The role of data in jobs

Julia Schmidt, Graham Pilgrim and Annabelle Mourougane

Trade and Productivity Statistics Division

Statistics and Data Directorate

5th International Seminar on Big Data for Official Statistics

May 2024

Roadmap

1. What is the OECD Innovation LAB?

2. Project “the role of data in jobs”

What is the Innovation LAB?

- The LAB was conceived as an incubator of innovative projects that can be directly applied to OECD work.

- 2 facets:
  - Build up the technical capacity of OECD staff
  - Help a community of researchers develop and deliver innovative projects

- Scope: innovative projects that can enrich the breath of our analysis, improve the timeliness and the granularity of data.

The role of data in jobs

Paper by Julia Schmidt, Graham Pilgrim and Annabelle Mourougane (Statistics and Data Directorate, OECD), financed by the UK Department of Business and Trade

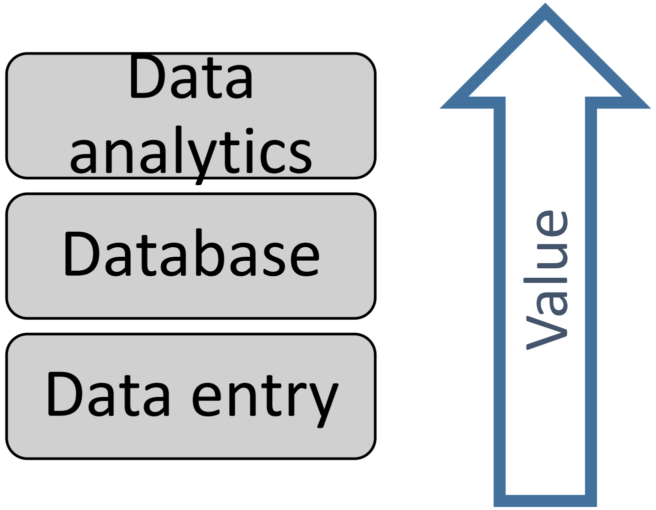
Link to the paper: https://doi.org/10.1787/fa65d29e-en

1. Methodology
2. Main insights
3. Conclusions

What have we done?

The approach

- Use NLP and online job advertisements to derive estimates of the data intensity of jobs in the United Kingdom, Canada and the United States, for the year 2020.
- We identify jobs that are involved in data production using the skills/tasks that are reported in the job advertisement.
- Follow the framework set out by Corrado et al. (2022) and Statistics Canada (2019)



A four-step approach

1. Extract from the job advertisement the skills/tasks that are related to the production of data using natural language processing

2. Classify the job as data-intensive or not, based on a set of rules including the extent to which the job description refers to data entry, database or data analytics activities

3. Aggregate data-intensive jobs to get estimates of shares by occupation, industry and at the economy-wide level

4. Inject those shares into a sum of cost approach to derive estimates of investment in data

The pros and cons of using online job advertisements (Lightcast)

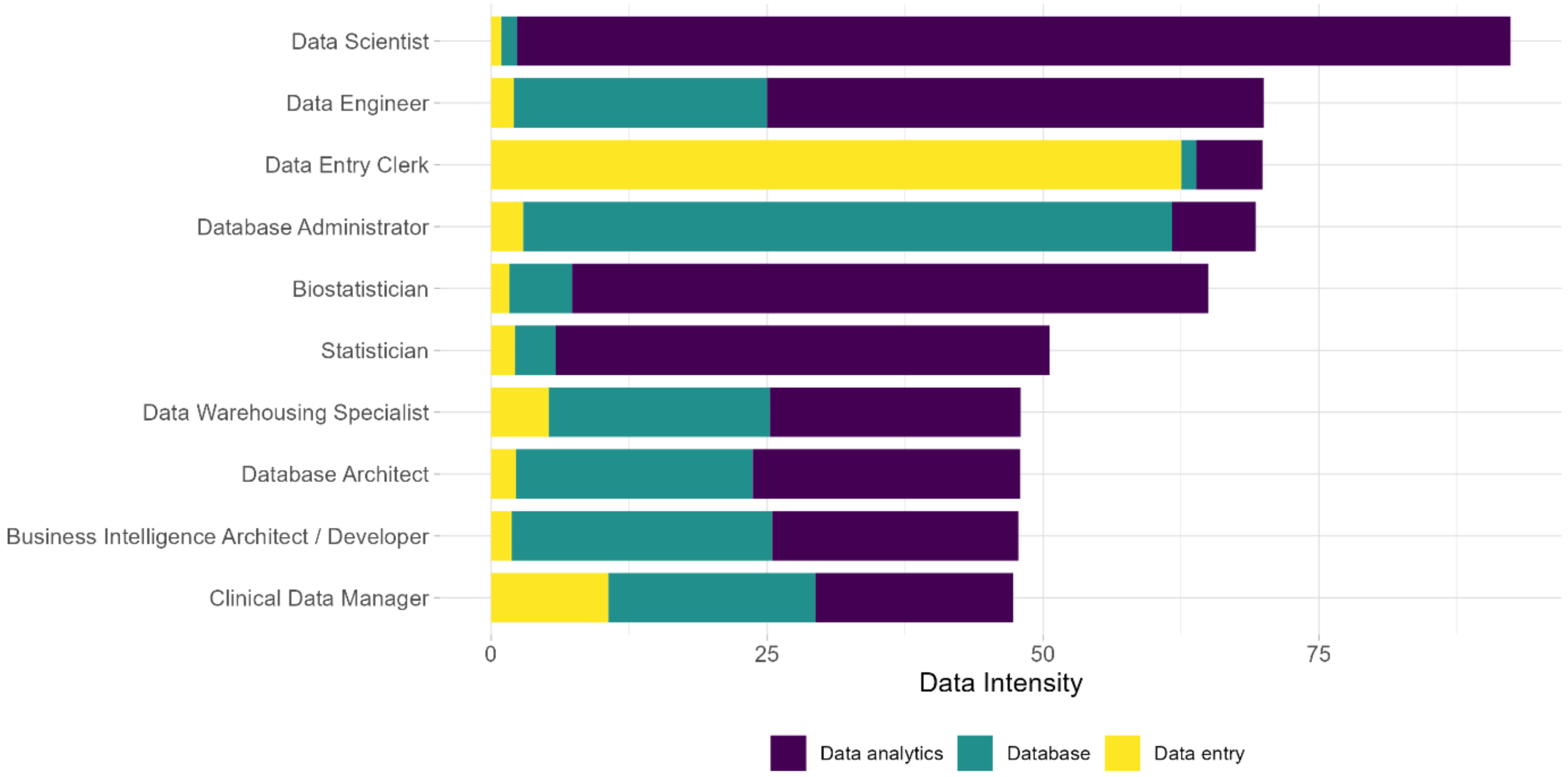
Job online advertisements are a measure of labour demand (flow as opposed to labour stock)

Advantages	Disadvantages
Timely data (2012 – present)	**Country coverage is limited** (GBR, CAN, USA, NZL, AUS as well as EU countries)
Linkage to firm-level and regional data	Limited coverage depending on year and country, no insights on how firms hire
Standardised occupation and industry classifications	**Representativeness** is heterogeneous (industry, occupation level; white collar jobs)
Identify **skill demands beyond standard labour market statistics**	No information about quantity of hiring Recruitment agencies cause duplications

What have we found?

The most data-intensive occupations are those that use data analytics skills

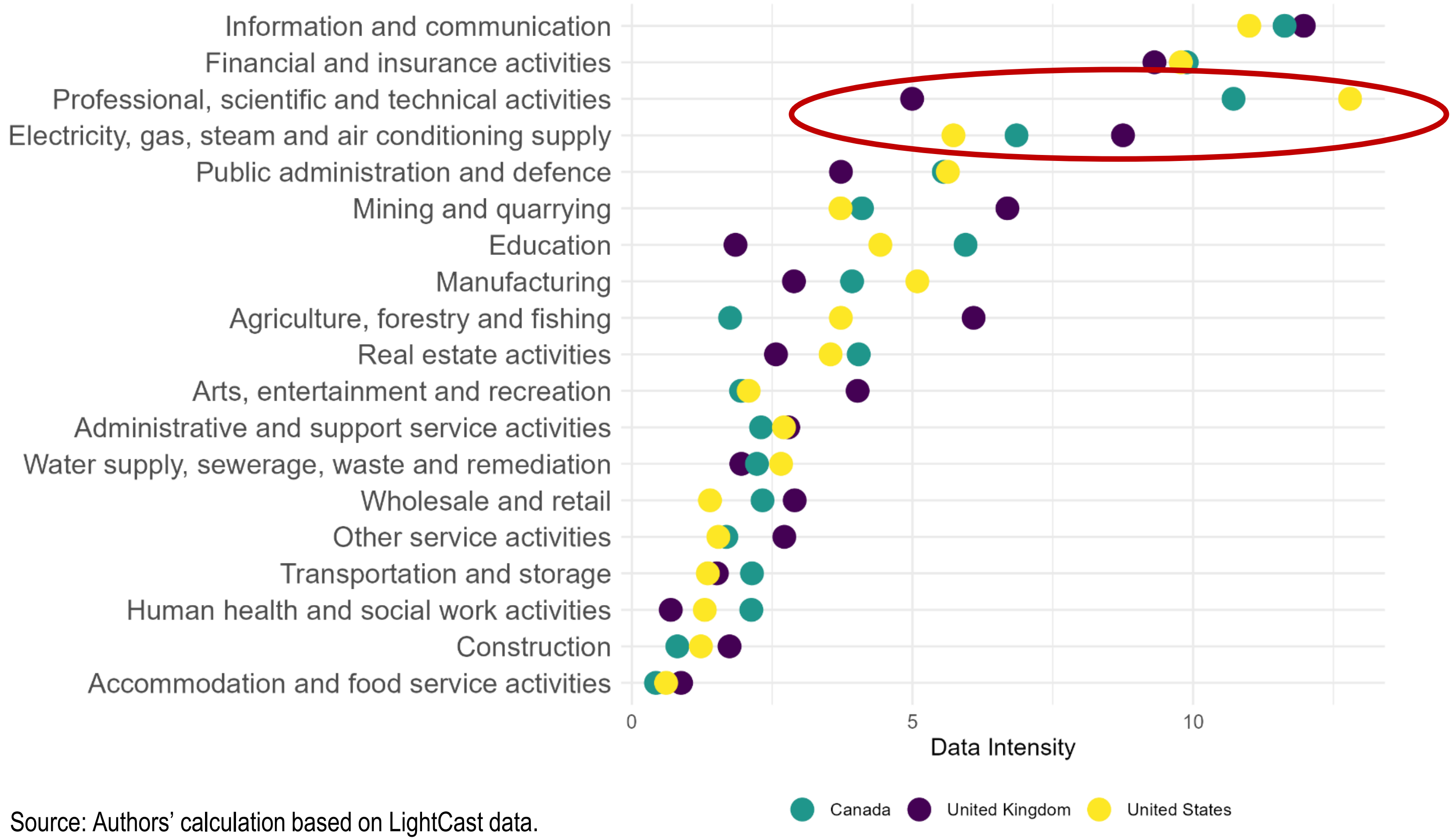
Top 10 data-intensive occupations in the United Kingdom, per cent, 2020



Source: Authors' calculation based on LightCast data.

Differences in data intensity across the countries are concentrated in a handful of industries

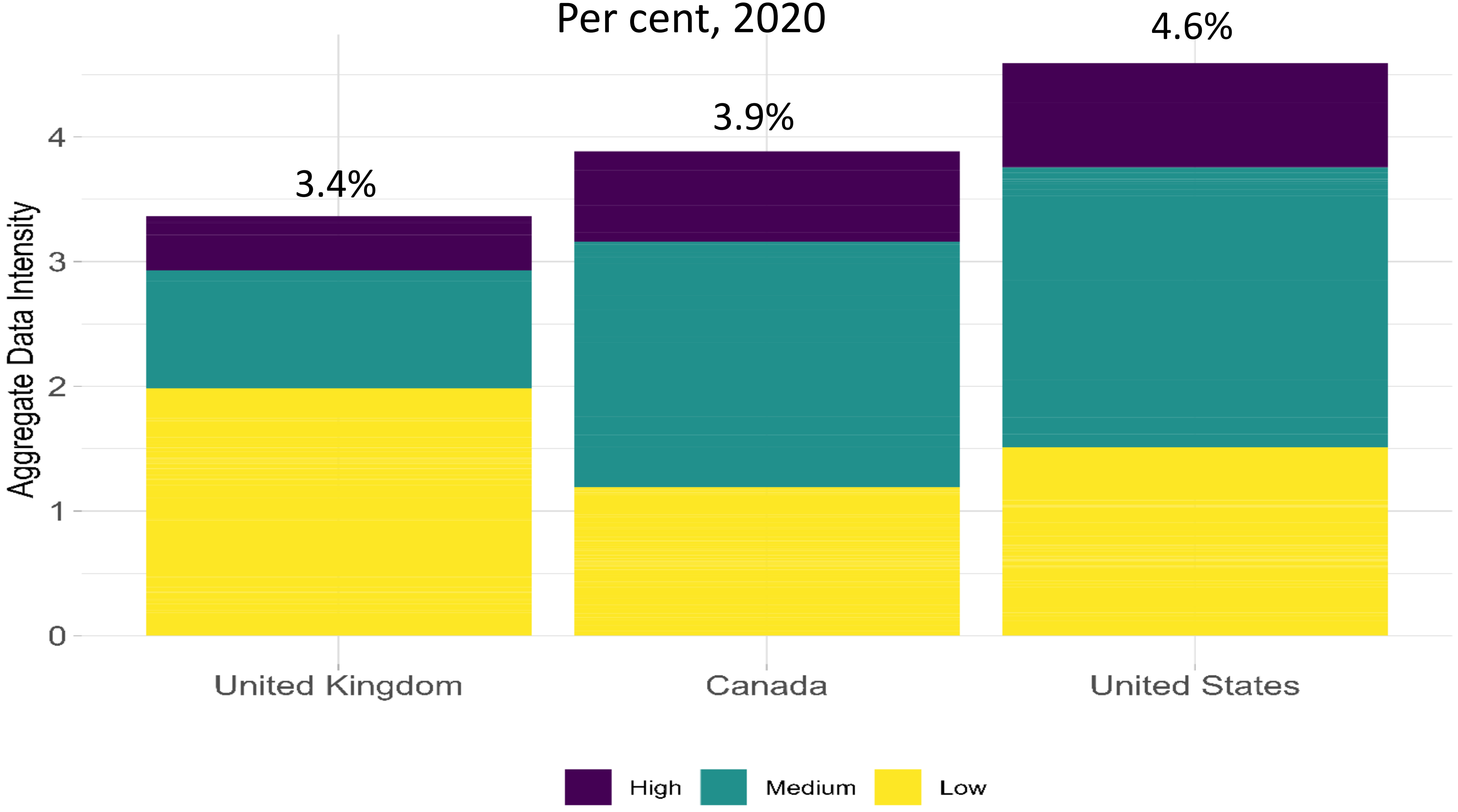
Data intensity at industry level, per cent, 2020



Source: Authors' calculation based on LightCast data.

The United Kingdom and Canada appear to be less data-intensive than the United States

Per cent, 2020



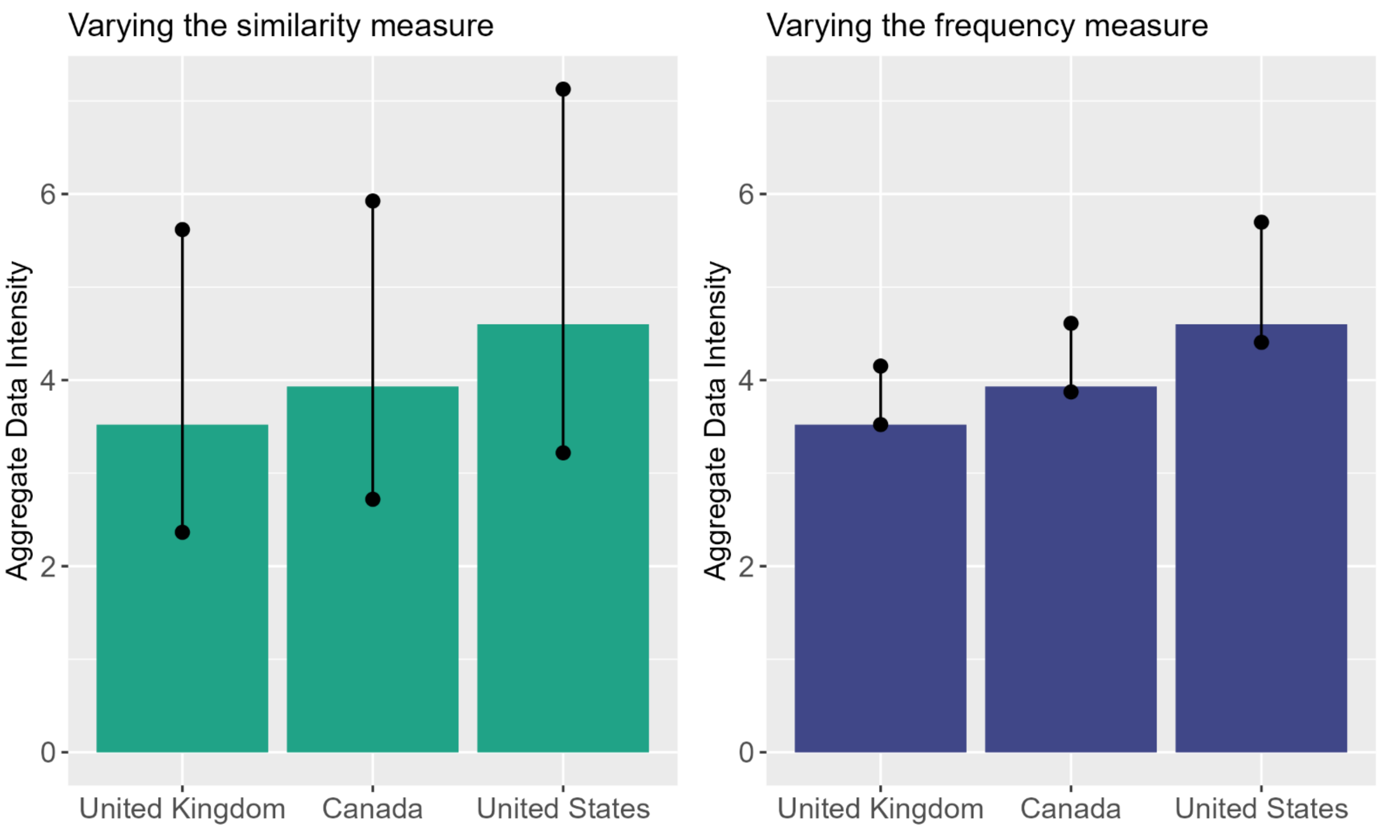
Notes: Data intensity takes values between 0 and 100. Low data-intensive occupations: 0<10%, medium data-intensive occupations: 10-50% and high data-intensive occupations > 50%

Results at aggregate level are sensitive to changes in the classification rule

- Order of magnitude of results broadly similar

- Estimates seem to depend on the similarity parameter we use in the model, but not on the other parameters

- The ranking between countries does not seem to be affected



Source: Authors' calculation based on LightCast data.

A sum-of-cost approach for investment in data

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

d : type of job (data entry, database or data analytics), i : industry

α : mark-up (non-wage cost and a margin for capital services) → use a range

Remark: Use a data intensity share based on the frequency of skills requirements and tasks contained in job ads rather a time-use factor.

Caveats:

- Capital and intermediary consumption not considered
- Uniform wages within industry

Estimates of investment in data using data intensity shares

Investment in data in 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, pp	0.7 - 1.5	0.3 - 0.7	0.5 - 1.1
Database, pp	1.0 - 2.2	0.9 - 2.0	2.0 - 4.1
Data analytics, pp	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 STATCAN (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)

Conclusions

- The paper has developed an NLP methodology using online job postings to derive estimates of data intensity of jobs, which can be used to get insight on investment in data.

- The approach is subject to a number of limitations, including sensitivity to some of the assumption in the NLP classification rules and in the use of a sum-of-cost approach, although the country ranking appears to be robust.

- Further research:
  - Expanding the time and country coverage
  - Use the methodology to derive estimates of AI-hiring intensity

Thanks!

@OECD_Stat in OECD Statistics www.oecd.org/sdd www.stats.oecd.org

OECD
BETTER POLICIES FOR BETTER LIVES