OECD Statistics Working Papers 2023/05

What is the role of data in jobs in the United Kingdom, Canada, and the United States? A natural language processing approach

Julia Schmidt,

Graham Pilgrim,

Annabelle Mourougane

https://dx.doi.org/10.1787/fa65d29e-en

OECD

Organisation for Economic Co-operation and Development

SDD/DOC(2023)5

Unclassified

English - Or. English

13 September 2023

STATISTICS AND DATA DIRECTORATE

What is the role of data in jobs in the United Kingdom, Canada, and the United States? A natural language processing approach

SDD Working Paper No. 119

Contacts: Julia Schmidt (Julia.schmidt@oecd.org); Graham Pilgrim (graham.pilgrim@oecd.org), Annabelle Mourougane (Annabelle.mourougane@oecd.org).

JT03525106

OECD STATISTICS WORKING PAPER SERIES

The OECD Statistics Working Paper Series – managed by the OECD Statistics and Data Directorate – is designed to make available in a timely fashion and to a wider readership selected studies prepared by OECD staff or by outside consultants working on OECD projects. The papers included are of a technical, methodological or statistical policy nature and relate to statistical work relevant to the Organisation. The Working Papers are generally available only in their original language – English or French – with a summary in the other.

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the authors.

Working Papers describe preliminary results or research in progress by the authors and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed and may be sent to the Statistics and Data Directorate, OECD, 2 rue André Pascal, 75775 Paris Cedex 16, France.

This document, as well as any statistical data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The release of this working paper has been authorised by Paul Schreyer, OECD Chief Statistician and Director of the OECD Statistics and Data Directorate.

https://www.oecd-ilibrary.org/economics/oecd-statistics-working-papers_18152031

What is the role of data in jobs in the United Kingdom, Canada, and the United States?

Julia Schmidt, Graham Pilgrim, Annabelle Mourougane

Abstract / Résumé

This paper estimates the data intensity of occupations/sectors (i.e. the share of job postings per occupation/sector related to the production of data) using natural language processing (NLP) on job advertisements in the United Kingdom, Canada and the United States. Online job advertisement data collected by Lightcast provide timely and disaggregated insights into labour demand and skill requirements of different professions. The paper makes three major contributions. First, indicators created from the Lightcast data add to the understanding of digital skills in the labour market. Second, the results may advance the measurement of data assets in national account statistics. Third, the NLP methodology can handle up to 66 languages and can be adapted to measure concepts beyond digital skills. Results provide a ranking of data intensity across occupations, with data analytics activities contributing most to aggregate data intensity shares in all three countries. At the sectoral level, the emerging picture is more heterogeneous across countries. Differences in labour demand primarily explain those variations, with low data-intensive professions contributing most to aggregate data intensity in the United Kingdom. Estimates of investment in data, using a sum of costs approach and sectoral intensity shares, point to lower levels in the United Kingdom and Canada than in the United States.

Keywords: Data intensity, data asset, natural language processing, job advertisements, data economy, United Kingdom, Canada, United States.
JEL codes: C80, C88, E01, J21.

****

Cet article estime l'intensité des données des professions/secteurs (c'est-à-dire la part des professions/secteurs dans la production de données) en utilisant le traitement automatique du langage naturel (TALN) sur les offres d'emploi au Royaume-Uni, au Canada et aux États-Unis. Les données des annonces d'emplois en ligne collectées par Lightcast fournissent des informations ponctuelles et désagrégées sur la demande de travail et les compétences requises dans différentes professions. Ce document apporte trois contributions majeures. Premièrement, les indicateurs créés à partir des données de Lightcast contribuent à la compréhension des compétences numériques sur le marché du travail. Deuxièmement, les résultats peuvent faire progresser la mesure des données dans les statistiques des comptes nationaux. Troisièmement, la méthodologie de TALN peut traiter jusqu'à 66 langues et être utilisée pour mesurer des concepts au-delà des compétences numériques. Les résultats fournissent un classement de l'intensité des données entre les professions, avec les activités d'analyse de données contribuant le plus à l'intensité globale des données dans les trois pays. Au niveau sectoriel, on observe une plus grande hétérogénéité selon les pays. Ces variations s'expliquent principalement par la demande de travail, les professions à faible intensité de données contribuant le plus à l'intensité globale des données au Royaume-Uni. Les estimations de l'investissement dans les données, utilisant une approche de somme des coûts et des parts sectorielles d'intensité, indiquent des niveaux inférieurs au Royaume-Uni et au Canada par rapport aux États-Unis.

Mots-clés : Intensité des données, données, traitement automatique du langage naturel, offres d'emploi, économie des données, Royaume-Uni, Canada, États-Unis.
Codes JEL : C80, C88, E01, J21.

Table of contents

Tables

Figures

Boxes

What is the role of data in jobs in the United Kingdom, Canada, and the United States?

A natural language processing approach

Julia Schmidt, Graham Pilgrim, Annabelle Mourougane[1]

OECD Statistics and Smart Data Directorate

1. Introduction

1. Understanding and assessing the share of jobs involved in data production, here referred to as data intensity of jobs, is critical to better understand the increasingly digital economy. Data collection, processing and analysis skills are becoming more and more relevant across different industries (Sostero and Tolan, 2022[1]). The people who master those skills are in high demand in the labour market and typically earn a wage premium.

2. Identifying the location of data-intensive jobs can support the design of labour and education policies. The increasing use of data and advanced analytics across countries has driven the demand for new types of skills and jobs (Acemoglu and Restrepo, 2017[2]). Employers search for increasingly specialised skillsets, and technologies are changing fast, making traditional competency and occupation taxonomies less useful to grasp these emerging trends in labour markets (Sostero and Tolan, 2022[1]). Insights into the distribution of data-intensive jobs across occupations and sectors can influence spending on job-search support, and more generally employment and social policies. They may also help in understanding structural changes and longer-term transitions in labour markets and the reallocation from the digital and green transition.

[1] This paper benefited greatly from comments, guidance, and support by Sebastian Barnes, Ben Conigrave, Asa Johansson, Joseph Grilli, Polina Knutsson, Simon Lange, Molly Lesher, John Mitchell, Minsu Park, Samuel Pinto-Ribeiro, Paul Schreyer, Lea Samek, Rudy Verlhac and Jorrit Zwijnenburg (OECD), Nikos Tsotros, Borislav Gargov, James Wignall and Akash Kohli (UK DBT), as well as Rosanna White, Aris Xylouris and Berkeley Zych (UK DSIT). It also benefited from discussions at the Working Party for Trade in Goods and Services (WPTGS), the Informal Advisory Group on Measuring GDP in a Digitalised Economy (6-7th June 2023) and the ESCOE Conference of Economic Measurement (2023). The authors would like to thank Virginie Elgrably for excellent support in formatting the document.

The financial support for the research presented in this paper was provided by the Department for Business and Trade, United Kingdom.

3. Information on data-intensive jobs can indicate where to best advance innovation and help policymakers target support to those firms that need it most. Highly productive firms in digital-intensive sectors appear to have advantages when harnessing the digital transformation (Berlingieri et al., 2020[3]). Indeed, firms employing people with data analytics skills grow faster than their less data-intensive counterparts (Harrigan, Reshef and Toubal, 2021[4]), sometimes at the expense of smaller and medium-sized firms, where the internal capacities to adopt digital tools are often limited (OECD, 2023[5]). Understanding data-intensive jobs and the occupations and industries they work in, can also help to monitor trends in market concentration (Schoch, 2020[6]) and mark-up trends in digital-intensive sectors (Calligaris, Criscuolo and Marcolin, 2018[7]).

4. Insights into the data intensity of jobs may improve the design of digital trade agreements and the growing framework of digital provisions supporting digital trade and the free flow of data (López González, Sorescu and Kaynak, 2023[8]). Such statistics can contribute to assessing the competitive advantage of specific industries or regions. They can help to identify professions whose importance is going to grow or those which are going to disappear with digital transformation and assist analysis of barriers affecting the cross-border movement of those professions (OECD, 2022[9]). Finally, as data-intensive jobs may involve the creation and utilisation of intellectual property, the agreements to promote responsible data practices and governance, encourage innovation and foster technology exchange can be further examined.

5. Against this background, it is important to capture the role of data in the measurement of economic indicators (Corrado et al., 2022[10]). The data intensity of jobs is a key input to measuring the economic value of data. While data are often assumed to have become more important in economic production, statistical frameworks are not yet equipped to define and capture the economic value of data adequately. An update of the System of National Accounts (SNA) is expected for 2025, which makes proposals on how to improve the approaches to measuring the value of data relevant and timely. Measuring the extent to which jobs are data-intensive could also aid in better identifying the main drivers of productivity, as the contribution of new technologies, including machine learning and big data analytics, might not be adequately captured in productivity statistics (Brynjolfsson, Rock and Syverson, 2021[11]).

6. This paper contributes in three major ways to existing research and can aid policy analysis beyond the topic of measuring data assets:

- First, it puts forward a novel methodology using natural language processing (NLP) to online job advertisement text from Lightcast (LC), previously Burning Glass Technologies and Emsi Burning Glass, to generate occupation- and industry-level estimates of data intensity. In doing so, it builds on a paper by Calderón and Rassier (2022[12]) who use machine learning to better identify data-related skills clusters from online job advertisements for the United States.
- Second, the methodology can be used to advance cross-country comparable results on measuring the value of data assets in the data economy and the evolution of digital skills in the labour market. Discussions during the Informal Advisory Group on Measuring GDP in a Digitalised Economy in June 2023 signalled great interest in developing a harmonised conceptual approach to measuring data assets. Preliminary estimates using LC data showed that the data source is sufficiently stable, with a coverage from 2012 to real-time.
- Third, the NLP algorithm is flexible and can be applied to concepts that are difficult to capture in traditional labour market statistics, such as green and AI-related jobs. The algorithm can also be adapted to over 66 languages, thus allowing for an inclusion of other countries. However, in such cases, the job advertisement text would either need to be manually web-scrapped or provided by alternative sources.

7. The paper is structured as follows. The next section discusses the main challenges in estimating the data intensity of jobs, relying on the literature. Section 3 presents the methodology and Section 4 the data used. Key insights from the empirical work are detailed in Section 5. Section 6 concludes.

2. Advancing the measurement of data intensity of jobs

8. Earlier estimates of the data intensity of jobs have been derived from occupation classifications, such as O*NET, a large database provided by the U.S. Bureau of Labor Statistics. These estimates rely on identifying jobs where data analysis and processing are central to the work performed, based on a subjective threshold (U.S. Department of Commerce, 2015[13]). Other studies have used O*NET to develop measures of knowledge and work activity related to computers to calculate the digital intensity for occupations in the United States (Muro et al., 2017[14]). Yet, most of these papers acknowledge that these data are limited in their granularity and timeliness.

9. Most recent analyses have combined different data sources. The Asian Development Bank (2022[15]) used a mix of labour market, survey and national accounts data to better understand digital skills in the Asia-Pacific region. Calvino et al. (2018[16]) created a digital taxonomy at the industry level using a wider range of indicators on components of development and adoption of the most advanced "digital" technologies, the human capital needed to embed them in production and the extent to which digital tools are used to deal with clients and suppliers.

10. Job vacancy data have gained in popularity to measure the data intensity of jobs. A first strand of the literature uses those data to create skills taxonomies of emerging professions. Lassébie et al. (2021[17]), for instance deploy a machine-learning approach to map job vacancy data to O*NET and create a taxonomy of emerging skillsets, such as digital skills. Other studies focus on understanding shifts in labour demand and discover a concentration of demand for jobs in digital industries (Garasto et al., 2021[18]).

11. A second strand of the literature focuses on skill requirements for digital professions. Sostero and Tolan (2022[1]) analyse vacancy data from 2012-20 in the United Kingdom and identify clusters related to digital skills. They also find these skillsets at the core of some non-digital domains, like the administrative and clerical cluster. Beblavy, Fabo and Lenaerts (2016[19]) focus on IT skills, while Samek, Squicciarini and Cammeraat (2021[20]) examine the demand for artificial intelligence skills in the labour market.

12. Soh et al. (2022[21]) combine O*NET and job vacancy data in the United States and find that regions that were hit harder by the COVID-19 recession experienced a larger increase in the share of digital occupations in both employment and newly posted vacancies. This was however driven by a smaller decline for the employment and vacancy share of digital workers compared to non-digital workers, and not by an absolute increase in the demand for digital workers. Instead, digital occupations were more insulated from the COVID-19 shock to the labour market. Bellatin and Galassi (2022[22]) find a similar effect for jobs related to digital skills in Canada, using job advertisements.

13. Text mining and natural language processing have helped to exploit the massive amount of data contained in online job advertisements. A range of studies from the computer science literature analyse the algorithms and their performance when extracting information from the raw text of job online advertisements (Boselli et al., 2018[23]; Zhang et al., 2022[24]). Studies in this domain also explored several approaches to classify skills extracted from vacancy data (Sayfullina, Malmi and Kannala, 2018[25]; Tamburri, Van Den Heuvel and Garriga, 2022[26]). NLP approaches have been used to analyse occupation changes in the United States (Rock, Bana and Brynjolfsson, forthcoming[27]), understand skill requirements for managerial positions (Hansen et al., 2021[28]) or skill requirements in emerging job areas (Kortum, Rebstadt and Thomas, 2022[29]). Recent work by the UK ONS and Nesta have explored the wealth of information in job advertisements (Kanders and Sleeman, 2021[30]; Vassilev, Romanko and Evans, 2021[31]).

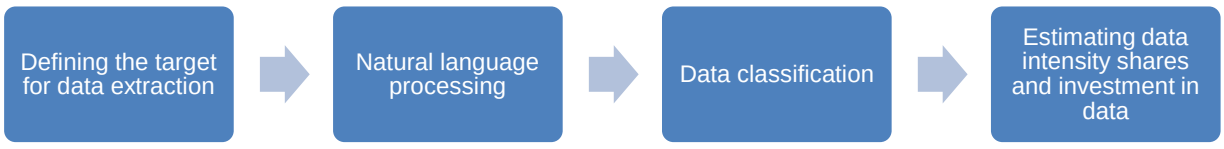
3. A natural language processing approach

3.1. Overview of the empirical approach

14. The approach presented uses NLP to derive the share of jobs involved in data production, referred to as "data intensity", by occupations and sectors (Figure 1.). The main data source is text data retrieved from online job advertisements data provided by Lightcast in the United Kingdom, Canada and the United States. The data provide good coverage of the labour demand in each of the three countries, and allow for insights into the skills and task requirements for each job advertisement (for an extensive discussion of its coverage, representativeness and timeliness see Section 4):

1. **Defining data-intensive jobs:** Job advertisements are a rich source of information to understand the labour demand. The data value chain put forward by Statistics Canada (2019[32]) and Corrado et al. (2022[33]) is used as a conceptual framework to determine whether a job is data-intensive.
2. **Deploying Natural Language Processing:** The text data are cleaned of noise, and quality and consistency checks are deployed to check the properties of the data. Subsequently, the NLP algorithm performs the text feature extraction, which transforms text data into a mathematical object that can be classified.
3. **Classifying data:** The parts of the online job advertisement identified as data-related skills and tasks are classified into data entry, database and data analytics related capabilities to allow for a breakdown by these types of data-related production activities.
4. **Deriving the data intensity by occupation and sector and estimates of investment in data:** In a final step, the data-intensive jobs are aggregated to derive a data intensity share by occupation and sector and matched onto national accounts data to calculate an estimate of investment in data at sector and economy level.

Figure 1. Steps of the empirical approach



Source: Authors' illustration.

3.2. Defining data-intensive jobs

15. In this study a data-intensive job is identified via the skills and tasks required to produce data along a value chain. Data is defined as "*Information content that is produced by accessing and observing phenomena; and recording, organising and storing information elements from these phenomena in a digital format, which provide an economic benefit when used in productive activities"* (ISWGNA, 2022[34]). A value chain describes the sequence of activities that a firm performs to deliver a valuable product to the end customer. Recognising the specific economic characteristics of data, recent papers have established that data value creation involves the application of successive layers of data technologies to generate data assets (Corrado et al., 2022[33]; Statistics Canada, 2019[32]).

16. Building on Corrado et al. (2022[33]), the methodology aims at distinguishing between value creation at three stages of data production and defines three types of data assets, forming a "modern data stack":

- Data entry: These tasks relate to work with raw records that have been stored but not yet cleaned, formatted or transformed for analysis (e.g. data scraped from the web). Raw records also refer to data collected from experiments, statistical surveys or administrative records.
- Databases: Jobs in this category work on transformed raw data, records that have been cleaned, formatted, and structured to be used in using data analytics or visualisation.
- Data analytics: These tasks reflect jobs using advanced tools to analyse data (e.g. machine-learning algorithms).

17. Advantages of the data value chain framework lie in its complementarity to the existing definition of databases in the System of National Accounts. It also creates a clear distinction between the different stages of data production, commonly used in the industry to structure data processes and aligns with the premise that data analytics related jobs create a higher value (linked to a higher wage premium) compared to data entry related jobs.

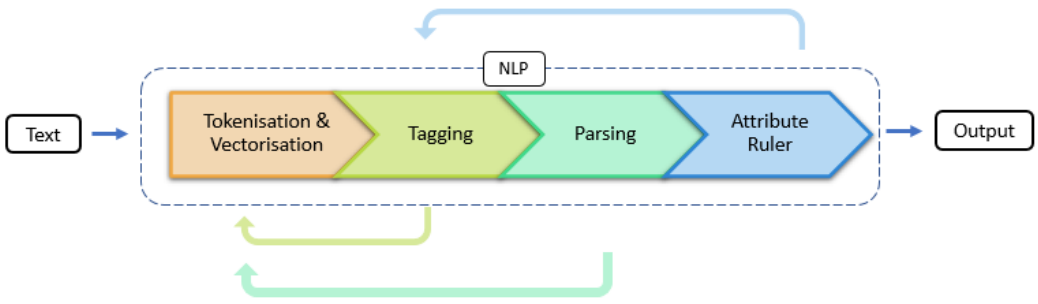
18. Despite these advantages, the framework falls short of differentiating between activities along the value chain that are more rival, and potentially more 'valuable' with regards to firm profitability, and activities that are non-rival. For example, data have potential economic value if widely shared, yet many firm operations involve exclusive data access due to privacy concerns (Corrado et al., 2022[33]). These externalities have not been factored in. Moreover, while the concept of the data stack assumes that more value is added as data move from data entry to data analytics stages, it is not clear which layer of the data stack is the most essential to the economy, and whether these differences exist in all sectors or are subject to changes in labour demand.

3.3. Constructing a measure of data intensity at occupation and sector level

Deploying an NLP algorithm

19. This paper uses NLP techniques to extract relevant information from online job advertisements and construct a measure of data intensity at occupation and sector level. It relies on an open-source NLP pipeline provided by the 'spacy' python library (spaCy, 2022[35]). The pipeline combines different NLP models to efficiently perform advanced text processing operations in an iterative fashion (Figure 2). The output of the pipeline can be used for tasks such as text classification or analysing phrase frequencies.

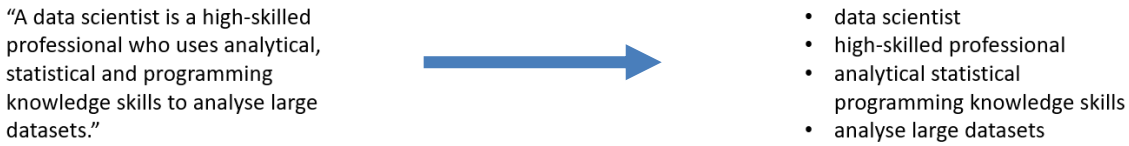
Figure 2. A natural language processing pipeline



Source: Authors' illustration based on (spaCy, 2022[35]).

20. Tokenisation and vectorisation are the most relevant steps to the approach used in this paper. Tokenisation is the process of breaking down a text into smaller units, called tokens. These tokens can be words, phrases, or even individual characters, depending on the specific method used. In this paper, text from the job advertisements is processed and subsequently split into noun chunks (visible on the right in Figure 3). This allows to harmonise the text for each job advertisement, as the chunks are consistent in their length and converted into lower case. It also automatically removes stop words (e.g. "the" or "and") and noise, such as special characters. This step is usually the first step in an NLP process and takes information from functions in the pipeline that identify the types of words (tagging), capture their grammatical structure (parsing) and customise rules for specific words in a sentence based on their characteristics (attribute ruler).

Figure 3. Example of a tokenisation process



Source: Authors' illustration.

21. In a second step, called vectorisation, the pipeline transforms text into numerical vectors (embeddings). It is possible to represent each noun chunk as a vector of word frequencies or word embeddings (see Box 1). One advantage of the spacy pipeline is the availability of pre-trained models to generate embeddings that capture the meaning of the text more efficiently. The specific model has been pre-trained on large amounts of text data and can produce high-quality representations for a wide range of NLP tasks as shown in previous research (Carrasco and Rosillo, 2021[36]; Reddivari and Wolbert, 2022[37]). This saves time and computing power and allows for a quick deployment of spacy models on a variety of NLP tasks. The resulting embeddings can be used to perform mathematical operations on the text, such as similarity measures. The outcomes of the pre-trained model were checked on a validation set of over 10 000 manually classified job advertisements.

Box 1. Embeddings – Creating a numerical representation of a vector

An embedding is a numerical representation of a word using a vector. The process of generating a vector for a text object is called vectorisation. Each word is assigned a unique vector that encodes its meaning and relationships with other words. The position and orientation of the vector in the multi-dimensional space capture the semantic similarities and differences between words. In this application, vectors are generated by a machine-learning model, and typically have 300 dimensions. The dimensions of the vector depend on the machine-learning model chosen that generates the vector (Jurafsky and Martin, 2023[38]).

How are embeddings created?

Embeddings are generated such that two text tokens with similar linguistic usage will have vectors that are close to each other, i.e. the distance between them in vector space is small. However, the features of the embedding no longer have any meaning. In this paper, the spacy neural network model uses a machine-learning component "tok2vec" that learns how to produce vectors for each token to generate the embeddings (spaCy, 2023[39]).

The example below shows vectors for the noun chunks "data analysis", "data analytics" and "your information". The generated vectors for "data analysis" and "data analytics" are almost identical as the words represented are very close to each other compared to the noun chunk "your information".

```
Data analysis    = [1.5, -0.4, 7.2, 19.6, 3.1, ..., 20.2]
Data analytics   = [1.5, -0.4, 7.2, 19.5, 3.2, ..., 20.8]
your information = [7.5, -1.0, 7.2, 14.8, 2.8, ..., 19.0]
```

What are they used for?

Converting words into vectors can support mathematical operations, such as calculating the similarity between two vectors. A common measure used is cosine similarity, defined as the inner product of two vectors (x and y) divided by the product of their length (Equation 1).

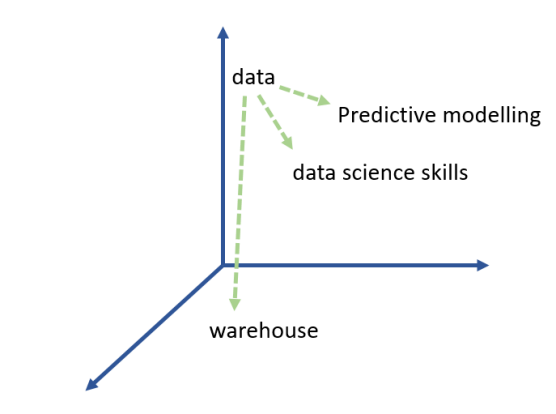
Equation 1: Cosine similarity measure

$$Cos\ (x, y) = \frac{x\ .\ y}{||x|| * ||y||}$$

x	vector representation of noun chunk (x)
y	vector representation of noun chunk (y)

When the similarity score is 1 then the two vectors are similar, or 0 then the two vectors are orthogonal. Geometrically, cosine similarity measures the cosine of the angle between two vectors projected in a multidimensional space (Figure 4). The smaller the angle, the higher the similarity (Kotu and Deshpande, 2019[40]).

Figure 4. Illustrating similarity measures



Source: Authors' illustration.

Classifying noun chunks into data entry, database and data analytics activities

22. After the processing by the NLP pipeline, the chunks are classified as data entry, database and data analytics activities based on three criteria, namely their similarity to the term "data", their dispersion across occupations and the frequency within certain landmark occupations (Box 2). The criteria chosen were selected from a set of measures commonly used in NLP research and to ensure estimates are robust to small changes in thresholds (see below).

Box 2. Explaining the classification rule

Noun chunks are classified as related to data entry, database and data analytics related skills and tasks based on three distinct measures:

- ***Cosine similarity measure*:** The vectors specific to each noun chunk and created by the machine learning model (spaCy, 2023[39]) are used to calculate how similar the chunk is to the target word 'data'. The threshold for a specific chunk to be data intensive is set at 50%, a threshold widely applied in the existing NLP literature indicating that the words are similar to each other (Crocetti, 2015[41]; Manning and Schutze, 1999[42]).
- ***Dispersion measure*:** The noun chunk must be specific to a few occupation classes in the dataset to be labelled as data intensive. This ensures that the word is not widely used, such as common phrases e.g. 'your skillset' or 'your personal data', but instead likely to be a specific term only found in certain job advertisements, such as 'predictive modelling'. The dispersion measure describes the frequency of a noun chunk in one occupation relative to the frequency of the same noun chunk in all job advertisements (Equation 2). The parameter is chosen based on a threshold x_c to adjust for biases in the sample of text data and avoid that the number of total noun chunks, which significantly varies across countries, drive the classification results of the algorithm.

Equation 2 Relative frequency measure

$$rel\_freq_n = \frac{Count\_by\_n\_occ/Count\_by\_occ}{Count\_by\_n/Count\_Total} > x_c$$

n	noun chunk
occ	occupation class
x_c	data-specific threshold

- ***Occurrence measure*:** The noun chunks need to appear in one of three landmark occupations (namely data entry clerk, database administrator and data scientist) to ensure they describe either data entry, database or data analytics skills and tasks. The respective landmark occupations are chosen based on existing research (Statistics Canada, 2019[43]; Calderón and Rassier, 2022[44]).

A noun chunk is classified as data intensive, if all three criteria of the classification rule are met. This is specified for data entry, database and data analytics respectively. The criteria for the three types of data-related roles are independent from each other, meaning noun chunks identified in one landmark occupation do not overlap with the others.

Classification rule

Noun chunk is data intensive IF:

Cosine similarity	> 0.5	AND
$\text{relative\_frequency}_{\text{noun chunk}}$	$> x_c$	AND
noun chunk	$\epsilon\ landmark$ (= data entry clerk, database administrator and data scientist)	

ELSE: 0

3.4. Deriving data intensity shares and estimates of data investment

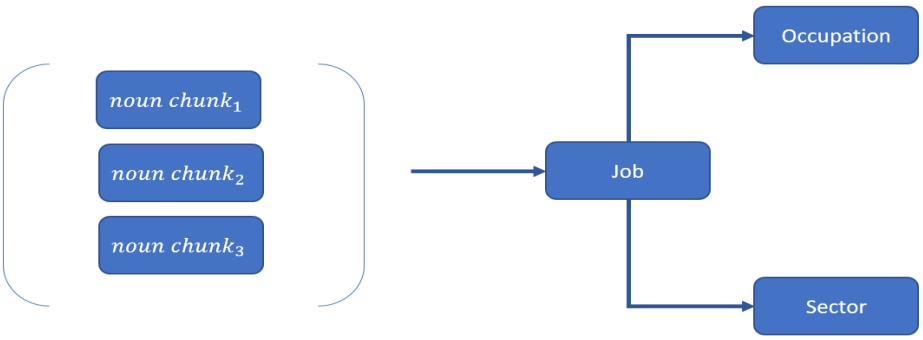
Aggregating from noun chunks at job level to occupation and sector level

23. The classified noun chunks are used to construct an indicator of data intensity at occupation and sector level (Figure 5.).

24. In a first step the noun chunks are classified to identify whether a job is data intensive or not, and to what extent the data intensity is driven by data entry, database, and data analytics skillsets. If more than three noun chunks meet the classification criteria, the job advertisement is considered as a data-intensive job (1); otherwise, it is considered as non-data-intensive (0). The breakdowns into data entry, database and data analytics activities are expressed as shares of total chunks classified and always sum to 1.

25. In a second step, the jobs are aggregated by occupations (SIC, NOC and NAICS respectively), and sectors (ISIC version 4) using a weighted average (by the count of jobs per occupation/sector) to calculate the share of data-intensive jobs in each grouping (for an illustrative example, see Annex A). Throughout all stages of aggregation, data integrity checks were implemented to ensure the consistency of classifications and the appropriate treatment of missing values in the data.

Figure 5. Creating data intensity measures at occupation and sector level



Source: Authors' illustration.

Deriving estimates for the value of data assets

26. While data are often assumed to have become more important in economic production, measuring their value has proven challenging. Unlike other goods and services, data are rarely traded in markets (OECD, 2022[45]; Koutroumpis, Leiponen and Lleweellyn, 2020[46]). Estimates of the value of these assets can be derived from their production costs, the sum of costs approach, using the data intensity as an input.

27. The valuation of data is complicated by its characteristics (Coyle et al., 2020[47]; OECD, 2022[45]). Data are a non-rival good, which allows usage by multiple parties simultaneously without exhausting the asset. However, technical and legal restrictions can make access to data excludable (Jones and Tonetti, 2020[48]), such as in the case of physicians where they are free to access their patients' data, but the patients' employers are not usually permitted to do so. While non-rivalry is inherent to data, ownership or use rights over data are typically not very well defined. In the absence of a governance framework that encourages shared use, current technologies make it easy to exclude others from their use.

28. Statistical frameworks are not yet equipped to define and capture the economic value of data adequately. The current Systems of National Accounts (SNA 2008) contains no definition of "data", but records the costs related to the creation of databases. This includes cost spent on data processing, human resources, capital, and intermediary consumption, but excludes the costs for the actual data contained in a database (van de Ven, Ahmad and Schreyer, 2018[49]). Furthermore, the SNA 2008 proposes an "inconsistent" treatment of data assets depending on whether they are for own use or for sale (Rassier, Kornfeld and Strassner, 2019[50]). In the process of updating the SNA, expected for 2025, the recommendation to incorporate data into the production and asset boundary has been endorsed (UNSTATS, 2023[51]), which makes proposals on how to improve the measurement of the value of data relevant and timely (OECD, 2020[52]).The creation and use of data in the absence of observable market transactions have resulted in a variety of empirical approaches to approximate their value (Coyle and Manley, 2022[53]; Corrado et al., 2022[10]).

29. For most assets the best representation of their value is their price on an open market. However, data assets are usually produced in-house and, currently, there exist very few open and observable data markets to estimate the value of data with precision, despite some attempts in the literature using limited observations (Koutroumpis, Leiponen and Lleweellyn, 2020[46]). Even if some datasets are traded on markets, using that information to value datasets in general is difficult as data are highly heterogenous and there is no registry of data. Hence, market-based valuations are not a reliable option at the moment.

30. A second approach focuses on the market capitalisation of firms (Coyle and Li, 2021[54]; Ker and Mazzini, 2020[55]). However, this approach is unsuitable as it likely overestimates the data value by

including firms' reputation and other intangible assets due to double-counting (Baruch and Feng, 2016[56]; Mitchell, Ker and Lesher, 2021[57]). A third approach derives value from the income the asset will generate in the future. Such income-based approaches are useful in cases where the future value can be predicted. For data however, the uncertainty around data future use and value affects predicted future income streams, making this approach unsuitable.

31. Against this background, a consensus has emerged in the statistical community to use the sum of costs approach. In the SNA (2008) the sum of costs approach is used to measure the output of goods and services provided for free (or at insignificant prices) by government or collective non-market output (e.g. social security), or the production retained by the producer for its own final consumption or capital formation.

32. The calculation presented in this paper follows this consensus and is in line with Statistics Canada (2019[43]) and Rassier, Kornfeld and Strassner (2019[50]). Investment in data is calculated as the product of a mark-up (α) capturing non-wage cost and capital service margins (see Section 5.4), the employment compensation specific to sector i and the share of data-intensive jobs in that sector (Equation 3). It is calculated separately for data entry, database, and data analytics categories (d) which are then summed to obtain a value of total investment in data.

Equation 3. Sum of costs approach to calculate investment in data assets

$$investment_{d,\,i} = \alpha * compensation\ of\ employees_i * \frac{number\ of\ data\_intensive\ jobs_d\ in\ i}{number\ of\ jobs_d\ in\ i}$$

d	type of job (data entry, database or data analytics)
i	industry
α	mark-up (non-wage cost and a margin for capital services)

33. This approach has some limitations. Firstly, it is difficult to determine the intermediate consumption and the capital costs that are needed to generate data assets. A lack of harmonised reporting standards and inadequate valuation mechanisms compound these difficulties.

34. Secondly, the approach is highly sensitive to the type of occupations classified as contributing to data production, and the assumed time spent on these production activities, referred to as "time-use factor". Previous studies have relied on subjective expert judgement to estimate this time-use factor (Statistics Canada, 2019[32]). Whilst this may be feasible for some occupations, deriving a realistic time-use factor for all data related activities is challenging. As job advertisements do not provide insights into the actual time spent on a certain task, this study derives a data intensity share, instead of a time-use factor, based on the frequency of skills requirements and task references contained in job advertisements (see Section 3.3). The advantage of this approach is that data intensity shares allow to identify data-related tasks and skills even in industries or occupations typically associated with non-data related roles.

35. Finally, to achieve comparability across countries, national accounts data on employment compensation at sectoral level are chosen rather than labour market data that account for differences in salaries between data and non-data-intensive occupations. As a result, the equation above assumes uniform wages within sectors, as opposed to data jobs earning a premium.

36. Overall, the results derived using this approach are likely to underestimate the actual economic value of data. On the one side, the use of job advertisements as a proxy of labour demand may lead to an upward bias in the data intensity levels, as it neglects the existing employee stock working in data or non-data related production activities. At the same time, the calculation does not consider wage premia of data intensive workers, likely underestimating the contribution of high data intensive jobs. More generally, the sum-of-cost approach does not consider any value created by data use and is thus regarded as conservative approach in measuring the economic value of data.

4. Using online job advertisement data

37. Online job advertisement data are increasingly used as a proxy of job vacancy and of labour demand. They only represent part of the total jobs advertised, although the share of jobs posted online has been growing particularly since the start of the COVID-19 pandemic (Strohmeier, 2020[58]). Additionally, a single job announcement can be used to fill in several vacancies or not filled at all. It is also difficult to infer a churn rate, that is how often a firm needs to repost a specific type of job as employees may move on quickly from their post (Tsvetkova et al., forthcoming[59]; Cameraat and Squicciarini, 2021[60]).

38. Despite these limitations, online job advertisements provide timely and granular information. In many cases, they have distinct advantages over official labour market data, such as O*NET and are often complementary to official employment or vacancy rates collected via surveys (Box 3). Job advertisement data allow close to real-time monitoring of labour market developments and provide details on skills requirements. Furthermore, if treated properly, the data can be linked at a geographical level and firm level, albeit with heterogeneous quality and coverage (Cameraat and Squicciarini, 2021[60]; Lancaster, Mahoney-Nair and Ratcliff, 2019[61]; Tsvetkova et al., forthcoming[59])

39. The online job advertisements data are provided by Lightcast (LC), a private data provider previously known as Emsi Burning Glass or Burning Glass Technologies. They gather job postings from close to 40,000 online sources, such as job boards, employer sites, newspapers, and public agencies using web scraping techniques. LC covers Australia, Canada, New Zealand, Singapore, the United Kingdom and the United States as well as EU countries. The data are available annually for 2012-present (May 2023 at the time of writing) and, since 2020, in monthly and weekly formats. Data for 2007, 2010 and 2011 are available only for the United States. The data for EU countries is available from 2018 onwards. In 2020, the reference year for this study, the dataset includes 6.4 million job advertisements for the United Kingdom, 1.3 million for Canada, and 36.4 million for the United States.

40. The LC data provide one unstructured text entry per job advertisement available in HTML format, that can be broken down into noun chunks. Each job advertisement has on average 140 noun chunks, with slight variation across countries and is identifiable with a unique identification number (Job ID). The text entry can be matched with variables of interest from the structured LC dataset (e.g. occupation class, sector etc.). This means, each job advertisement has a unique Job ID, and can be linked to an occupation class and a sectoral class, depending on data coverage. For the United Kingdom, around 700 occupation classes (SIC 2010) and 21 sectors (ISIC 4 level) exist for 6.4 million job advertisements. LC processes the web-scraped advertisements to avoid counting the same posting multiple times or counting the same job advertised on multiple sites. As one of the key contributions, this study operates on the raw text format next to utilising the structured data to estimate data intensity at occupation and sector level.

41. However, the occupational and sectoral distributions provided by LC come with shortcomings. Announcements which can match several occupation classes are more likely misclassified (Tsvetkova et al., forthcoming[59]). Cultural or institutional differences in the way those job postings are drafted may also make it harder for the algorithm to classify certain advertisements correctly (Devlin et al., 2018[62]).

42. Existing studies show that LC data can differ from official data (Carnevale, Jayasundera and Repnikov, 2014[63]; Hershbein and Kahn, 2018[64]; Cameraat and Squicciarini, 2021[60]).

43. At the occupation level, Tsvetkova et al. (forthcoming[59]) show that, in Canada, professions working in business, finance and administration, management as well as natural and applied sciences are over-represented for 2016, 2019 and 2022, while sales and service professions as well as manufacturing and utilities are the most under-represented sectors. Cameraat and Squicciarini (2021[60]) found that for the United Kingdom, the representativeness of LC data is very good, for 2019, with relatively higher numbers of job openings in relation to “Managers”, “Professionals”, and “Technicians and associate professionals”, as compared to lower skilled occupational groups. In the United States, the overall representativeness in 2019 is lower than in the United Kingdom. US job postings related to “Professionals”

emerge as the largest group, whereas the group “Skilled agricultural, forestry and fishery workers” accounts for the smallest number of job adverts.

44. At the industry level, the representativeness of LC data is broadly consistent with official data, yet differences across countries exist (Tsvetkova et al., forthcoming[59]). Representativeness of the most under-represented sectors in all three countries has improved between 2016 and 2022, albeit based on industry groupings slightly diverging from the ISIC Rev4. standard applied in this study.

45. In the version of LC used in this study, the coverage is uneven, with two-digit industry codes only available for half of the United Kingdom dataset, compared to 60% for the United States and 70% for Canada (Table 1.). Job advertisements with missing classifications at occupational and sectoral level have been dropped. At the occupational level, the sample is assumed to be representative, considering the very few missing values (0% for the United Kingdom, 4% for Canada and 5% for the United States). At the one-digit industry classification, the use of weighting factors generally ensures representativeness, except for lower-level classifications for which this treatment could not be applied.

46. For the United Kingdom, education, health, public administration and safety services are the most over-represented sectors, while the most under-represented sectors are accommodation, foods, arts and recreation, and trades, transport and warehousing. For digital sectors such as information, media, and telecommunication, finance and insurance as well as professional, scientific and technical activities and administrative and other services, differences between the data sources are small. In Canada, finance, insurance and real estate are the most over-represented sectors, while accommodation foods, arts and recreation; construction as well as professional, scientific, technical, administrative and other services are under-represented. Representativeness in the remaining sectors is high. In the United States, the most over-represented sectors are public administration and safety services, while the most under-represented sectors are natural resources and information, media and telecommunications.

Table 1. Share of missing values in Lightcast data

Per cent, 2020

	United Kingdom	Canada	United States
Job ID	0	0	0
National occupational code	0	5	4
LightCast occupational group	9	9	7
Industry code - two digits	47	32	40
Industry code - three digits	60	45	40
Industry code - four digits	62	51	44
Industry code - five digits	94	77	70
Industry code - six digits	na	77	71

Source: Authors’ calculations based on the Lightcast data.

Box 3. Reviewing the use of online job advertisements from Lightcast

Data on online job postings represent a new source of information on labour demand in the economy. Evidence shows that the use of online portals to seek employment has risen in the past decade in Europe and the United States (Vermeulen and Amaros, forthcoming[65]; Sostero and Tolan, 2022[1]; Soh et al., 2022[21]). Many countries lack statistics on labour demand that are sufficiently up-to-date and disaggregated across regions, sectors and occupations. Web-scraped data from online job postings can provide further insights on labour market trends. Using such data requires a careful evaluation of their advantages and disadvantages, depending on the objective of the analysis (Table 2).

This study uses Lightcast (LC) data, a private sector data provider, whose data is being increasingly used to inform policy analysis (Cameraat and Squicciarini, 2021[60]; Grabner and Tsvetkova, 2022[66]; Squicciarini and Nachtigall, 2021[67]; Samek, Squicciarini and Cameraat, 2021[20]; Borgonovi et al., 2023[68]).

The LC online job advertisements offer advantages, especially considering the lack of granular and timely statistics on labour demand. LC data span from 2012-present (May 2023 at the time of writing), providing the raw text of online job advertisements in a weekly frequency from 2020 onwards. For instance, O*NET data, while widely recognised for its completeness and coverage of occupations and skill requirements, are not as timely, with the latest version being updated in 2019.

Another key advantage of LC data is their granularity. They are linkable at firm- and regional level and include standard occupation and industry classifiers. In addition, LC data provide the raw text of online job advertisements. This feature of the LC data enables a skills extraction for data-related jobs specifically, contrary to classic occupation taxonomies such as O*NET (Lassébie et al., 2021[17]).

Nevertheless, LC data also present limits in representativeness across occupations and sectors. The representativeness at occupation level strongly varies over time and country. At the industry level, the differences are broadly consistent with official data sources for the countries analysed in this study (Tsvetkova et al., forthcoming[59]). In European countries, large differences between countries, regions, sectors and occupations remain when validating the data with official vacancy data from Eurostat. For Netherlands, Germany and Sweden, the LC data align most with official sources (Vermeulen and Amaros, forthcoming[65]).

Table 2. Main advantages and disadvantages of using Lightcast data

Advantages	Disadvantages
Timely data (2012-present)	Country coverage is limited (United Kingdom, Canada, United States, New Zealand, Australia, Singapore as well as EU countries at the time of writing)
Linkage to firm-level and regional data	Limited coverage depending on year and country; no insights on how firms hire (e.g. churn rate)
Standardised occupation and industry classifications	Representativeness is heterogeneous (depending on country, sector, region)
Identify skill demands beyond standard labour market statistics	

Source: Authors' compilation.

5. Data intensity in the United Kingdom, Canada and the United States

5.1. Data intensity by occupation is broadly similar across the three countries

Ranking occupations by data intensity paints a consistent picture across the United Kingdom, Canada, and the United States

47. Digitalisation changes the type of skillsets companies seek, and the way people work. The presented methodology allows to identify the type of occupations that are involved in data production, the degree to which they are data-intensive and the type of data-related activity they are involved in.

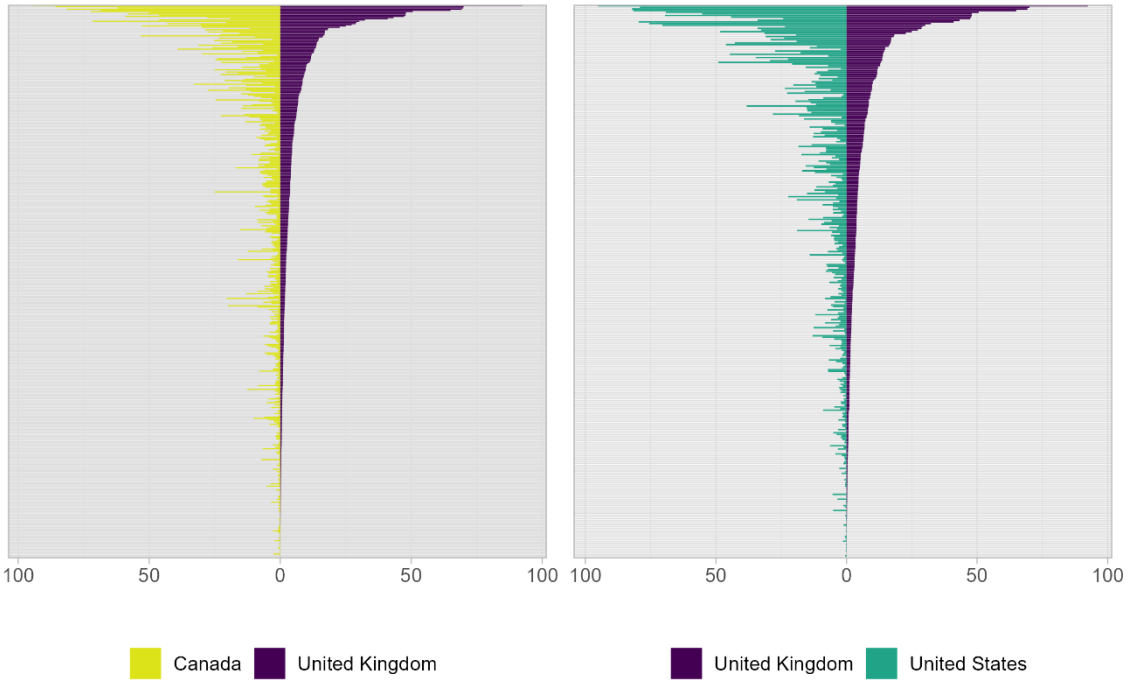
48. The ranking of occupations according to their share of data-intensive jobs is similar across all three economies (Figure 6.). The Spearman rank coefficient is around 80% for the comparison of rankings between the United Kingdom and Canada, as well as the United Kingdom and the United States, suggesting a strong positive relationship (Figure 7).

49. Minor differences occur when zooming into high data-intensive occupation classes. The top ten data-intensive occupations are very similar in all three countries for most professions, yet the highly data-intensive professions in Canada and the United States are more data intensive than their respective British counterparts. A statistician for instance is estimated to have a data intensity of 50% in the United Kingdom, compared to close to 70% in the United States and Canada.

50. Differences are more pronounced for low data-intensive occupation classes (Figure 6). Those in the United Kingdom are often less data-intensive than in Canada and the United States. For instance, a drill operator is 2.4% data intensive in the United Kingdom and in the United States, compared to 12.4% in Canada. Similarly, a librarian is 2.9% data intensive in the United Kingdom, compared to 8.6% in Canada and 7.3% in the United States. The largest differences exist between specific occupations, such as mathematicians, which are highly data-intensive in the United States (around 80%), compared to only 40% in the United Kingdom and 20% in Canada. Similarly, social science researchers are more data-intensive in Canada (50%), compared to 25% in the United Kingdom.

Figure 6. Data intensity across occupation classes

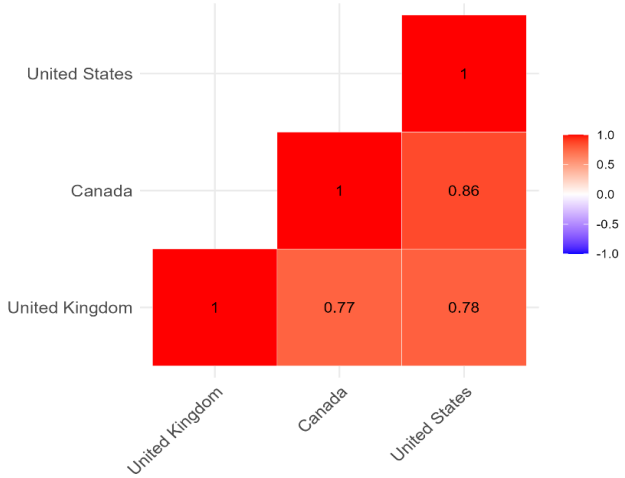
Per cent of labour demand, 2020



Note: The chart shows the data intensity for each occupation class in the United Kingdom, Canada and the United States. Data intensity takes values between 0 and 100. The y axis is sorted by the descending data intensity of the United Kingdom.
Source: Authors' calculation based on Lightcast data.

Figure 7. Spearman rank coefficient of data intensity across occupations

2020



Note: The Spearman rank coefficient is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function. The coefficient takes values between +1 (positively correlated) or −1 (negatively correlated).
Source: Authors' calculations based on Lightcast data.

51. Estimates of data intensity at occupation level largely align with existing studies. According to Statistics Canada (2019[32]), the data entry clerk is considered 100% data-intensive, followed by the database administrator (90-100%) and the financial quantitative analyst (60-70%). In the medium range, the economist and social policy researcher were placed at 20-30% data intensity. The results obtained through the NLP approach used in this study align with Statistics Canada's ranking but point to consistently lower levels of data intensity. For example, data intensity of the database administrator is estimated at 73%, higher than those of the data entry clerk at 59%, while those of the social policy researcher is estimated lower at 53%. This difference from the Statistics Canada approach may be due to the fact that the set of skills related to data production are less consolidated (Sostero and Tolan, 2022[1]). In addition, employees in these professions, despite being highly involved in data work, may also have other responsibilities and tasks to attend to.

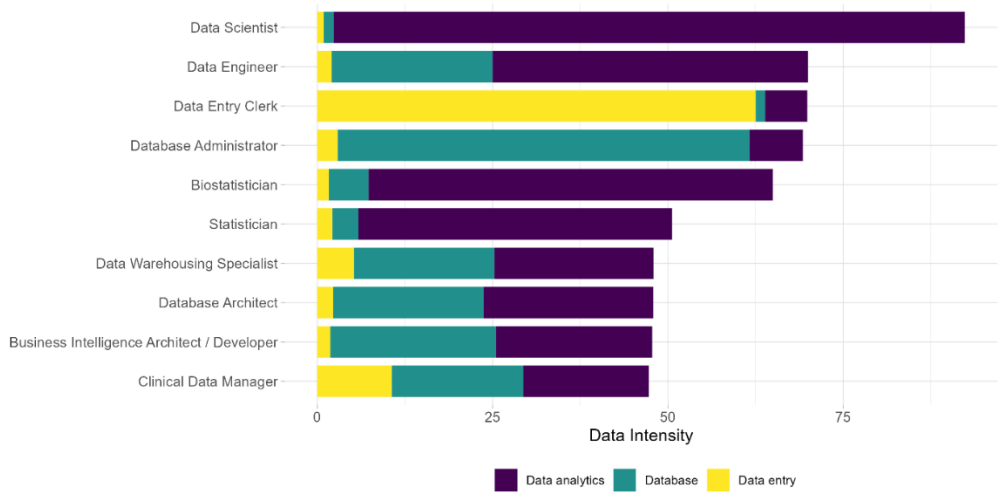
52. Estimates in this paper are also broadly consistent with the findings of Caldéron and Rassier (2022[44]), who used machine learning on LC skills clusters to estimate data intensity for occupations in the United States and with Sostero and Tolan (2022[1]), who analyse the evolution of digital skills profiles across occupation classes in the United Kingdom.

The most data-intensive occupations are linked to data analytics skills

53. In the United Kingdom, the occupation with the highest level of data intensity is the data scientist, with a rate of 92.3%. Following closely are the data engineer at 69% and the data entry clerk at 68% (Figure 8.). Most of these occupations primarily revolve around data analytics skills, although there are a few exceptions. For instance, the data entry clerk and the database administrator exhibit data intensity primarily linked to respectively data-entry and database-related capabilities. In general, the highly data-intensive occupations tend to be specialised, technology-oriented professions, with occupations like biostatistician and clinical data manager showing connections to fields such as medicine and biology. This finding holds when aggregating the occupation classes along the SOC classification in the United Kingdom (Box 4) and similar trends are observed in Canada and the United States (see Annex B, Figure B.1 and Figure B.2).

Figure 8. Data intensity at occupation level in the United Kingdom is linked to data analytics skills

Per cent of labour demand, 2020



Note: The Lightcast data provide occupation classifications. Data intensity takes values between 0 and 100.
Source: Authors' calculations based on Lightcast data.

Box 4. Data intensity by SOC groups in the United Kingdom

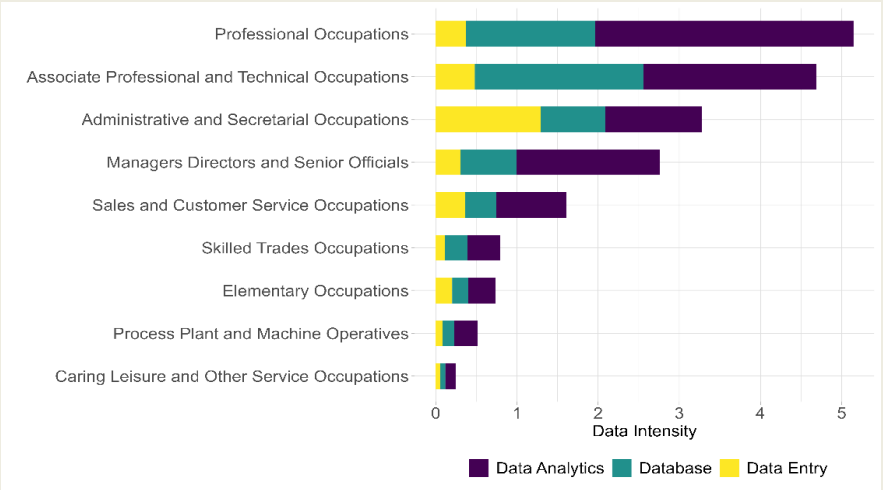
The LC dataset includes variables on national classification systems for each job advertisement, such as the Standard Occupation Classification (SOC) in the United Kingdom. This enables insights into jobs that are similar in terms of qualifications, training and skills. The SOC is used in the United Kingdom to classify and organise job titles into different occupational groups based on the skills, knowledge and tasks required for each role. The SOC classification system is hierarchical, with four levels of detail: major group, sub-major group, minor group, and unit group. Each level provides more details about the occupations included in that group. The SOC system is regularly updated to reflect changes in the labour market (ONS, 2023[69]).

Figure 9. shows the aggregate data intensity for the nine UK SOC major groups. The group of professional occupations for instance, is overall only 5% data intensive, as the figure displays the aggregate share of data-related job advertisements divided by the labour demand in this specific group. Professional occupations, and associate professional and technical occupations, both highly specialised occupations with graduate education requirements, tend to be most involved in data production activities. Administrative and secretarial occupations, as well as managers, directors, and senior officials, the occupation class with the highest qualification requirements, appear fairly data intensive. Elementary occupations, on the other hand, the class with the lowest requirements for general education, are more data intensive than caring, leisure and other service occupations. Sostero and Tolan (2022[1]) find very similar results in analysing digital skills for the United Kingdom.

While the professional occupations are mostly linked to data analytics and database related skillsets, associate professional occupations are both database and data analytics centred. Administrative, professional and secretarial occupations, on the other hand, display the highest share of data entry related skills. The remaining occupation classes are mostly driven by data analytics components. Jobs working in elementary occupations, plant, process, and machinery, and caring, leisure, and other service occupations appear to have equal shares of all three components.

Figure 9. Specialised occupations have a high level of analytical skills in the United Kingdom

Per cent of labour demand per major occupation group, 2020



Note: Data intensity takes values from 0 to 100. The data intensity at major group level is low as it is the mean of data intensive jobs compared to the total jobs in this specific group, which are much higher than those of sectors. As the result, the order of magnitude of the data intensity of occupations is lower than those of sectors, closer to the economy-wide data intensity. As data intensity shares are aggregated along the SOC system, the data intensity shares of the groups approach the total data intensity of the economy.
Source: Authors' calculations based on Lightcast data.

5.2. Data-intensive jobs exist in almost every sector, but their importance varies

Technology related, service-oriented as well as resource-intensive sectors are among the most data-intensive industries in the United Kingdom

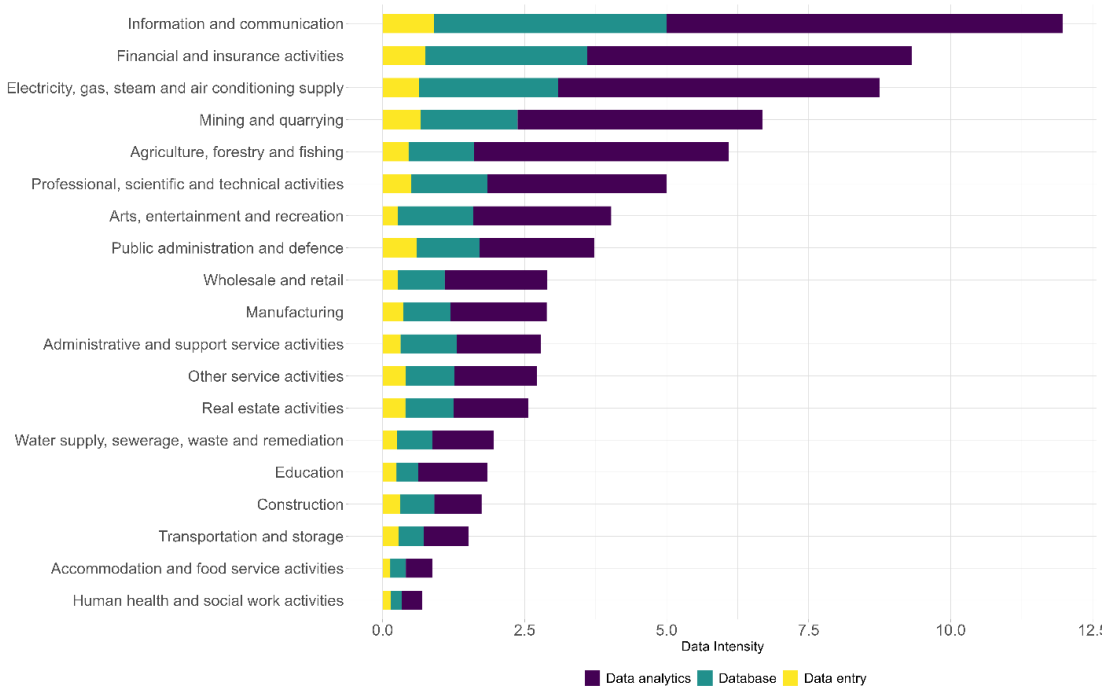
54. About one in ten job advertisements are data intensive in ICT and financial and insurance sectors in the United Kingdom. Those are the most data-intensive sectors, with data intensity shares at 11.9% and 9.3% respectively, which likely reflects the presence of several multinational corporations (MNEs) (OECD-UNSD, 2023[70]). Shares are lower but still significant in electricity, gas steam and air conditioning supply (8.8%), mining and quarrying (6.7%). Agriculture, known to be labour intensive and highly mechanised in the United Kingdom by European standards (UK Department for Environment Food and Rural Affairs, 2022[71]), is estimated to be relatively highly data intensive, with 6.7% of the jobs being involved in data production activities.

55. Most service-oriented sectors in the United Kingdom have a medium to low data intensity. Professional, scientific and technical activities, arts, entertainment and recreation as well as wholesale and retail are exhibiting data intensity shares between 2.5-5.0%. By contrast, manufacturing, which contributes to 10% of UK gross value added, and wholesale and retail, the sector with the highest number of jobs in the economy, are less data intensive than many of the service-oriented sectors. Accommodation and food service activities, as well as health and social work activities, are the least data-intensive sectors.

56. Data analytics activities are by far the most prevalent across all sectors, followed by database and data entry. This suggests that data-intensive jobs are essentially high-value added, high-skilled positions.

Figure 10. Data intensity of sectors in the United Kingdom

Per cent of labour demand, 2020



Note: Sectors are based on the ISIC vers. 4 classification. Activities of extra-territorial organisations and activities of households are excluded. Data intensity takes values between 0 and 100.
Source: Authors' calculation based on Lightcast data.

Differences in data intensity across the United Kingdom, Canada and the United States are concentrated in a handful of sectors

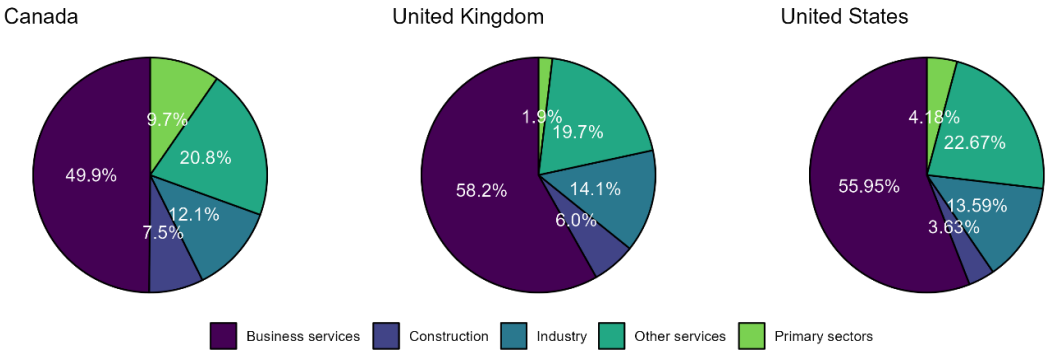
57. The sectoral structure of the United Kingdom, Canada and the United States is broadly similar. The three countries are all highly developed, industrialised economies with large services sectors (Figure 11.). The size of the services sector is larger in the United States and in the United Kingdom than in Canada, where the primary sector accounts for a comparatively larger share of the economy.

58. While the overall structure of the economy is similar, financial and insurance as well as information and communication activities are the two most data-intensive industries in all three countries, with shares of the same order of magnitude, close to or above 10% (Figure 12.). Shares are also very similar in most sectors with low data intensity, in particular accommodation and food service activities, construction, or transportation and storage. This is consistent with Calvino et al. (2018[16]), which use a different methodology.

59. Those numbers can however mask some structural differences across countries. For instance, in the finance and insurance sector, the United Kingdom's share is almost at par with the United States and Canada, with data mining analysts making the largest contribution to the data intensity of the sector in all three countries. However, the high demand for data mining analysts in the sector more than compensates the lower average data intensity of the profession in the United Kingdom (30% as compared to 70% in the United States and Canada). Overall, the contribution of the profession to the data intensity of the sector is about twice as high in the United Kingdom (0.8 percentage point), compared to Canada (0.3 percentage point) or of the United States (0.4 percentage point).

Figure 11. Sectoral decomposition of Canada, the United Kingdom and the United States

Per cent, 2020



Source: National accounts data (OECD, 2023[72]).

60. Differences in the level of data intensity of individual industries are noticeable in all the three countries in the industry professional, scientific and technical activities, where data intensity is much higher in the United States, and to a lesser extent in Canada than in the United Kingdom. Interestingly, the software developer is the profession that contributes most to the overall data intensity of this industry in all three countries. Yet, there are differences in the average data intensity of this profession across the countries. In the United Kingdom, a software developer is on average 10% data intensive, while in Canada and the United States scores amount to 25% and 24% respectively. One tentative explanation could be that in the United Kingdom, software developers are hired for a broader variety of tasks (e.g. front end development of applications), but they work very little directly with data. Weighted by the demand for employees in the profession, the contribution of the United Kingdom to the overall data intensity of the

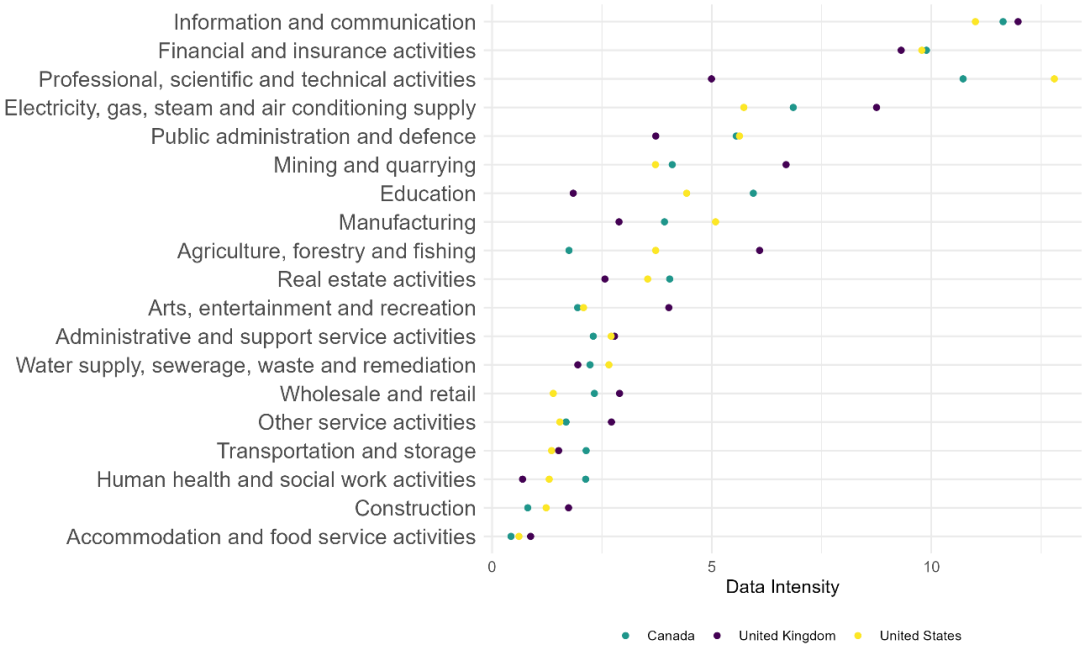
sector is only 0.2 percentage point, as compared to 1.2 percentage points in Canada and 1.5 percentage points in the United States.

61. Differences across countries are also marked regarding the extent of data intensity in agriculture and forestry and electricity, gas, steam and air conditioning supply, whose labour demand is more data intensive in the United Kingdom than in Canada or the United States. It is also important to note that these sectors are usually underrepresented in LC, potentially inflating the shares for these sectors.

62. In a few sectors, such as mining and quarrying and arts, entertainment and recreation, and public administration and defence, the United States and Canada exhibit very similar data intensity scores, which are much lower than in the United Kingdom (for an overview including breakdowns by data-related activities, see Annex B, Table B.1).

Figure 12. Data intensity in the United Kingdom, Canada, and the United States per industry

Per cent of labour demand, 2020



Note: Sectors are based on the ISIC version 4 classification. Activities of extraterrestrial organisations and activities of households are excluded. Data intensity takes values between 0 and 100.
Source: Authors' calculations based on Lightcast data.

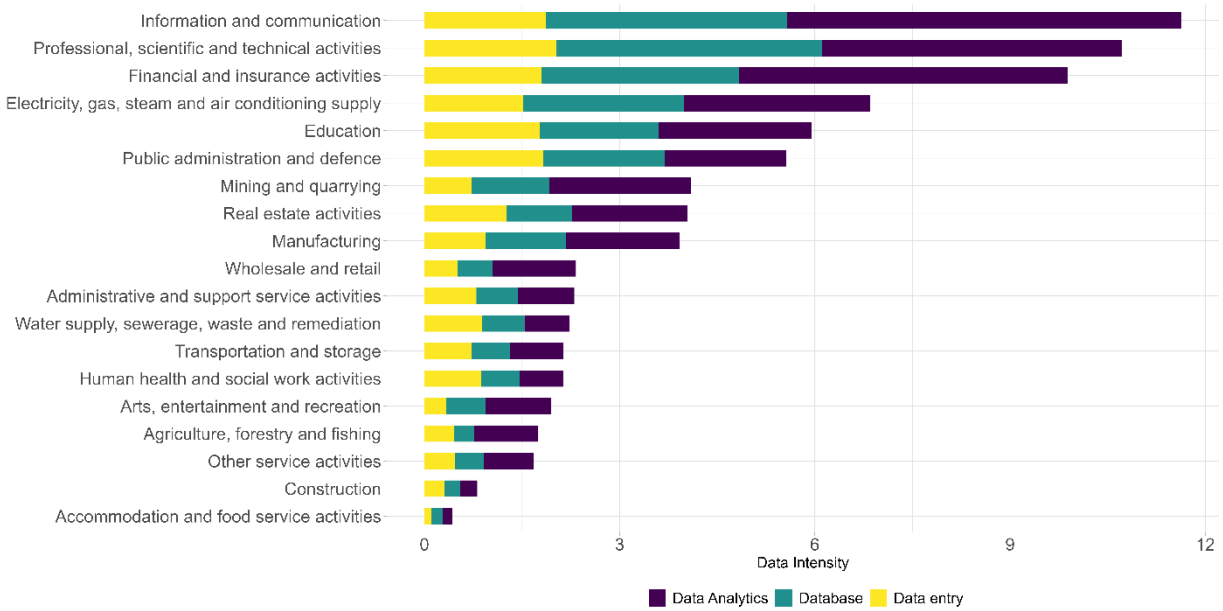
63. Data analytics appears to be the main contributor to sectoral data intensity scores in all three countries (Figure 13). This is in line with Corrado et al. (2022[33]). In financial services and insurance for instance, the data analytics component contributes to about half of the total data intensity of the sector in the three countries. In most sectors, however, the contribution of data analytics is much more important than those of database in the United Kingdom while the contribution of the two categories is closer in the United States, and sometimes even reverse with database contributing more to the score than data analytics (for example in professional scientific and technical activities).

64. In most sectors, data analytics contribute significantly more to the overall intensity score in the United Kingdom than in Canada or the United States. Notable exceptions are finance and insurance, manufacturing, real estate, professional scientific and technical activity, public administration, health and social works.

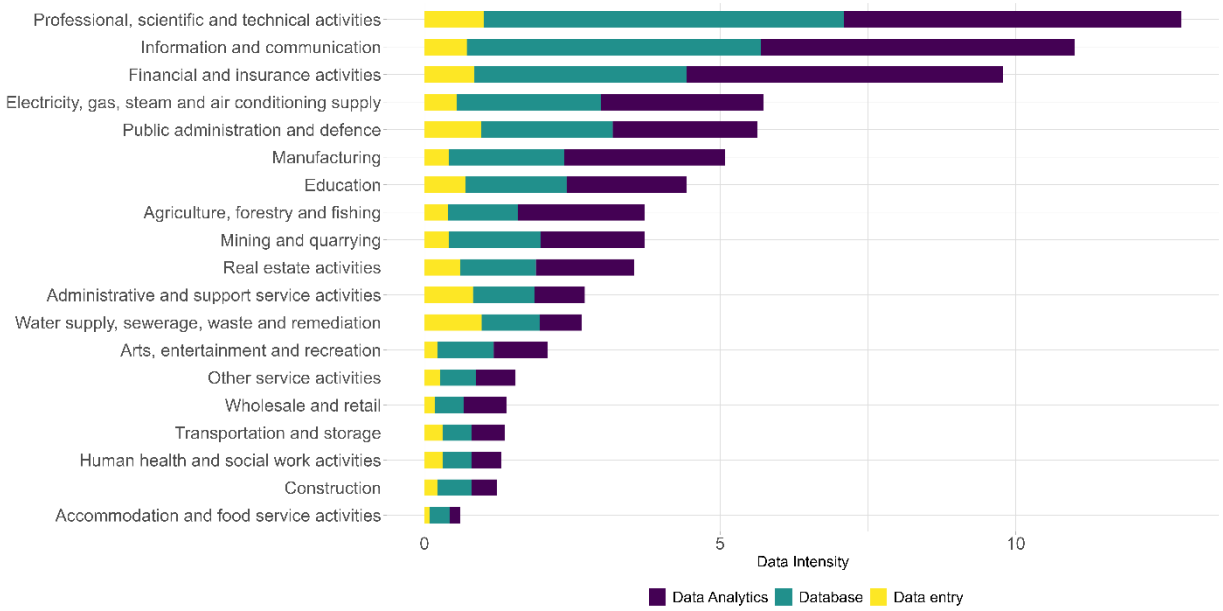
Figure 13. Data intensity by sector

Per cent of labour demand, 2020

A – Canada



B – United States



Note: Sectors are based on the ISIC version 4 classification. Activities of extraterrestrial organisations and activities of households are excluded. Data intensity ranges from 0 and 100.
Source: Authors' calculation based on Lightcast data.

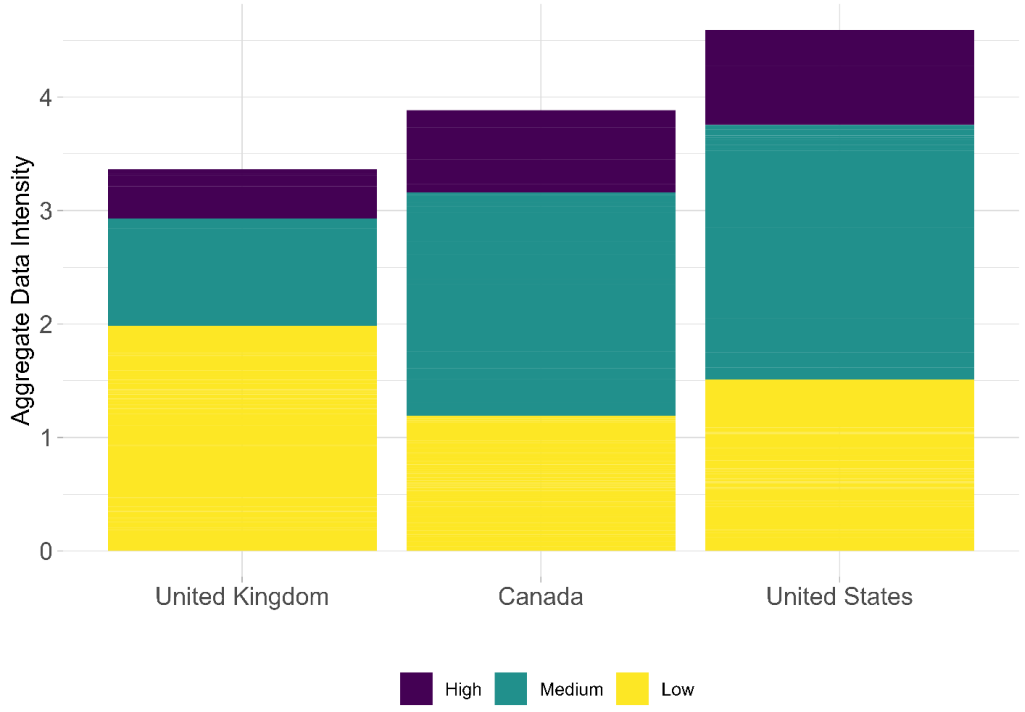
5.3. At the aggregate level, data intensity is relatively low in the United Kingdom

Professions with a low level of data intensity contribute most to the aggregate data intensity in the United Kingdom

65. At the economy-wide level, the flow of jobs in the United Kingdom and Canada appear to be less data-intensive than the United States (Figure 14). The overall share of data-intensive jobs in the United Kingdom lies at 3.4%, weighting the data intensity at occupation level by the number of job advertisements posted in 2020. This compares to 3.9% in Canada and 4.6% in the United States.

Figure 14. Low data-intensive occupations contribute most to data intensity in the United Kingdom

Per cent, 2020



Notes: Data intensity takes values between 0 and 100. Low data-intensive occupations: 0<10%, medium data-intensive occupations: 10-50% and high data-intensive occupations > 50%.
Source: Authors' calculations based on Lightcast data.

66. In the United Kingdom, the low data-intensive occupations are those that count most for the overall data economy – more than medium and highly data-intensive jobs (Figure 14). In Canada and the United States, medium data-intensive occupation classes contribute the largest proportion to the overall data intensity. Reweighting the United Kingdom's labour demand with the aggregate data intensity of Canada, would make the United Kingdom's labour demand much more data intensive, with a score of 6.0%.

Demand for low data-intensive jobs such as office administrative assistants and software developers are those which contribute the most to the UK overall score

67. The distribution of the demand data-intensive jobs is highly unequal for all three countries. A few highly data-intensive occupations exhibit higher data intensity than all other professions, while much of the labour force are potentially hired in professions that are medium or low data intensive.

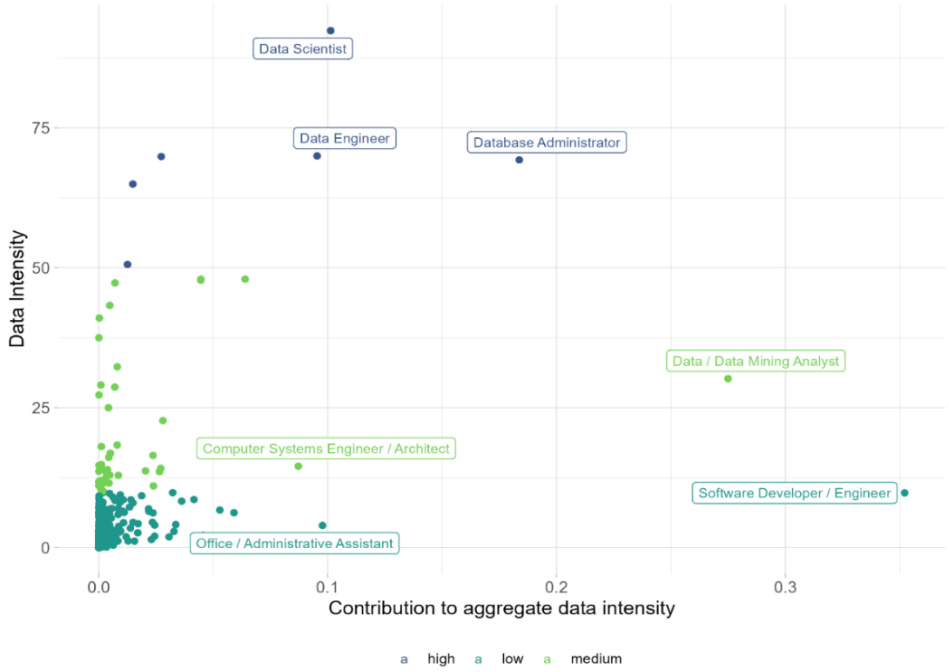
68. In the United Kingdom, most occupations are classified as low data intensive and thus employ a workforce that is only marginally involved in data production (Figure 15. Panel A). There exist only very few highly data-intensive occupations, such as the data scientist, data engineer, or data administrator. In some cases, professions with a low level of data intensity contribute more to the observed labour demand than many of the high data-intensive professions. For instance, office assistants, part of a low data-intensive occupation class, contribute as much to the overall data intensity of the economy as data scientists, as many more people may be potentially hired and ultimately work in this profession.

69. For Canada and the United States, the distribution of labour demand for data-intensive jobs is very similar, yet the level of data intensity across professions is generally higher. A data scientist, the occupation with the highest data intensity in all three countries, has a data intensity score of 94.5% in Canada and 95.1% in the United States, compared to 92.3% in the United Kingdom. In addition, more medium and high data- intensive occupations related to data production (Figure 15 Panel B and C). Among the high data-intensive professions in Canada, data entry clerks, database administrators, and data mining analysts contribute most to aggregate data intensity, next to professions such as the business management analyst and software developer at the medium level. The United States have the widest range of professions contributing to aggregate labour demand at the medium and high data intensity level, amongst them the network system analyst and the computer system engineer.

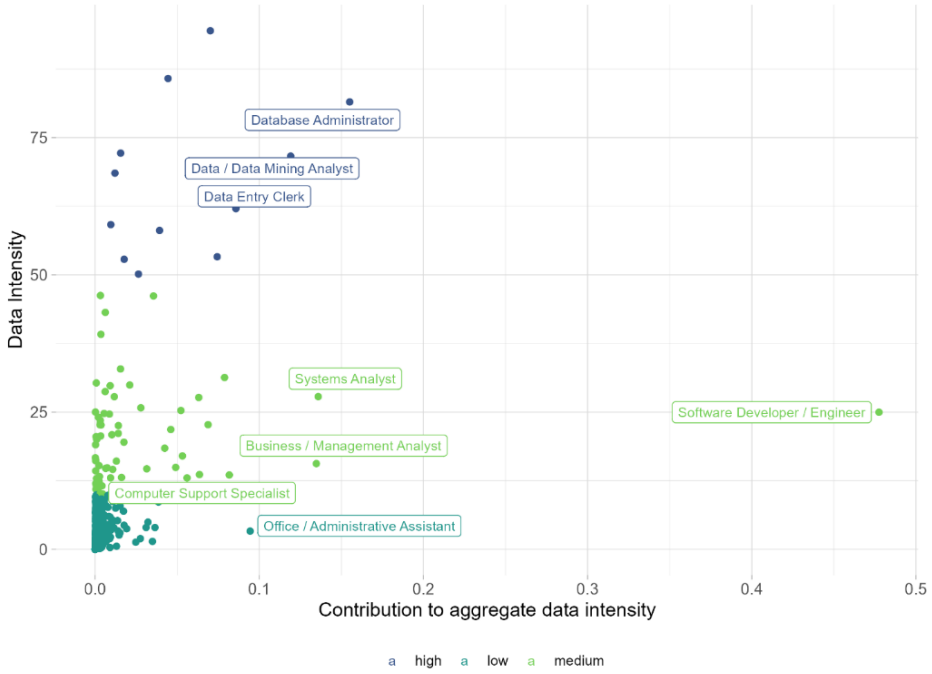
Figure 15. Data intensity across occupations

Data intensity of an occupation in per cent, contribution to aggregate data intensity in percentage points, 2020

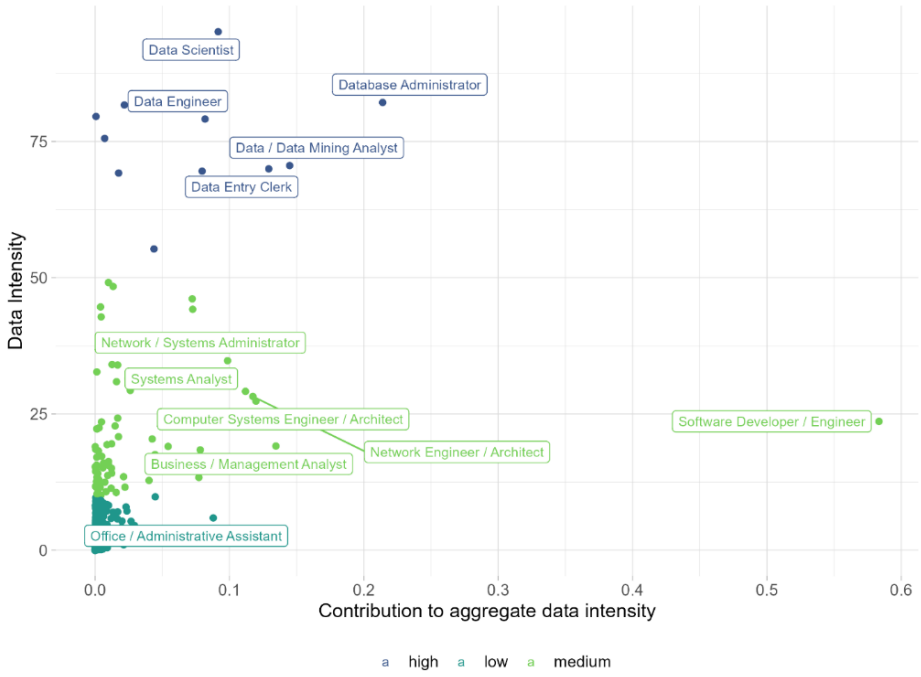
A - United Kingdom



B – Canada



C- United States



Notes: Contributions are computed as the data intensity of occupation classes weighted by their share of employees. Data intensity takes values between 0 and 100. Low data-intensive occupations: 0<10%, medium data-intensive occupations: 10-50% and high data-intensive occupations: > 50%. Contribution to aggregate data intensity is displayed in percentage points.
Source: Authors' calculation based on Lightcast data.

Estimates are sensitive to the calibration of the classification rules

70. Estimates of data intensity at the occupational level are sensitive to changes in the calibration of the classification rules. To assess this sensitivity, robustness checks were conducted on all classified job advertisements by varying the similarity and frequency measures by 10% around their baseline values. Overall, the order of magnitude for the aggregate data intensity at the economy level remains unchanged, although estimates vary with the threshold chosen for the similarity measure (Figure 16).

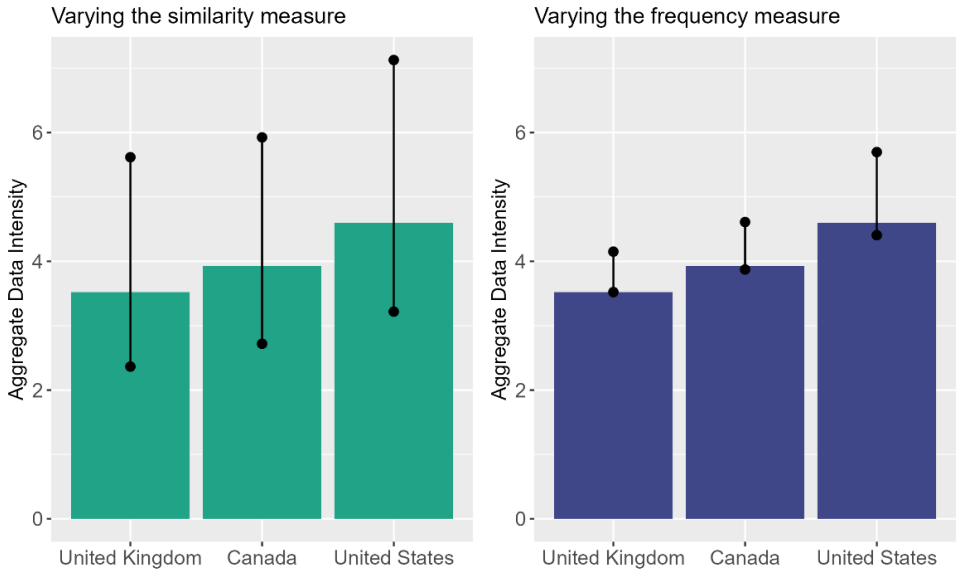
71. The results across countries tend to react strongly to a change in the similarity measure. Estimates for the United Kingdom range between 2.4-5.6% of aggregate data intensity when varying the similarity measure from 0.45 to 0.55. This compares to 2.7-5.9% for Canada and 3.2-7.1% for the United States. In other words, the algorithm becomes more restrictive in identifying noun chunks as data related when increasing the similarity measure. Above similarity measures of 0.5, the algorithm excludes words such as "strong quantitative analytical skills" or "predictive modelling". Below 0.5, the algorithm includes a large share of words that are generally not similar to the term "data", such as "strategies", "consumers" or "industry activity". This, together with the objective to design an inclusive but clear classification rule in line with NLP literature, justify a baseline value of 0.5.

72. In contrast, a proportional change of the frequency measure leads to comparatively smaller variations of aggregate data intensity, albeit with slight differences across countries. The United Kingdom estimates range from 3.5-4.2%, while the estimate for the United States ranges from 4.4-5.7% and 3.9-4.7% in Canada. This measure is critical to ensure that the number of noun chunks per job advertisement itself does not drive the selection of jobs labelled as data intensive and had to be adjusted according to

the sample of text. The measure was calibrated on the size of the text corpus of the United Kingdom, which explains why the interval is asymmetric around the baseline value. As the United States has the largest text corpus, estimates for this country react more to an adjustment in its frequency measure than in Canada or the United Kingdom.

Figure 16. Sensitivity analysis over the aggregate data intensity

Per cent, 2020



Note: The similarity measure refers to the cosine similarity described in the classification rule in Section 3.3. The baseline cosine similarity is 0.5. The frequency measure is a relative frequency share and is country specific.
Source: Authors' calculations based on Lightcast data.

5.4. Data intensity shares can be used to derive estimates of investment in data

73. Sectoral data intensity shares of jobs, computed in the previous sections, can be used to derive estimates of the labour cost share of data production activities, which is a proxy of economy-wide investment in data. The calculation is performed for the United Kingdom, Canada and the United States. It follows the sum of costs approach put forward by Statistics Canada (2019[32]).

74. Estimates of data investment are highly sensitive to the value of a markup (α in section 3.4), which captures non-wage cost and capital services margins. Those markups are hard to estimate, especially at a very granular level and for some specific sectors, as it is difficult to determine which intermediate inputs are needed to generate data assets. In most studies, the same markup as for software assets or an approximation thereof is chosen. Given the uncertainties around this parameter, it was judged preferable to present a range rather than point estimates. The lower bound estimates are derived using a value of 1.5 for α, in line with Statistics Canada (2019[32]). The upper bound estimates are computed by recalculating α using information on gross value added, compensation of employee, intermediate consumption, consumption of fixed capital and net operating surplus from national accounts for each of the respective countries in 2020. Resulting α's are around 3.0-3.3.

75. Unlike the traditional sum of costs method, which uses a time-use factor, the calculation is derived from the skill requirements contained in job advertisements. In addition, national accounts data on employment compensation at sectoral level are chosen rather than labour market data that account for

differences in salaries between data and non-data-intensive occupations. This could underestimate the actual value of data, as data-intensive workers usually earn a wage premium.

Table 3. Investment in data at economy level, 2020

	Canada	United Kingdom	United States
Billion, national currency	69.3 – 147.9	63.4 – 141.7	901.1 – 1902.2
As a share of GVA, per cent	3.1 – 6.7	3.0 - 6.7	4.4 - 9.4
Of which			
Data entry, p.p	0.7 - 1.5	0.3 - 0.7	0.5 - 1.1
Database, p.p	1.0 - 2.2	0.9 - 2.0	2.0 - 4.1
Data analytics, p.p	1.4 – 2.9	1.7 - 3.9	2.2 - 4.7

Note: The estimates are derived using the equation in section 3.4. The lower bound estimates apply a markup of 1.5 following Statistics Canada (2019). The upper bound estimates use a country-specific markup = (compensation of employees + intermediate consumption (excluding materials) + consumption of fixed capital + net operating surplus)/ compensation of employees. For Canada, data on intermediate consumption was not available in 2020 and was approximated by applying the growth rate of the GVA to the 2019 estimate.
Source: Authors' calculations based on Lightcast data and national accounts data (OECD, 2023[72]).

76. Estimates for 2020 point to an investment in data of the same magnitude in the three economies, between 3 to 9% of GVA (Table 3). Estimates for the United Kingdom are relatively lower than those of the United States, with all three categories related to data production contributing to this difference.

Table 4. Estimates of data investment from the literature

	Rassier, Kornfeld, Strassner (2019)	Statistics Canada (2019)	De Bondt and Mushkudiani (2021)	Smedes, Nguyen and Tenburren (2022)	Santiago Calderón, Rassier (2022)	Corrado et al. (2022)
Countries	United States	Canada	The Netherlands	Australia	United States	Group of 9 European economies*
Period	2012-2017	2005 - 2018	2001-2017	2011 -2016	2003-2020	2010-2018
Sectors	Private sector	Private + public sector	Private sector	Private + public sector	Private sector	Private sector
Investment in data (flow)	USD 52.8 billion (2012) USD 74.9 billion (2017)	CAD 29 – 40 billion CAD (2018)	EUR 15.6 – 20 billion (2017)	AUD 35 – 47 billion (2016)	USD 82.6 billion (2003) USD 159.5 billion (2020)	-
Investment in data (flow)/GDP	0.3% (2012) 0.4% (2017)	1.3% - 1.8% (2018)	2.1% - 2.7% (2017)	2.2-2.7% (2016)	0.7% (2003) 0.8% (2020)	5.3% (2010-2018)

Note: * The group of nine European economies includes Germany, Denmark, Spain, Finland, France, United Kingdom, Italy, the Netherlands, Sweden.
Source: Authors' compilation.

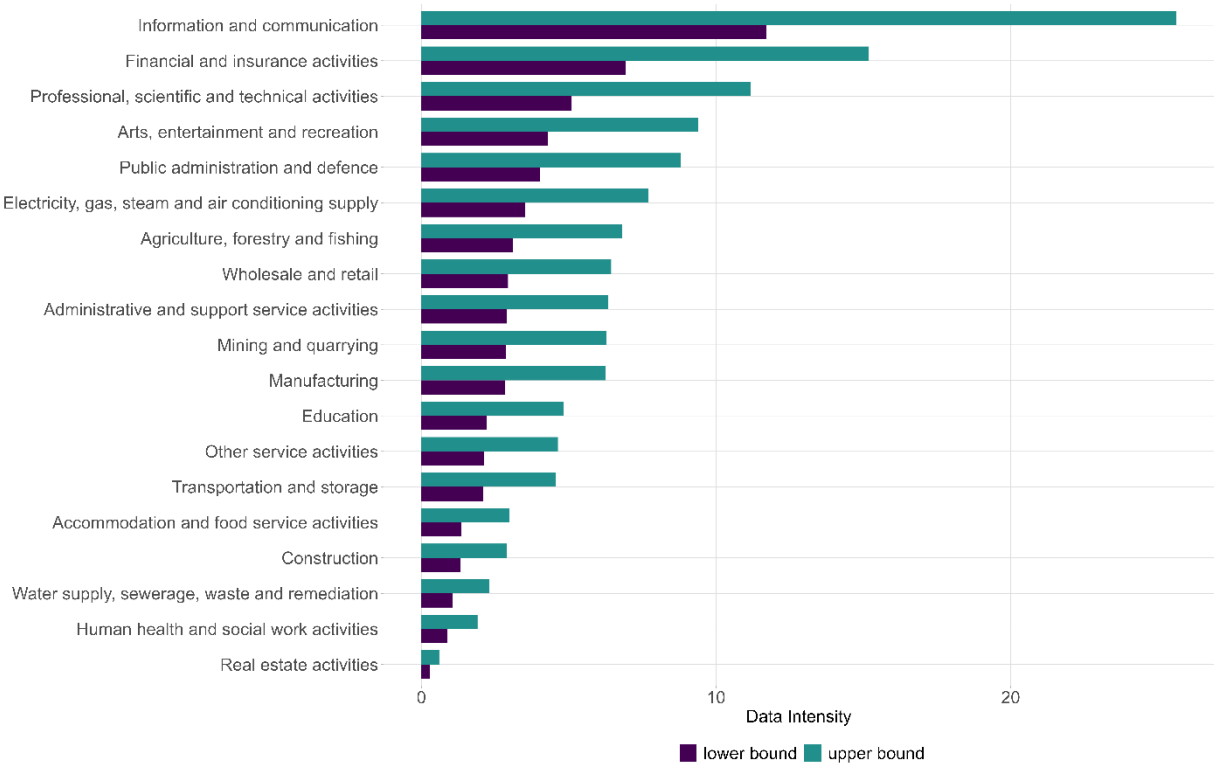
77. Estimates are within the range of those found in most recent analyses from the literature, though higher than what a number of past studies have found (Table 3). Estimates for the United Kingdom are in line with the recent results presented by Corrado et al. (2022[33]), whose estimates amount to 6.5%. Differences across studies seem to reflect to a large extent differences in the period under consideration, but also the sectoral coverage (whether it is limited to the private sector or not) and the regional coverage. Data sourcing and the methods employed also matter. Estimates reported in this study are higher than

past results, as the NLP approach allows for better capturing the data-intensive tasks/responsibilities/skillsets across occupations and the sum of costs approach is highly sensitive to the selection of occupations chosen that contribute to data production.

78. At the industry level, the information, finance and scientific industries are found to be the most data intensive sectors in the United Kingdom (Figure 17). The lowest data investment tends to come from real estate activities, human health and water supply sewerage waste management activities. Most interestingly, administrative and support services mining and quarrying as well as manufacturing appear to invest to similar extent in data assets. While the upper and lower bound estimates vary due to the uncertainty around non-wage and capital services costs, the order of magnitude across sectors remains stable.

Figure 17. Investment in data assets at sectoral level in the United Kingdom

Investment in data as share of GVA, industry level (ISIC version 4), per cent, 2020



Note: The estimates are derived using the equation in section 3.4. The lower bound estimates, apply a markup of 1.5 following Statistics Canada (2019). The upper bound estimates use a country-specific markup = (compensation of employees + intermediate consumption (excluding materials) + consumption of fixed capital + net operating surplus)/ compensation of employees.
Source: Authors' calculations based on Lightcast data and national accounts data (OECD, 2023[72]).

6. Conclusion

79. This paper has developed an NLP methodology to estimate the data intensity of occupations and sectors in the United Kingdom, Canada and the United States. It uses online job advertisements, a timely and granular source of data that enables detailed insights into the skills requirements of professions. The approach allows for a breakdown of data intensity into data entry, database and data analytics activities and a comparison of the estimates across countries while capturing country-specific variations. In addition, by providing a key input to the sum of costs approach, it contributes to advancing the measurement of investment in data consistent with the System of National Accounts (SNA).

80. Despite those important benefits, the approach is subject to several caveats and limitations. First, it remains quite IT- and time-consuming, given the massive amount of information that needs to be treated. Second, estimates remain sensitive to the calibration of thresholds chosen for the similarity measure, highlighting the need to calibrate this parameter carefully. Third, the skillsets for digital skills are often less consolidated than for more traditional datasets, making it more difficult for the NLP algorithm to extract all skills potentially contributing to data production. Related to that, cultural differences in how the job advertisements are framed, as well as implicit versus more explicit language, are a known hurdle to NLP methods (Sostero and Tolan, 2022[1]).

81. The approach could be extended in several directions.

82. First, it could be implemented from 2012 to real time to assess whether job data intensity has significantly evolved over time. Preliminary analysis using data for 2019 point to similar results. Furthermore, the analysis could be expanded to other countries where LC data coverage is good, such as Australia, New Zealand and Singapore and/or run at the regional or subnational level, data permitting.

83. Second, the approach is flexible and can be applied to capture concepts outside of the realm of traditional occupation classifications, such as green or AI-related jobs. The deployed NLP algorithm can handle over 66 languages, thereby enabling analysis of online job advertisements in non-English-speaking countries. However, in such cases, the job advertisement text would either need to be manually web-scrapped or provided by alternative sources.

84. Finally, the findings in this paper hold the potential to provide valuable insights for policy research on the contribution of data assets to the digital economy and for studying labour market developments related to digital skills. For instance, estimates on data intensity of industries could be incorporated into theoretical models examining how data in production processes influence economic growth (Farboodi and Veldkamp, 2021[73]) or on empirical studies investigating the productivity of data- intensive firms (Brynjolfsson and McElheran, 2016[74]).

References

Acemoglu, D. and P. Restrepo (2017), "Robots and Jobs: Evidence from US Labor Markets", *NBER Working Paper*, https://doi.org/10.3386/w23285. [2]

ADB (2022), *Digital jobs and digital skills. A shifting landscape in Asia and the Pacific*, http://dx.doi.org/10.22617/SPR220348. [15]

Baruch, L. and G. Feng (2016), *The End of Accounting and the Path Forward for Investors and Managers*, https://doi.org/10.1002/9781119270041. [56]

Beblavy, M., B. Fabo and K. Lenaerts (2016), *Demand for Digital Skills in the US Labour Market*, https://www.ceps.eu/ceps-publications/demand-for-digital-skills-in-the-us-labour-market-the-it-skills-pyramid/. [19]

Bellatin, A. and G. Galassi (2022), "What COVID-19 May Leave Behind: Technology-Related Job Postings in Canada", *Bank of Canada Staff Working Paper*, https://www.bankofcanada.ca/2022/04/staff-working-paper-2022-17/. [22]

Berlingieri, G. et al. (2020), "Laggard firms, technology diffusion and its structural and policy determinants", *OECD Science, Technology and Industry Policy Papers*, Vol. 86, https://doi.org/10.1787/281bd7a9-en. [3]

Borgonovi, F. et al. (2023), "The effects of the EU Fit for 55 package on labour markets and the demand for skills", *OECD Social, Employment and Migration Working Papers No 297*, OECD Publishing, Paris, https://doi.org/10.1787/6c16baac-en. [68]

Boselli, R. et al. (2018), "Classifying online Job Advertisements through Machine Learning", *Future Generation Computer Systems*, Vol. 86, pp. 319-328, https://doi.org/10.1016/j.future.2018.03.035. [23]

Brynjolfsson, E. and K. McElheran (2016), "Data in Action: Data-Driven Decision Making in U.S. Manufacturing", *US Census Bureau Center for Economic Studies Paper No. CES-WP-16-06; Rotman School of Management Working Paper No. 2722502*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2722502. [74]

Brynjolfsson, E., D. Rock and C. Syverson (2021), "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies", *American Economic Journal: Macroeconomics*, Vol. 13/1, pp. 333-72, https://doi.org/10.1257/mac.20180386. [11]

Calderón, J. and D. Rassier (2022), "Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings", *U.S. Bureau of Economic Analysis | NBER Working Paper*, https://conference.nber.org/conf_papers/f159271.pdf. [44]

Calligaris, S., C. Criscuolo and L. Marcolin (2018), "Mark-ups in the digital era", *OECD Science, Technology and Industry Working Papers*, OECD Publishing, Paris, https://doi.org/10.1787/4efe2d25-en. [7]

Calvino, F. et al. (2018), "A taxonomy of digital intensive sectors", *OECD Science, Technology and Industry Working Papers*, OECD Publishing, Paris, https://doi.org/10.1787/f404736a-en. [16]

Cameraat, E. and M. Squicciarini (2021), "BurningGlass Technologies data use in policy-relevant analysis: An occupation-level assessment", *OECD Science, Technology and Industry Working Papers*, OECD Publishing, Paris, https://doi.org/10.1787/cd75c3e7-en. [60]

Carnevale, A., T. Jayasundera and D. Repnikov (2014), "Understanding Online Job Ads Data: A Technical Report", *Washington, DC: Georgetown University Center on Education and the Workforce.*, https://repository.library.georgetown.edu/handle/10822/1050294. [63]

Carrasco, S. and R. Rosillo (2021), *Word Embeddings, Cosine Similarity and Deep Learning for Identification of Professions & Occupations in Health-related Social Media*, https://aclanthology.org/2021.smm4h-1.12.pdf. [36]

Corrado, C. et al. (2022), "Data, digitization and productivity", *NBER Working Papers*, https://www.nber.org/books-and-chapters/technology-productivity-and-economic-growth/data-digitization-and-productivity. [33]

Corrado, C. et al. (2022), "The value of data in digital-based business models: Measurement and economic policy implications", *OECD Economics Department Working Papers, No. 1723*, OECD Publishing, Paris, https://doi.org/10.1787/d960a10c-en. [10]

Coyle, D. et al. (2020), "The value of data. Policy implications", https://www.bennettinstitute.cam.ac.uk/media/uploads/files/Value_of_data_Policy_Implications_Report_26_Feb_ok4noWn.pdf. [47]

Coyle, D. and W. Li (2021), "The Data Economy: Market size and global trade", Paper prepared for the 36th IARIW Virtual General Conference. Session 23: National Accounts, https://iariw.org/wp-content/uploads/2021/08/coyle_li_paper.pdf. [54]

Coyle, D. and A. Manley (2022), "What is the value of data? A review of empirical methods", https://www.bennettinstitute.cam.ac.uk/publications/value-of-data/. [53]

Crocetti, G. (2015), "Textual Spatial Cosine Similarity", https://arxiv.org/ftp/arxiv/papers/1505/1505.03934.pdf. [41]

de Bondt, H. and N. Mushkudiani (2021), "Estimating the Value of Data in the Netherlands", *Paper prepared for the IARIW-ESCoE Conference*, https://iariw.org/wp-content/uploads/2021/10/bondt_paper.pdf. [75]

Devlin, J. et al. (2018), "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint* arXiv:1810.04805. [62]

Farboodi, M. and L. Veldkamp (2021), "A growth model of the data economy", *NBER Working Paper 28427*, https://www.nber.org/system/files/working_papers/w28427/w28427.pdf. [73]

Garasto, S. et al. (2021), "Developing experimental estimates of regional skill demand", https://www.escoe.ac.uk/publications/developing-experimental-estimates-of-regional-skill-demand/. [18]

Grabner, S. and A. Tsvetkova (2022), "Urban labour market resilience during the COVID-19 pandemic: what is the promise of teleworking?", *Regional studies*, https://doi.org/10.1080/00343404.2022.2042470. [66]

Hansen, S. et al. (2021), "The Demand for Executive Skills", *Havard Business School*, https://www.hbs.edu/ris/Publication%20Files/Working%20Paper%2021-133_27738927-867c-4b8c-9358-05a61ec6f40b.pdf. [28]

Harrigan, J., A. Reshef and F. Toubal (2021), "The March of the Techies: Job Polarization Within and Between Firms", *Research Policy*, Vol. 50/7, https://doi.org/10.1016/j.respol.2020.104008. [4]

Hershbein, B. and L. Kahn (2018), "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings", *American Economic Review*, Vol. 108/7, pp. 1737-1772, https://doi.org/10.1257/aer.20161570. [64]

ISWGNA (2022), *Recording of Data in the National Accounts*, https://unstats.un.org/unsd/nationalaccount/RAdocs/DZ6_GN_Recording_of_Data_in_NA.pdf. [34]

Jones, C. and C. Tonetti (2020), "Nonrivalry and the Economics of Data", *American Economic Review*, Vol. 110/9, pp. 2819-2858, https://doi.org/10.1257/aer.20191330. [48]

Jurafsky, D. and J. Martin (2023), *Speech and Language Processing*, https://web.stanford.edu/~jurafsky/slp3/. [38]

Kanders, K. and K. Sleeman (2021), "Open Jobs Observatory: Extracting skills from online job adverts", https://www.nesta.org.uk/project-updates/skills-extraction-ojo/. [30]

Ker, D. and E. Mazzini (2020), "Perspectives on the value of data and data flows", *OECD Digital Economy Papers, No. 299*, OECD Publishing, Paris, https://doi.org/10.1787/a2216bc1-en. [55]

Kortum, H., D. Rebstadt and O. Thomas (2022), *Dissection of AI Job Advertisements: A text mining-based analysis of employee skills in the disciplines computer vision and natural language processing*, https://www.researchgate.net/profile/Henrik-Kortum/publication/357740738_Dissection_of_AI_Job_Advertisements_A_Text_Mining-based_Analysis_of_Employee_Skills_in_the_Disciplines_Computer_Vision_and_Natural_Language_Processing/links/61fb9c33007fb50447311f7c/. [29]

Kotu, V. and B. Deshpande (2019), *Data Science. Concepts and Practice (Concepts and Practice)*, https://doi.org/10.1016/C2017-0-02113-4. [40]

Koutroumpis, P., A. Leiponen and T. Lleweellyn (2020), "Markets for data. The market for lemons: quality uncertainty and the market mechanism", *Industrial and Corporate Change*, Vol. 29/3, pp. 645-660, https://EconPapers.repec.org/RePEc:oup:indcch:v:29:y:2020:i:3:p:645-660. [46]

Lancaster, V., D. Mahoney-Nair and N. Ratcliff (2019), *Review of Burning Glass Job-ad Data*, https://biocomplexity.virginia.edu/sites/default/files/projects/Technical%20Report%20Review%20of%20BGT%20Job-ad%20Data.pdf. [61]

Lassébie, J. et al. (2021), "Speaking the same language: A machine learning approach to classify skills in Burning Glass Technologies data", *OECD Social, Employment and Migration Papers No. 263*, OECD Publishing, Paris, https://dx.doi.org/10.1787/adb03746-en. [17]

López González, J., S. Sorescu and P. Kaynak (2023), "Of bytes and trade: Quantifying the impact of digitalisation on trade", *OECD Trade Policy Papers 273*, OECD Publishing, Paris, https://doi.org/10.1787/11889f2a-en. [8]

Manning, C. and H. Schutze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, https://nlp.stanford.edu/fsnlp/. [42]

Mitchell, J., D. Ker and M. Lesher (2021), "Measuring the economic value of data", *OECD Going Digital Toolkit Note*, No 20, https://goingdigital.oecd.org/data/notes/No20_ToolkitNote_MeasuringtheValueofData.pdf. [57]

Muro, M. et al. (2017), *Digitalization and the American Workforce*, https://www.brookings.edu/research/digitalization-and-the-american-workforce/. [14]

OECD (2023), *Small and Medium Enterprise and Entrepreneurship Outlook*, OECD. [5]

OECD (2023), *Value added by activity (indicator)*, https://doi.org/10.1787/a8b2bd2b-en (Accessed on 13 February 2023). [72]

OECD (2022), "Measuring the value of data and data flows", *OECD Digital Economy Papers No. 345*, OECD Publishing, Paris, https://doi.org/10.1787/923230a6-en. [45]

OECD (2022), "Shedding new light on the evolving regulatory framework for digital services trade", OECD Publishing, Paris, https://issuu.com/oecd.publishing/docs/shedding-new-light-on-the-evolving-regulatory-fram. [9]

OECD (2020), *4th meeting of the Informal Advisory Group on measuring GDP in the Digital. Item 4b: Valuing the data economy: A labor costs approach using unsupervised machine*, https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=SDD/CSSP/WPNA/A(2020)1&docLanguage=En. [52]

OECD-UNSD (2023), *OECD-UNSD Multinational Enterprise Information Platform*, https://www.oecd.org/sdd/its/mne-platform.htm. [70]

ONS (2023), *SOC 2020 Volume 1: structure and descriptions of unit groups*, https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020/soc2020volume1structureanddescriptionsofunitgroups. [69]

Rassier, D., R. Kornfeld and E. Strassner (2019), "Treatment of Data in National Accounts", *Paper prepared for the BEA Advisory Committee*, https://www.bea.gov/system/files/2019-05/Paper-on-Treatment-of-Data-BEA-ACM.pdf. [50]

Reddivari, S. and J. Wolbert (2022), "Calculating Requirements Similarity Using Word Embeddings", *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 438-439, https://doi.org/10.1109/COMPSAC54236.2022.00079. [37]

Rock, D., S. Bana and E. Brynjolfsson (forthcoming), *Word2Vec: Learning the latent structure of the labour market*, https://escoe-website.s3.amazonaws.com/wp-content/uploads/2022/01/25103832/Daniel-Rock-Slides.pdf. [27]

Samek, L., M. Squicciarini and E. Cameraat (2021), "The human capital behind AI. Jobs and skills demand from online job postings", *OECD Science, Technology and Industry Policy Papers, No. 120*, OECD Publishing, Paris, https://doi.org/10.1787/2e278150-en. [20]

Santiago Calderón, J. and D. Rassier (2022), "Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings", *U.S. Bureau of Economic Analysis | NBER Working Paper*, https://conference.nber.org/conf_papers/f159271.pdf. [12]

Sayfullina, L., E. Malmi and J. Kannala (2018), "Learning Representations for Soft Skill Matching", *Analysis of Images, Social Networks and Texts. AIST 2018*, Lecture Notes in Computer Science, pp. 141-152, https://doi.org/10.1007/978-3-030-11027-7_15. [25]

Schoch, D. (2020), “Mergers and Acquisitions in the Data Economy”, http://dx.doi.org/10.2139/ssrn.3686247. [6]

Smedes, M., T. Nguyen and B. Tenburren (2022), *Valuing data as an asset, implications of economic measurement*, https://www.abs.gov.au/about/economic-implications-digital-economy#papers-session-1. [76]

Soh, J. et al. (2022), “Did the COVID-19 Recession Increase the Demand for Digital Occupations in the United States? Evidence from Employment and Vacancies Data”, *IMF Working Paper 2022/195*, https://www.imf.org/en/Publications/WP/Issues/2022/09/23/Did-the-COVID-19-Recession-Increase-the-Demand-for-Digital-Occupations-in-the-United-States-523606. [21]

Sostero, M. and S. Tolan (2022), “Digital skills for all? From computer literacy to AI skills in online job advertisements”, *JRC Working Papers Series on Labour Education and Technology*, https://joint-research-centre.ec.europa.eu/publications/digital-skills-all-computer-literacy-ai-skills-online-job-advertisements_en. [1]

spaCy (2023), *spaCy Model Architectures*, https://spacy.io/api/architectures. [39]

spaCy (2022), *Language Processing Pipelines*, https://spacy.io/usage/processing-pipelines. [35]

Squicciarini, M. and H. Nachtigall (2021), “Demand for AI skills in jobs. Evidence from job postings”, *OECD Science, Technology and Industry Working Papers*, OECD Publishing, Paris, https://doi.org/10.1787/3ed32d94-en. [67]

Statistics Canada (2019), *Measuring investment in data, databases and data science: Conceptual framework*, https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00008-eng.htm. [32]

Statistics Canada (2019), *The value of data in Canada: Experimental estimates*, https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm. [43]

Strohmeier, S. (2020), “digital human resource management: A conceptual clarification”, *German Journal of Human Resource Management*, Vol. 34/3, pp. 345-365, https://journals.sagepub.com/doi/pdf/10.1177/2397002220921131. [58]

Tamburri, D., W. Van Den Heuvel and M. Garriga (2022), “DataOps for Societal Intelligence: a Data Pipeline for Labor Market Skills Extraction”, *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, https://ieeexplore.ieee.org/document/9191408. [26]

Tsvetkova, A. et al. (forthcoming), *Representativeness of Lightcast web-scrapped vacancy data. An assessment for largest English-speaking countries*. [59]

U.S. Department of Commerce (2015), *The Importance of Data Occupations in the U.S*, https://www.commerce.gov/sites/default/files/migrated/reports/the-importance-of-data-occupations-in-the-us-economy_0.pdf. [13]

UK Department for Environment Food and Rural Affairs (2022), “Agriculture in the UK Evidence Pack”, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1106562/AUK_Evidence_Pack_2021_Sept22.pdf. [71]

UNSTATS (2023), "22nd Meeting of the Advisory Expert Group on National Accounts. DZ.6 Recording of data in the National Accounts", https://unstats.un.org/unsd/nationalaccount/RADOCS/ENDORSED_DZ6_Recording_of_Data_in_NA.pdf. [51]

van de Ven, P., N. Ahmad and P. Schreyer (2018), *How to deal with globalisation in the framework of national accounts*, https://www.oecd.org/iaos2018/programme/IAOS-OECD2018_Schreyer-vandeVen-Ahmad.pdf. [49]

Vassilev, G., O. Romanko and K. Evans (2021), "What's in a job? Measuring skills from online job adverts", *Paper prepared for the 36th IARIW Virtual General Conference*, https://iariw.org/wp-content/uploads/2021/08/evans_et_al_paper.pdf. [31]

Vermeulen, W. and F. Amaros (forthcoming), "How well do online job postings cover European regions, sectors and occupations? Benchmarking Lightcast data against national and European statistical and labour agency sources", *OECD Local Economic and Employment Development (LEED) Papers*. [65]

Zhang, M. et al. (2022), "Skillspan: Hard and Soft Skills Extraction from English Job Postings", *arXIV*, Accepted to NAACL 2022 Main conference, https://doi.org/arXiv:2204.12811. [24]

Annex A. Aggregating noun chunks to jobs, occupations, and sectors

This annex provides two examples on how the measures at noun chunk level are translated into an indicator of data intensity at occupation or sector level. In a first step the noun chunks are classified to identify whether a job is data intensive or not, and to what extent the data intensity is driven by data entry, database, and data analytics skillsets. In a second step, the jobs are aggregated by occupations (SIC, NOC and NAICS respectively), and sectors (ISIC version 4) using their weighted mean to calculate the share of data intensive jobs in each grouping.

From noun chunks to data-intensive jobs

A job is labelled as data intensive based on the noun chunks contained in the text of the job advertisement. The classification of noun chunks is described in Section 3.3. Table A.1 shows a hypothetical example of one job with five noun chunks in its job advertisement. A job (j) gets classified as data intensive (1) if more than three noun chunks in this specific advertisement were classified, and non-data-intensive (0) otherwise. The noun chunks (n) are classified as either data entry, database, or data analytics. The components are independent from each other, meaning a noun chunk identified as a data entry related noun chunk cannot be labelled as database component nor as data analytics component. For each job, the breakdown into data entry, database and data analytics is normalised by the total number of noun chunks classified so that all three components add up to one. In this example, four out of five noun chunks were classified as data-related, the job as thus identified as data-intensive. The job is also 25% data entry related, 25% database related, and 50% data analytics related.

Table A.1. Example of aggregation from noun chunk to job level

Job ID	Noun chunk	Classified as data entry	Classified as database	Classified as data analytics	Not classified	Count of classified chunks
j_1	n_1	1	0	0	0	4
j_1	n_2	0	0	0	1	4
j_1	n_3	0	1	0	0	4
j_1	n_4	0	0	1	0	4
j_1	n_5	0	0	1	0	4

Job ID	Data entry share	Database share	Data analytics share
j_1	1/4	1/4	2/4

Source: Authors' illustration.

From jobs to occupations and sectors

In a second aggregation step, the data intensity per occupation class is calculated by taking the arithmetic mean across all jobs in the respective class (see Table A.2) shows how the data intensity at occupation level is calculated for a sample of five jobs (j). All jobs belong to the same occupation class (O_1). Each job is either data-intensive (1) or non-data intensive (0), with the respective breakdowns by data entry, database and data analytics. The data intensity of the occupation (O_1) is calculated by taking the arithmetic

mean across all jobs in the specific occupation. The aggregation by sectors follows the same sequence as the aggregation by occupation illustrated in Table A.1.

Table A.2. Exemplary aggregation of data intensive jobs to occupations

Occ ID	Job ID	Data entry share	Database share	Data analytics share	Data intensity j	Job count
occ_1	j_1	¼	¼	2/4	1	5
occ_1	j_2	5/10	5/10	0	1	5
occ_1	j_3	0	5/5	0	1	5
occ_1	j_4	0	0	0	0	5
occ_1	j_5	0	0	0	0	5

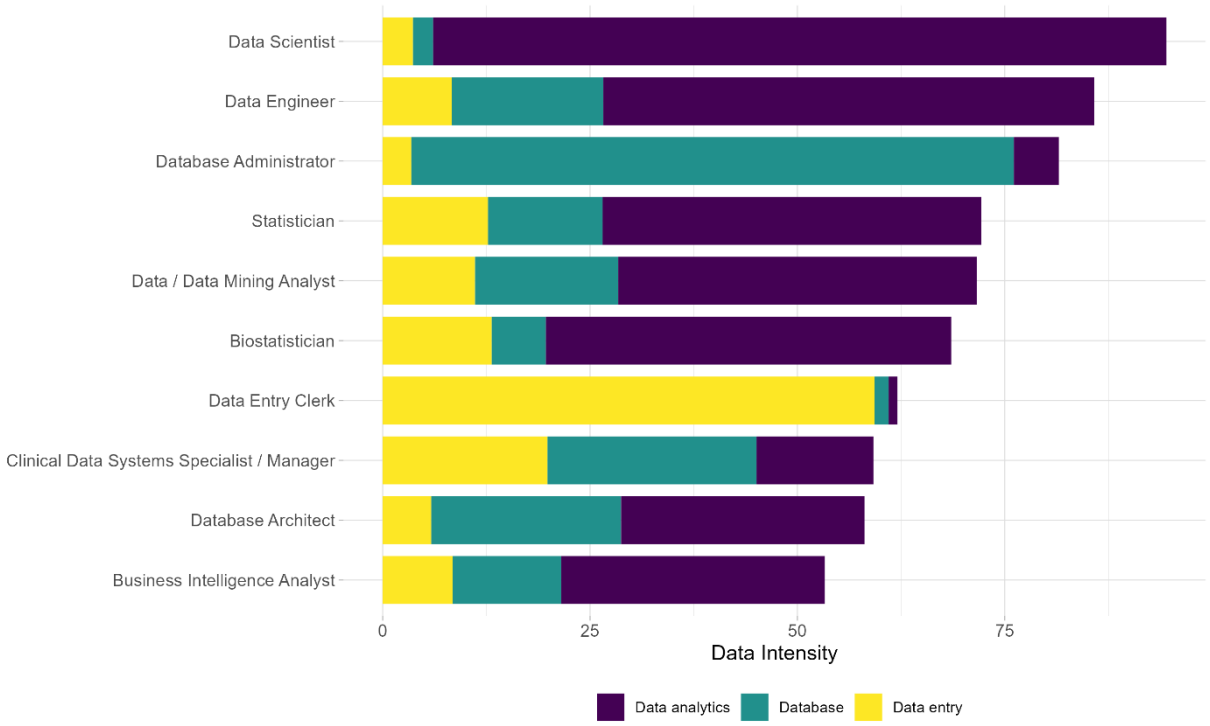
Occ ID	Data entry share	Database share	Data analytics share	Data intensity o
occ_1	0.15	0.35	0.1	0.6

Source: Authors' illustration.

Annex B. Additional results

Figure B.1. Top 10 data-intensive occupations in Canada

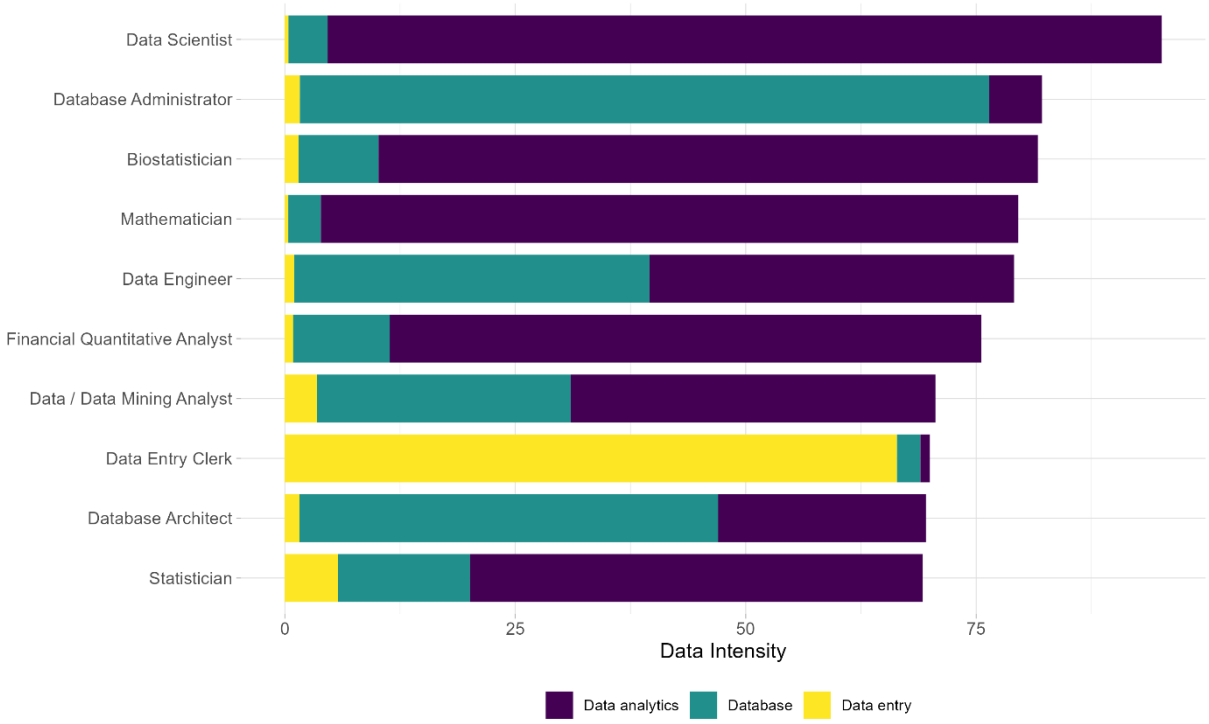
Per cent, 2020



Note: The Lightcast data provide occupation classifications. Data intensity takes values from 0-100.
Source: Authors' calculations based on LightCast data.

Figure B.2. Top 10 data-intensive occupations in the United States

Per cent, 2020



Note: The Lightcast data provide occupation classifications. Data intensity takes values from 0-100.
Source: Authors' calculations based on Lightcast data.

Table B.1. Data intensity at sectoral level for the United Kingdom, Canada and the United States, 2020

Sector	United Kingdom				Canada				United States			
	Data entry	Database	Data analytics	Data Intensity	Data entry	Database	Data analytics	Data Intensity	Data entry	Database	Data analytics	Data Intensity
Accommodation and food service activities	0.1%	0.3%	0.5%	**0.9%**	0.1%	0.2%	0.2%	**0.4%**	0.1%	0.3%	0.2%	**0.6%**
Administrative and support service activities	0.3%	1.0%	1.5%	**2.8%**	0.8%	0.6%	0.9%	**2.3%**	0.8%	1.0%	0.9%	**2.7%**
Agriculture forestry and fishing	0.5%	1.2%	4.5%	**6.1%**	0.5%	0.3%	1.0%	**1.8%**	0.4%	1.2%	2.1%	**3.7%**
Arts entertainment and recreation	0.3%	1.3%	2.4%	**4.0%**	0.3%	0.6%	1.0%	**1.9%**	0.2%	1.0%	0.9%	**2.1%**
Construction	0.3%	0.6%	0.8%	**1.7%**	0.3%	0.2%	0.3%	**0.8%**	0.2%	0.6%	0.4%	**1.2%**
Education	0.2%	0.4%	1.2%	**1.8%**	1.8%	1.8%	2.4%	**5.9%**	0.7%	1.7%	2.0%	**4.4%**
Electricity gas steam and air conditioning supply	0.6%	2.5%	5.7%	**8.8%**	1.5%	2.5%	2.9%	**6.9%**	0.5%	2.4%	2.7%	**5.7%**
Financial and insurance activities	0.8%	2.8%	5.7%	**9.3%**	1.8%	3.0%	5.0%	**9.9%**	0.8%	3.6%	5.3%	**9.8%**
Human health and social work activities	0.1%	0.2%	0.4%	**0.7%**	0.9%	0.6%	0.7%	**2.1%**	0.3%	0.5%	0.5%	**1.3%**
Information and communication	0.9%	4.1%	7.0%	**12.0%**	1.9%	3.7%	6.1%	**11.6%**	0.7%	5.0%	5.3%	**11.0%**
Manufacturing	0.4%	0.8%	1.7%	**2.9%**	0.9%	1.2%	1.8%	**3.9%**	0.4%	2.0%	2.7%	**5.1%**
Mining and quarrying	0.7%	1.7%	4.3%	**6.7%**	0.7%	1.2%	2.2%	**4.1%**	0.4%	1.6%	1.8%	**3.7%**
Other service activities	0.4%	0.9%	1.4%	**2.7%**	0.5%	0.4%	0.8%	**1.7%**	0.3%	0.6%	0.7%	**1.5%**
Professional scientific and technical activities	0.5%	1.3%	3.2%	**5.0%**	2.0%	4.1%	4.6%	**10.7%**	1.0%	6.1%	5.7%	**12.8%**
Public administration and defence	0.6%	1.1%	2.0%	**3.7%**	1.8%	1.9%	1.9%	**5.6%**	1.0%	2.2%	2.4%	**5.6%**
Real estate activities	0.4%	0.9%	1.3%	**2.6%**	1.3%	1.0%	1.8%	**4.0%**	0.6%	1.3%	1.7%	**3.5%**
Transportation and storage	0.3%	0.4%	0.8%	**1.5%**	0.7%	0.6%	0.8%	**2.1%**	0.3%	0.5%	0.6%	**1.4%**
Water supply sewerage waste management	0.2%	0.6%	1.1%	**2.0%**	0.9%	0.7%	0.7%	**2.2%**	1.0%	1.0%	0.7%	**2.7%**
Wholesale and retail trade	0.3%	0.8%	1.8%	**2.9%**	0.5%	0.5%	1.3%	**2.3%**	0.2%	0.5%	0.7%	**1.4%**
Total	0.37%	0.99%	2.04%	**3.4%**	0.94%	1.21%	1.76%	**3.9%**	0.60%	1.89%	2.11%	**4.6%**

Notes: Activities of households and activities of extraterritorial organisations have been excluded from the analysis. Data intensity takes values between 0-100%.
Source: Authors' calculations based on Lightcast data (2020).