebank format to WordNet format, enabling more accurate lemmatization. This step is crucial for normalizing words to their base form, considering their role in sentences, thereby improving the consistency and relevance of extracted keywords.
Advanced text cleaning and lemmatization	The `clean_text` function incorporates tokenization, stopword removal (including custom stopwords), and lemmatization using WordNet's POS tags. This comprehensive cleaning ensures the text is free of irrelevant words and standardized in terms of word forms, laying a solid foundation for effective keyword extraction.
TF-IDF for keyword extraction	Utilizes `TfidfVectorizer` to perform keyword extraction across different sections, emphasizing terms that are important and unique to each section relative to the entire corpus. This method helps in identifying keywords that best represent the thematic essence of each section based on their TF-IDF scores.
Removal of irrelevant sections	Irrelevant feature of keyword analysis such as Executive summary, Introduction and Current Innovation were eliminated as it was more relevant for our assignment to identify Canada's innovation policy going forward and their plan.
Qualitative analysis of feature importances	After TF-IDF analysis, a manual review of the top features (keywords) is conducted to filter out irrelevant or nonsensical terms. This step is necessary as automated stopword removal and lemmatization do not catch all contextually irrelevant terms.

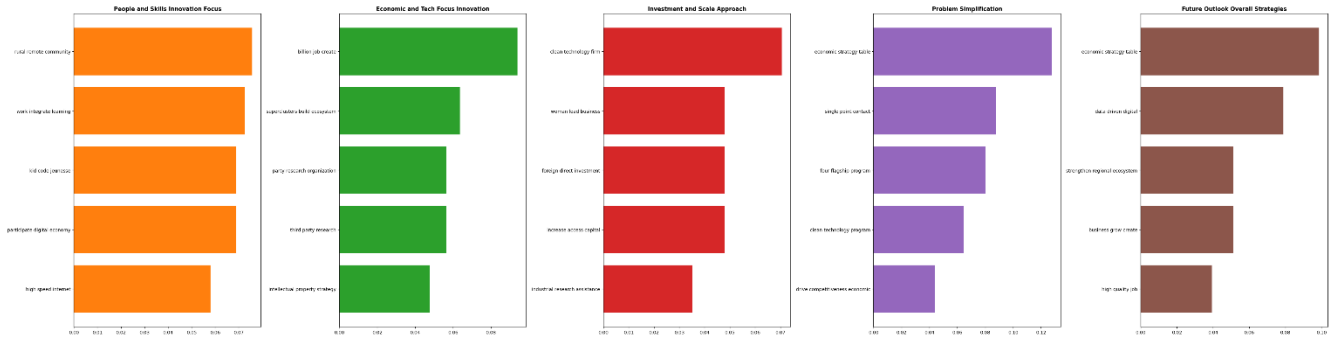

Figure 9: Output for N-grams word frequency for each section of ISEDC document

Appendix 2: Data Extraction and Cleaning of WIPO Document

Step	Method
Extracting PDF information	
PDF text extraction	A function used the 'fitz' library to read and load text from pages. Only the pages containing the country's innovation scores are required to be read-in.
Remove special characters	Due to the presence of several special characters including diamonds, circles, and clocks (representing unique attributes of the indices in the document), a function is required to remove these special characters.
Split lines which were concatenated	Individual data did not always appear on the same line in the text output. To clean this, two functions were required to identify unique patters in what made the text concatenate on the same line and separate the lines.
Iterate over all pages and combine into single .csv output	The functions are called and iterated over all relevant pages (page 76-207 of the WIPO document). All pages are combined into a single dataframe called "df_all_pages"
Print dataframe to .csv file	
Print .csv using UTF-16 encoding	The dataframe "df_all_pages was exported as ""innovation_scores.csv" using UTF-16 encoding to preserve special characters.

Appendix 3: Regression Models to Establish Feature Importance

Step	Method
Data Cleaning	
File import	Output file from the GII document extraction "innovation_scores_cleaned" was imported into the jupyter notebook.
Column removal	Columns (features) with more than 50% missing values were removed to reduce bias and as they were unlikely to contribute meaningful insights to the analysis.
Missing value imputation	For numeric columns, the median value of a country's continent group was used to impute missing values. This method allowed for relatively accurate imputation (as continents tend to share a similar economic status, at least more so than when considering the whole world). For categorical columns, the mode was used to impute missing values.
Re-calculate null rows	The number of null rows was re-calculated and confirmed to be zero before proceeding.
Prepare Target Variable	
Invert GII rank	Since the best GII score is 1, to improve interpretability, the best score was inverted to 132, with the worst being 1, using a simple N + 1 – GII score where N is 132 or the total number of participating countries. This also ensures higher values represent higher innovation, aligning the goal of associating greater numerical values with superior innovation.
Feature Importance	
Drop target and split data	After dropping the target variable from the dataset, the data was split into test and train sets (10% / 90% respectively).
Random Forest Regression	Feature importance was calculated using RandomForestRegressor which handles non-linear relationships and interaction between features. It evaluates how much each feature decreases the impurity of the split to identify feature importance. This model calculates the importance of a feature by looking at how much prediction errors increase when data for that feature is permuted while all the others are left unchanged. The feature importances were then ranked, providing initial insights into which variables most significantly influence the innovation rankings.
XGBoost	Feature importance was calculated using XGBoost. This algorithm offers a feature importance metric that quantifies each features's contribution to model performance. This is achieved through a "gain" measure, which identifies the average contribution of a feature to the model's predictions across all threes. By ranking these features based on their importance scores, we gained further insights into the predictive power of each variable in the context of innovation rankings. It should be noted that between Random Forest and XG Boost, the top 70 features were similar.
Ordinal Logistic Regression	Feature importance was calculated using Ordinal Logistic Regression with K-fold cross-validation. This method is suitable for ordinal target variables such as the GII ranking. Prior to model fitting, the features were standardized to ensure that all variables operate on the same scale, enhancing model fairness and convergence. Since this algorithm can involve multiple binary logistic models (one for each threshold in the ordinal scale), identification of feature importance is performed by summing the absolute coefficients for each feature.
Sensitivity analysis	Due to the non-linearity of random forest and XG boost algorithms, a sensitivity analysis was required to determine the increase in Y target variable based on a unit increase of feature X. The features in Canada's dataset were incrementally increased to make predictions on the GII rank. Common high-importance features were selected and increasing values were placed into the models to see what % increase would be required to put Canada into the top 5 in GII scores.

Appendix 3: Regression Models (continued)

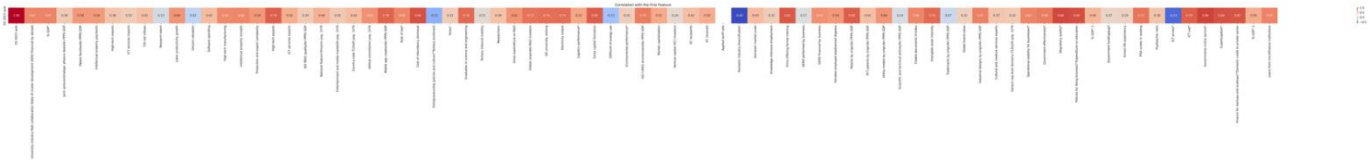

Figure 10: Feature importance by correlation matrix

Table 1: Change in feature performance which would propel Canada to top 5 in GII ranking

Feature	% change required to put Canada into top 5 GII rank
Logistics performance	338.5
ISO14001 environment/ billion PPP$ GDP	242.2
ICT use	322.6
ICT access	1001.0
Government effectiveness	281.6
Regulatory quality	-422.8
GitHub comments / million population aged 15-69	316.5
GERD financed by abroad, % GDP	-231.2
Patent families / billion PPP$ GDP	150.6
Unicorn valuation, % GDP	564.9
E-participation	426.9
Government online services	675.9
Expenditure on education, % GDP	413.5
Graduates in science and engineering, %	248.9
Finance for startups and scaleups	460.0
Policies for doing business	984.4

Table 2: Change in input indicators (logistics performance, policies for doing business, GERD financed by business, finance for startups and scaleups, E-participation, government online service, expenditure on education, patent families, graduates in science and engineering, ISO 14001 environment, ICT use, government effectiveness, regulatory quality, GERD financed by abroad) and ensuing GII rank

% change across input indicators	GII rank
20%	13
30%	8
36%	5
40%	2
50%	1

Appendix 4: News Scraping and Sentiment Analysis

Step	Method
News Scraping	
API	Register for API from https://scrape-it.cloud/
Keyword functions	Create a function that takes in keywords to search Google news. The API returns JSON data which can be extracted from the key "newsResults" to create a dataframe that consists of a title, link, source, snippint, and date of the news result. If the scrape results in a key error, it is likely that the search yielded zero results. A separate function saves data into a .csv with the same name as the search topic.
Loop through years	Create a loops through the years 2016-2024 and countries Canada, USA, Finland, Japan, Netherlands, Singapore, UK, and Germany. This was repeated for keywords "GERD in business" and "STEM graduates".
Additional queries for Canadian news	For Canadian news only, other keywords were queried: "gross capital formation", "ICT use", "utility model", "E-participation", "regulatory index", and "government online services".
Combine dataframes	Data for topics with multi-year searches were combined into one dataframe, any duplicates were removed, and then re-downloaded as a cleaned .csv file.
Sentimental analysis	
Sentiment function	Create a function that leverages sentimental analysis from the TextBlob library.
Data cleaning	Create a function that cleaned up the dates in a dataframe named "date" by extracting only the year part from the date strings. Two version of this function were created to solve inconsistencies in the API's result. The function looped through all rows of the dataframe and combined the news title and news snippet. Sentimental analysis was performed and appended to a list which then returns the results as a data frame.
Visualization and analysis	The sentimental analysis results were plotted as a scatterplot and heatmap. The average sentiment of news was calculated using panda's built-in mean function. Results were grouped by year and plot.

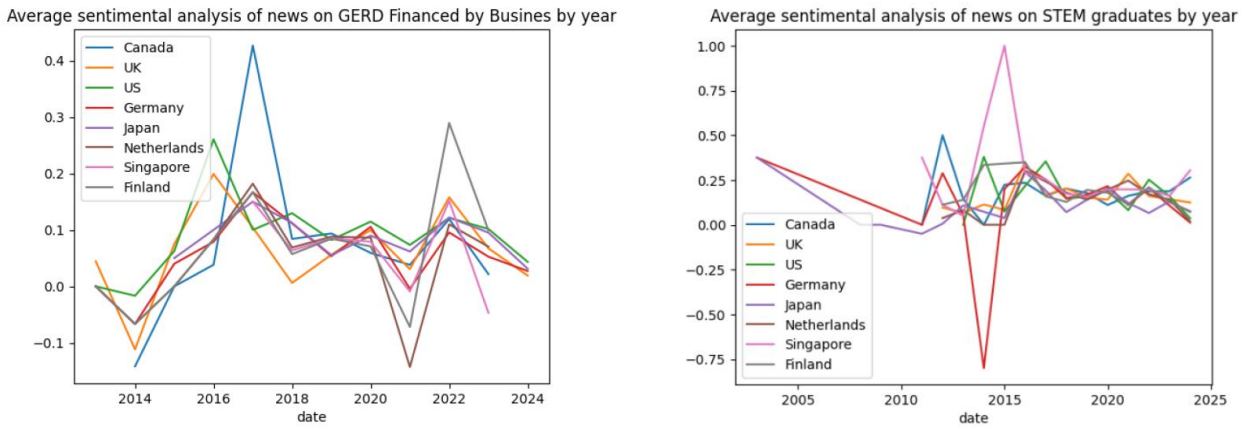

Figure 11: Average sentimental analysis across time for news articles collected from Canada, the UK, the US, Germany, Japan, the Netherlands, Singapore, and Finland, for keywords "GERD business" (left), and "STEM graduates" (right).

Appendix 4: News Scraping and Sentiment Analysis (continued)

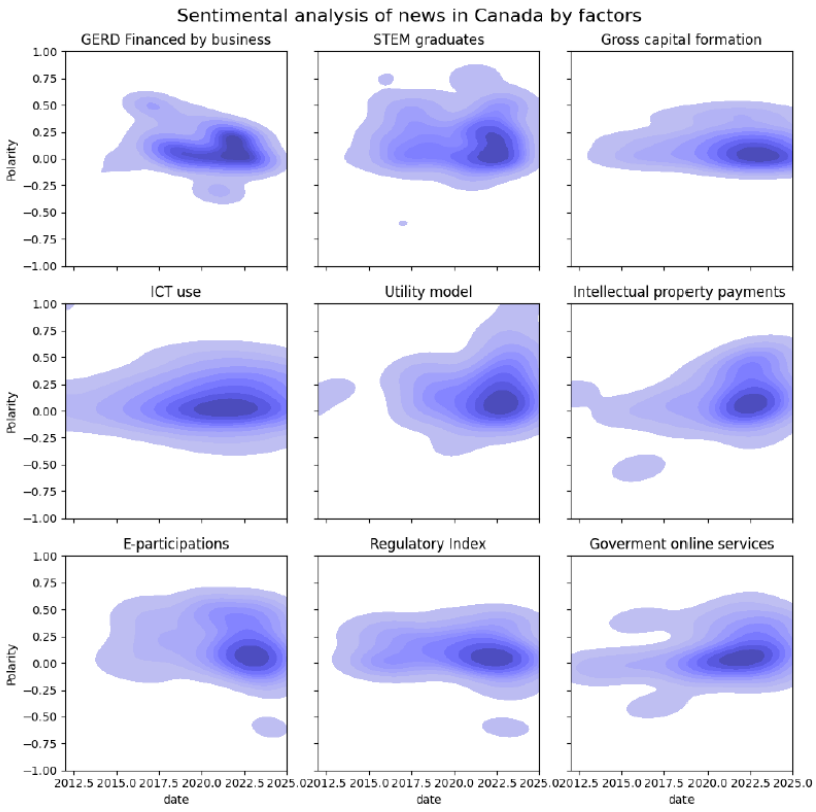

Figure 12: Sentimental analysis heat clouds for keywords found in Canadian news articles

Appendix 5: Frequency Analysis

Step	Method
Extracting PDF information and summarization using ChatGPT API	
The code is function dependent, to deal with a large sum of values and manage without losing uniqueness or hidden data within the keys position. The provided PDFs were manually renamed into their directory country/company for easier access and handling withing the code.	
PDF Text Extraction_A	Utilizes `PyPDF2` to extract text from PDF pages: by continuously defining global variables using a for loop, the PDFs contents were stored within these variables separately.
PDF Text Extraction_B	An additional library is needed to extract the texts from diagrams and charts, utilizing `Crypto.Cipher` from 'pycryptodome' helps `PyPDF2` to extract these texts.
GPT model configuration	Sets parameters for gpt-3.5-turbo-0125 and gpt-4, including output length (`max_tokens`) and required prompt to tailor the summary output.
Main keys extraction	Data for topics with multi-year searches were combined into one dataframe, any duplicates were removed, and then re-downloaded as a cleaned .csv file.
Main keys uniqueness assuring	The extracted keys were combined and fed to a function which would return the top 40 unique points to avoid repetition within the documents.
Canadian documents extraction	The desired output is to find strategies that are not utilized and can suggested for Canada to follow. From the documents published by Canadian Institutes the main keys mentioned were extracted separately, searched for uniqueness, and stored.
Suggested innovation keys generation	Key points from the Canadian documents and non-Canadian documents were fed to function that would compare them and return the top 10 strategies which were not mentioned in the Canadian documents.
Data storing	Each step was stored in a txt file to avoid random generation with each run and stored. Companies' keys strategies and countries' keys strategies were dealt with separately. Different sources had different innovation strategies approach, adding them within the same category would disregard important features.
Quantitative values extraction	
Set unique numbers	Combined countries' strategy keys were given unique numbers that correspond to the country of origin they were extracted from.
Sort into 5 categories	The complete list was fed to a function that would return them sorted into 5 categories (which were predetermined through multiple trials), keeping the numbering system intact.
Test for distinct features	The number of strategies in each category is used to indicate the focus in this category done by the countries in general. The returned text was separated into lines, which were tested through a loop function for distinct features to determine the final number using "re" library.
Sort per country of origin	A dictionary that follows the numbering sequence is used to sort the categories points into their country of origin. The amount was then calculated and normalized to be used as an indication of each where country's main efforts lie.
Infographics Generation	
Dall-E image generation	The combined countries' strategy keys were summarized into a simple text. The summarized keys were fed into a "dall-e-3" as a prompt to generate a bohemian image.