Using NLP to detect data/AI hiring intensive jobs and firms

Julia Schmidt (presenting), Graham Pilgrim, Annabelle Mourougane

13th June 2024

Motivation and scope

Objective

- Estimate AI/data-intensive jobs for the United Kingdom 2012-2022
- For 2015-2021 match data to two firm-level databases to generate insights on productivity/export behaviour

Methodology

- Develop a natural language processing algorithm on online job advertisements
- Classify jobs into data, as well as AI-related jobs

The pros and cons of using online job advertisements

- Provided by Lightcast data, previously BurningGlass Technologies
- Job online advertisements are a measure of labour demand (flow as opposed to labour stock)
- No information about the quantity of hiring
- Recruitment agencies cause duplications

Advantages	Disadvantages
Timely data (2012 – present)	Decreasing quality of the data the further back in the time (e.g., 2012 data are of worse quality than 2023 data)
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	

What are data/AI-intensive jobs?

- A data intensive job can be defined based on the data value chain concept (Corrado et al. (2022) and Statistics Canada (2019))
- An AI intensive job is related to generic and specific, AI related skills (Borgonovi et al. 2023)

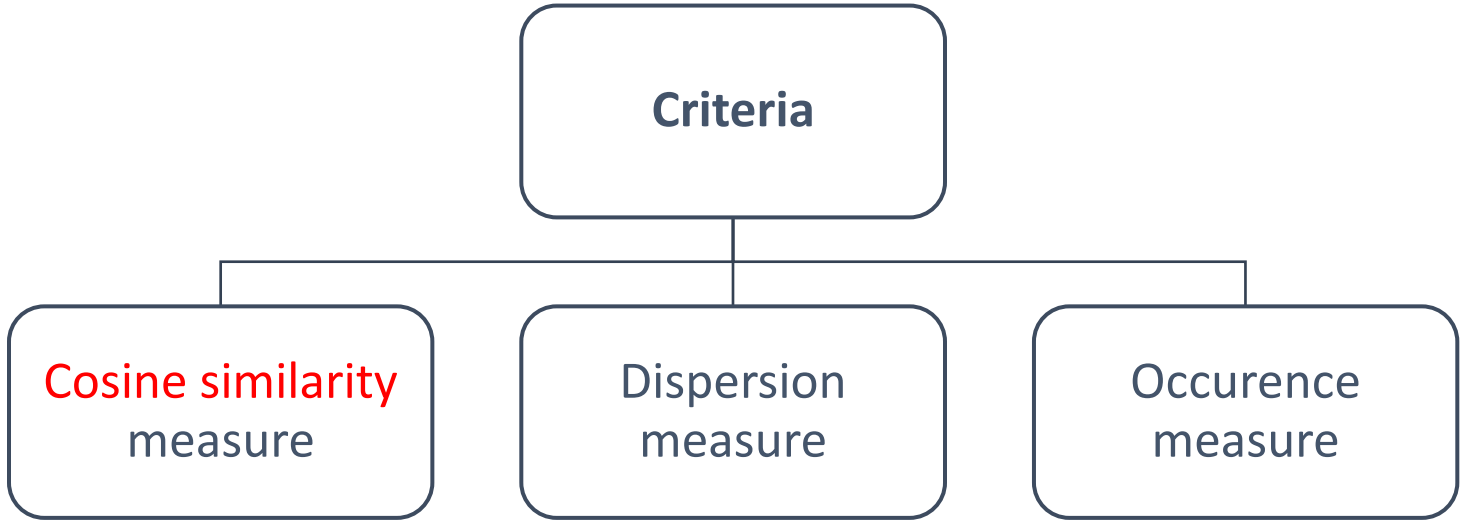
Methodology

Using natural language processing to estimate data/AI intensity

1. Process the text data from the job advertisements
2. Extract skills/tasks that identify the job as involved in data production/related to AI using natural language processing
3. Classify the job based on its link to data entry, database or data analytics activities or AI-related skills
4. Aggregate the jobs to occupation, firm, industry, and economy level (data/AI intensity score per occupation/firm/industry/economy)

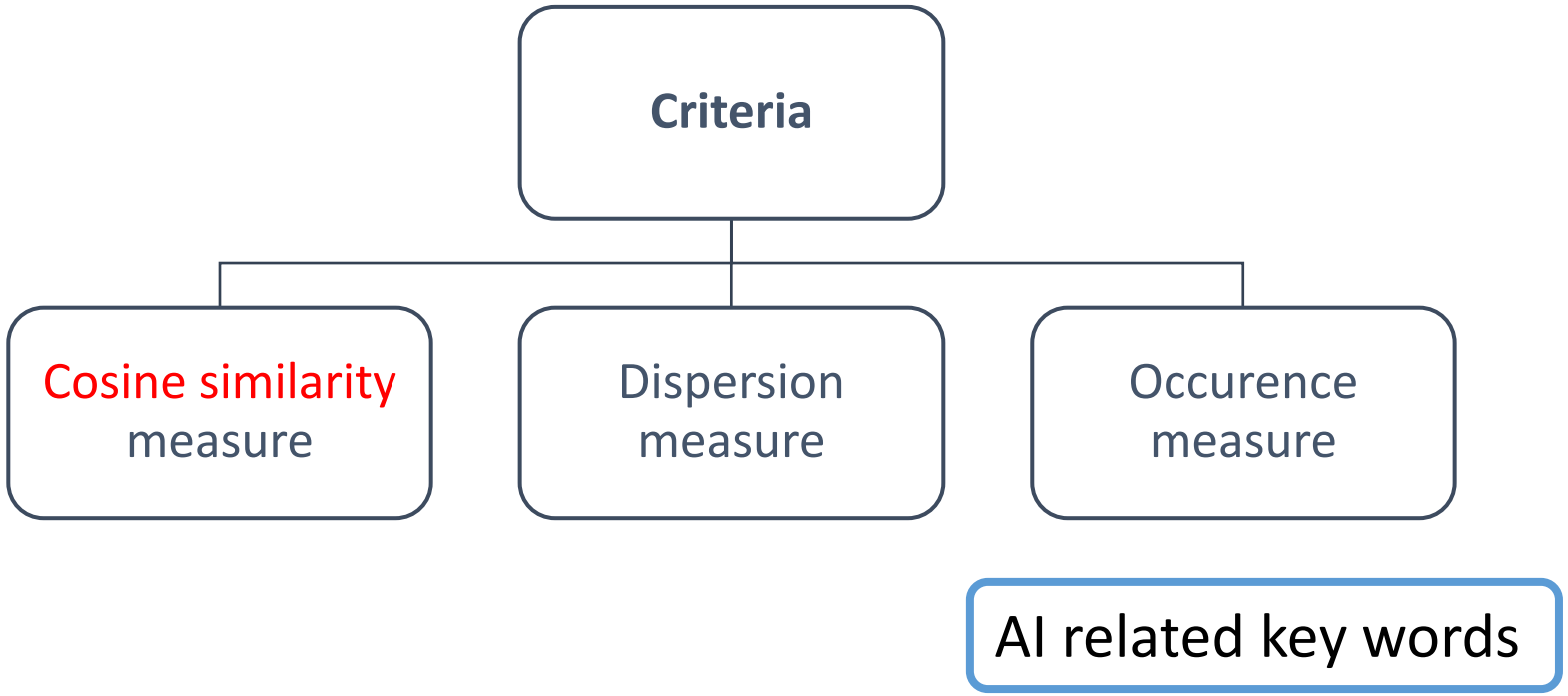
Classification and aggregation of data-intensive jobs

- A job is classified as data-intensive (1, else 0) if it passes the following criteria:



Classification and aggregation of AI-related jobs

- A job is classified as AI-intensive (1, else 0) if it passes the following criteria:

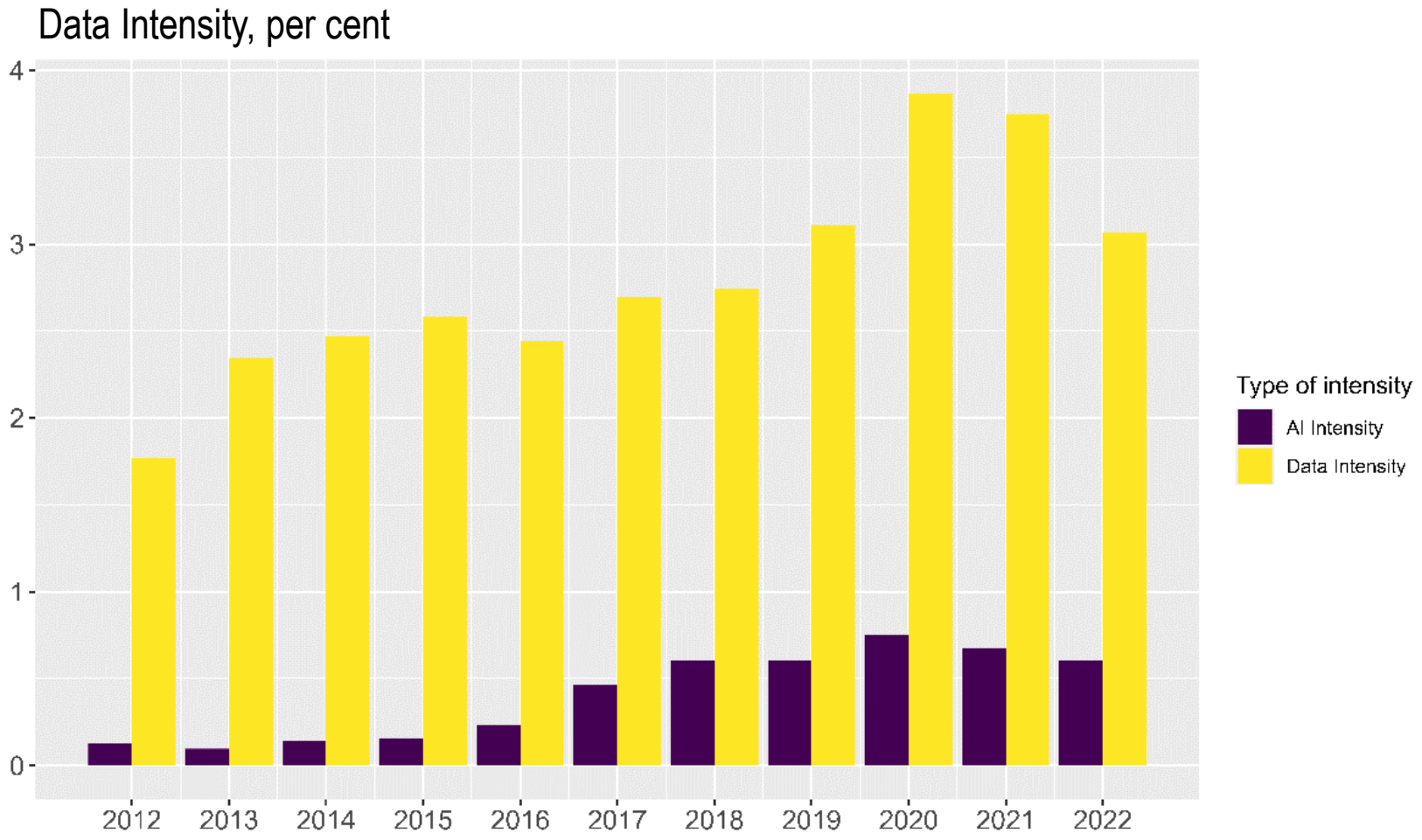


Matching Orbis to Lightcast data

- Matched Orbis data for the United Kingdom to Lightcast data
- Applied OpenCorporates OpenRefine Reconciliation API (version 0.4.8) to generate company IDs
- Facilitated matching (matching on ID vs. matching on names)
- Matching rates for 2022: validation via OpenCorporates (retained 77% of original Lightcast firms → 22% when matching to Orbis)

Results

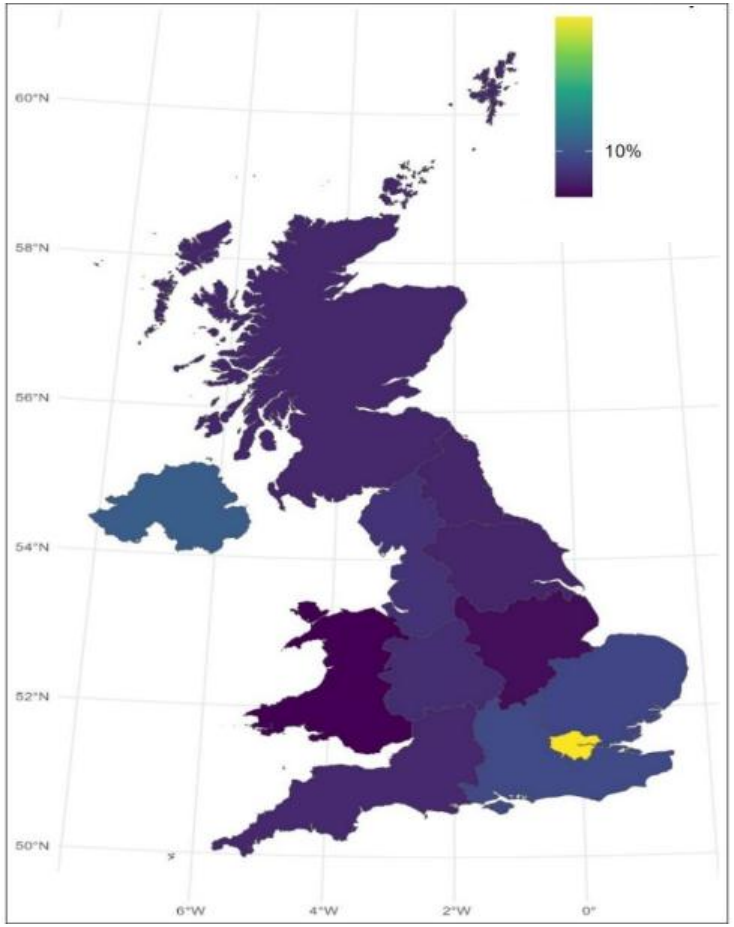
Data and AI-hiring intensity peaked during COVID in the UK



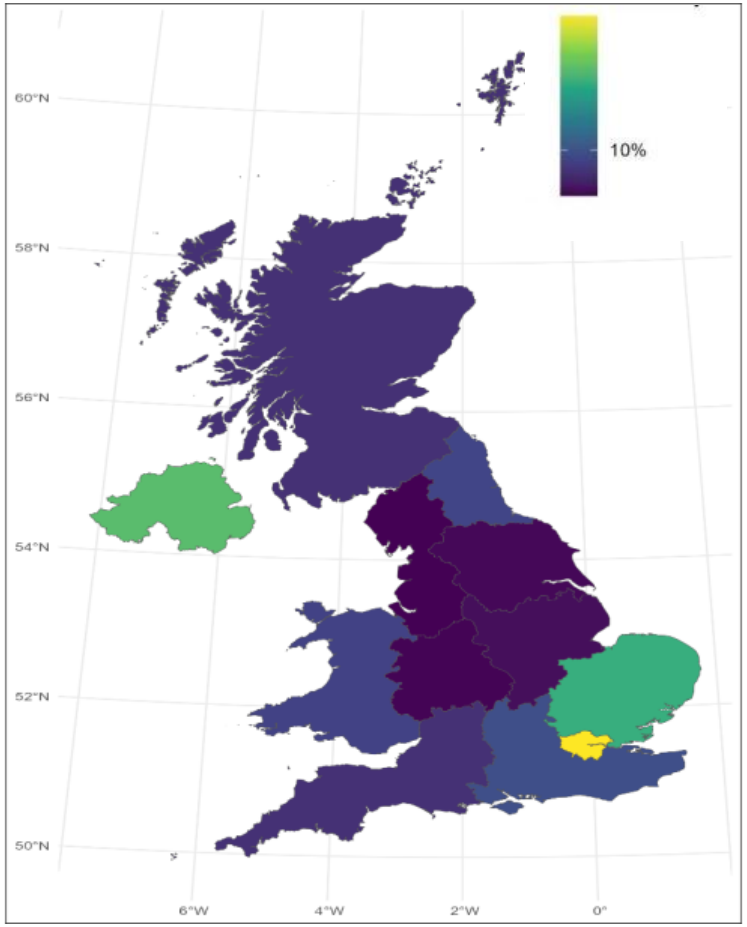
Source: Authors' calculations based on Lightcast data.

Data/AI intensive jobs are concentrated in London

UK regions, demand for data skills, 2022



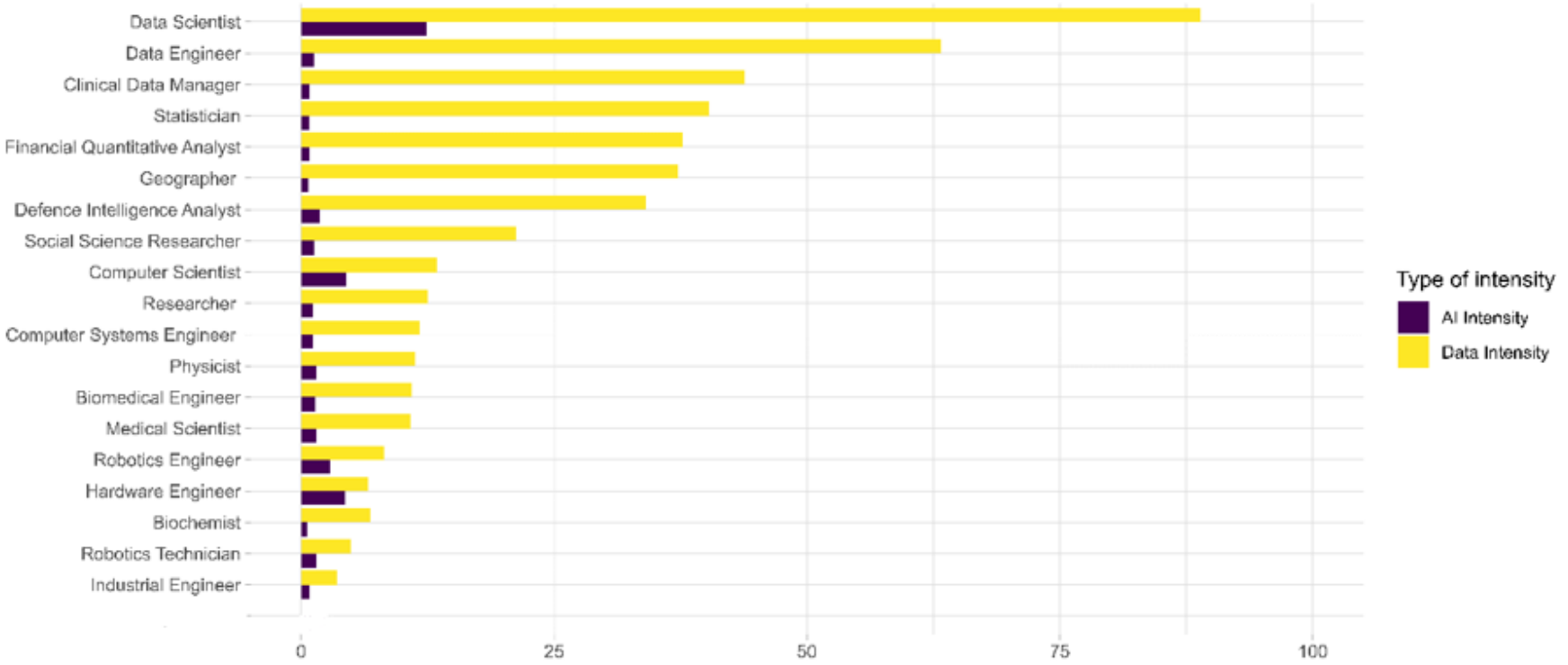
UK regions, demand for AI skills, 2022



Source: Authors' calculations based on Lightcast data.

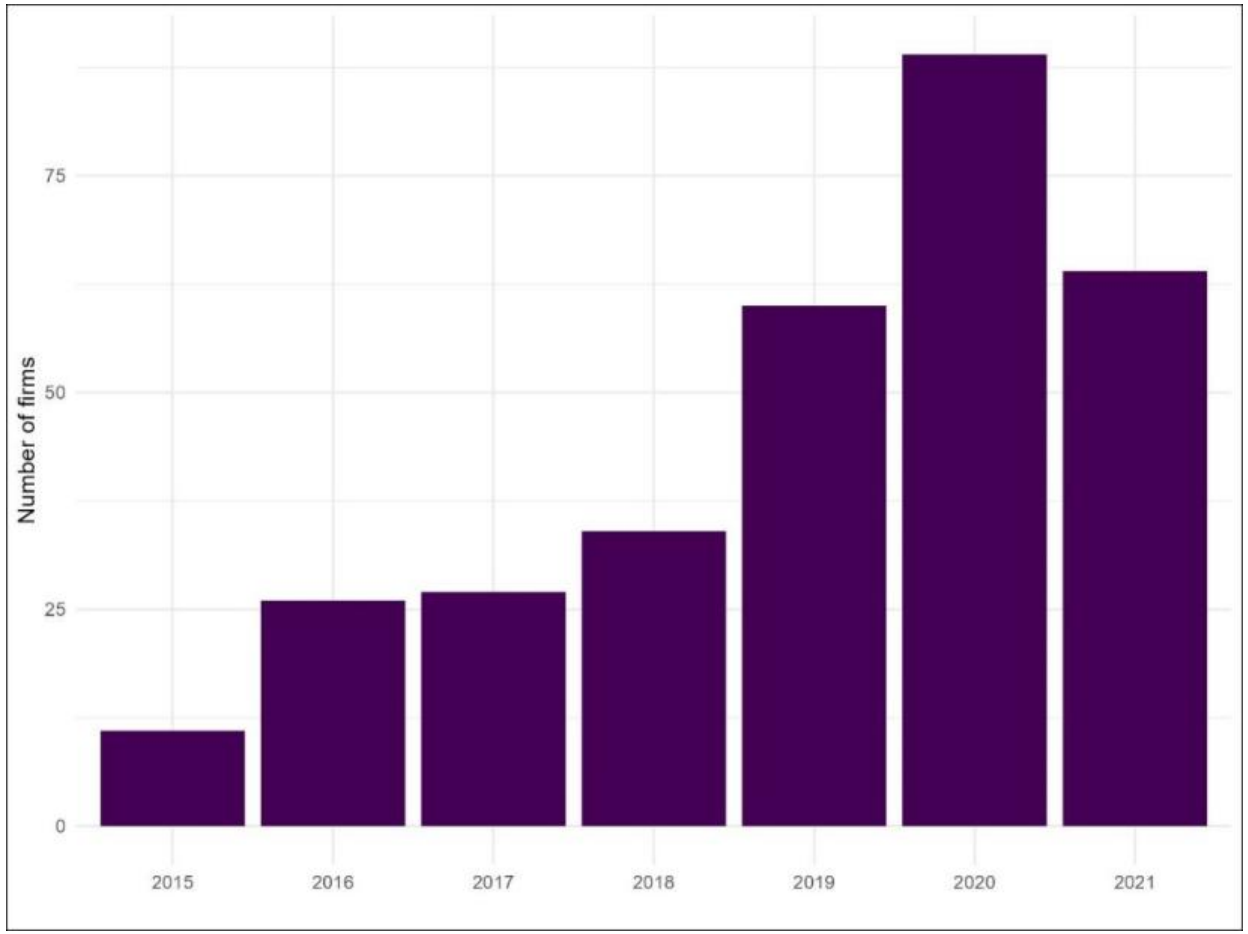
Demand for data/AI skills by occupation differs

Data and AI-hiring intensity in per cent, 2022



Source: Authors' calculations based on Lightcast data.

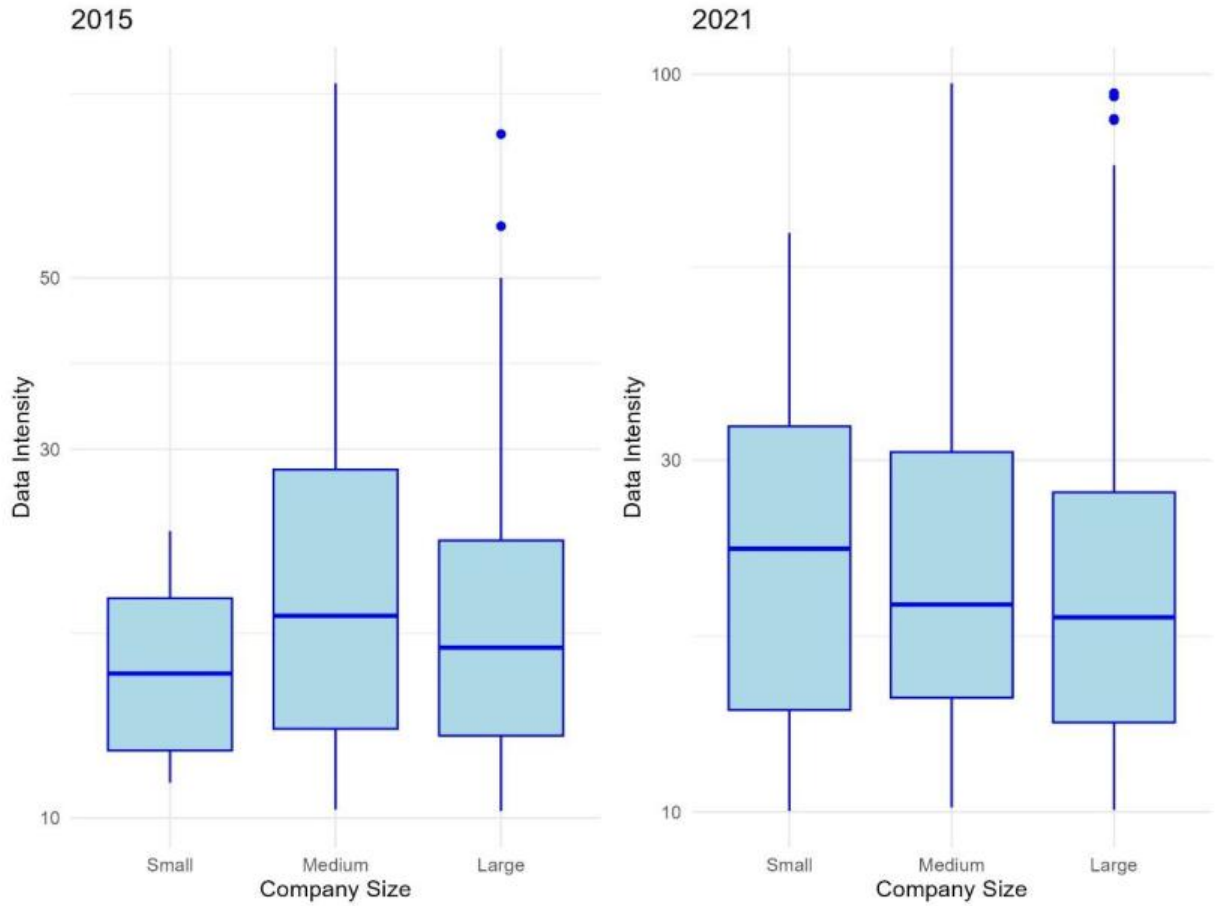
Number of highly data intensive firms increased



Source: Authors' calculations based on Lightcast data.

The group of data-intensive firms is heterogenous

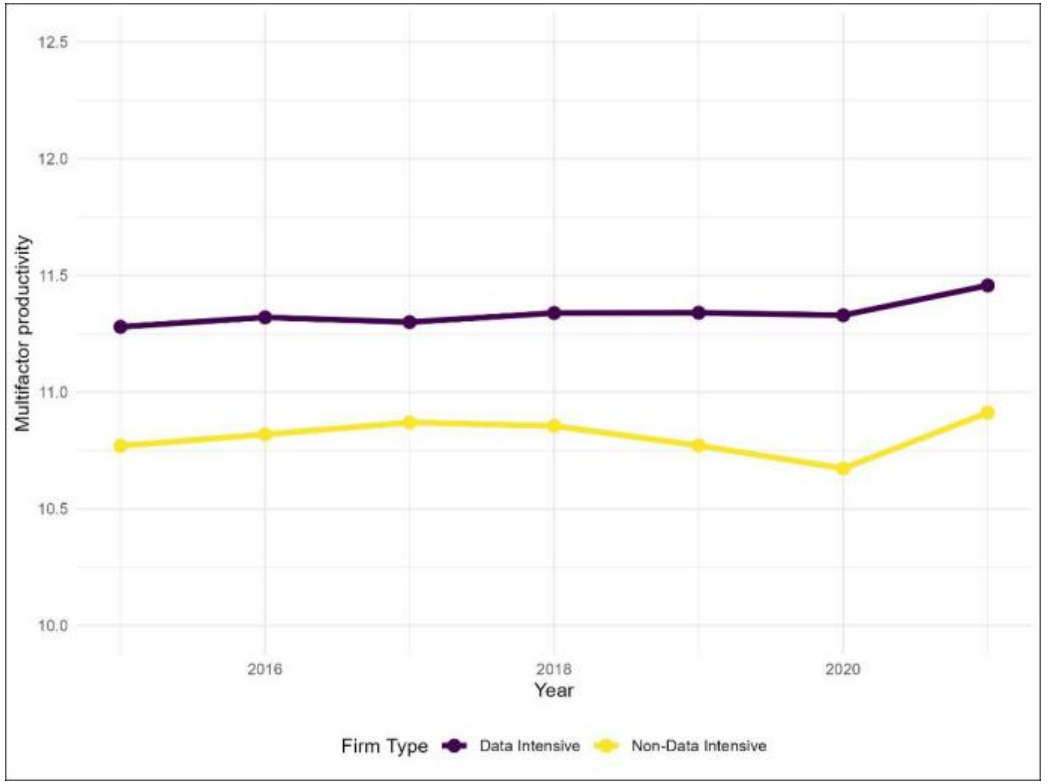
Data intensity in per cent



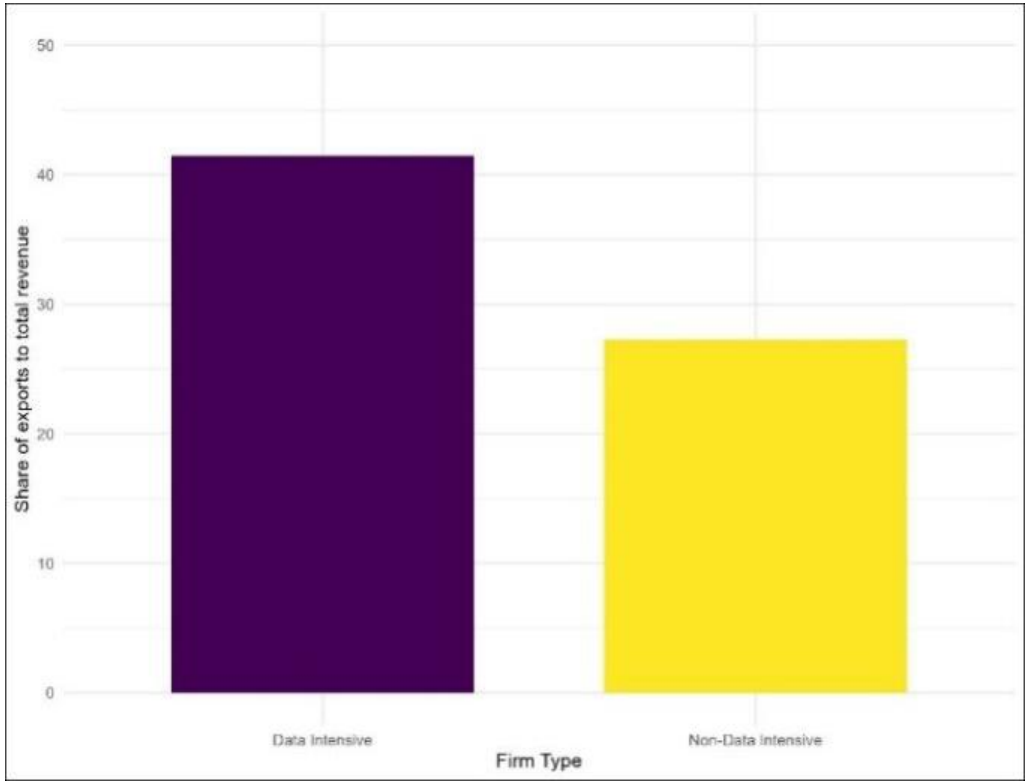
Source: Authors' calculations based on Lightcast and Orbis data.

Data-intensive companies are more productive and export

Average multifactor productivity levels, index

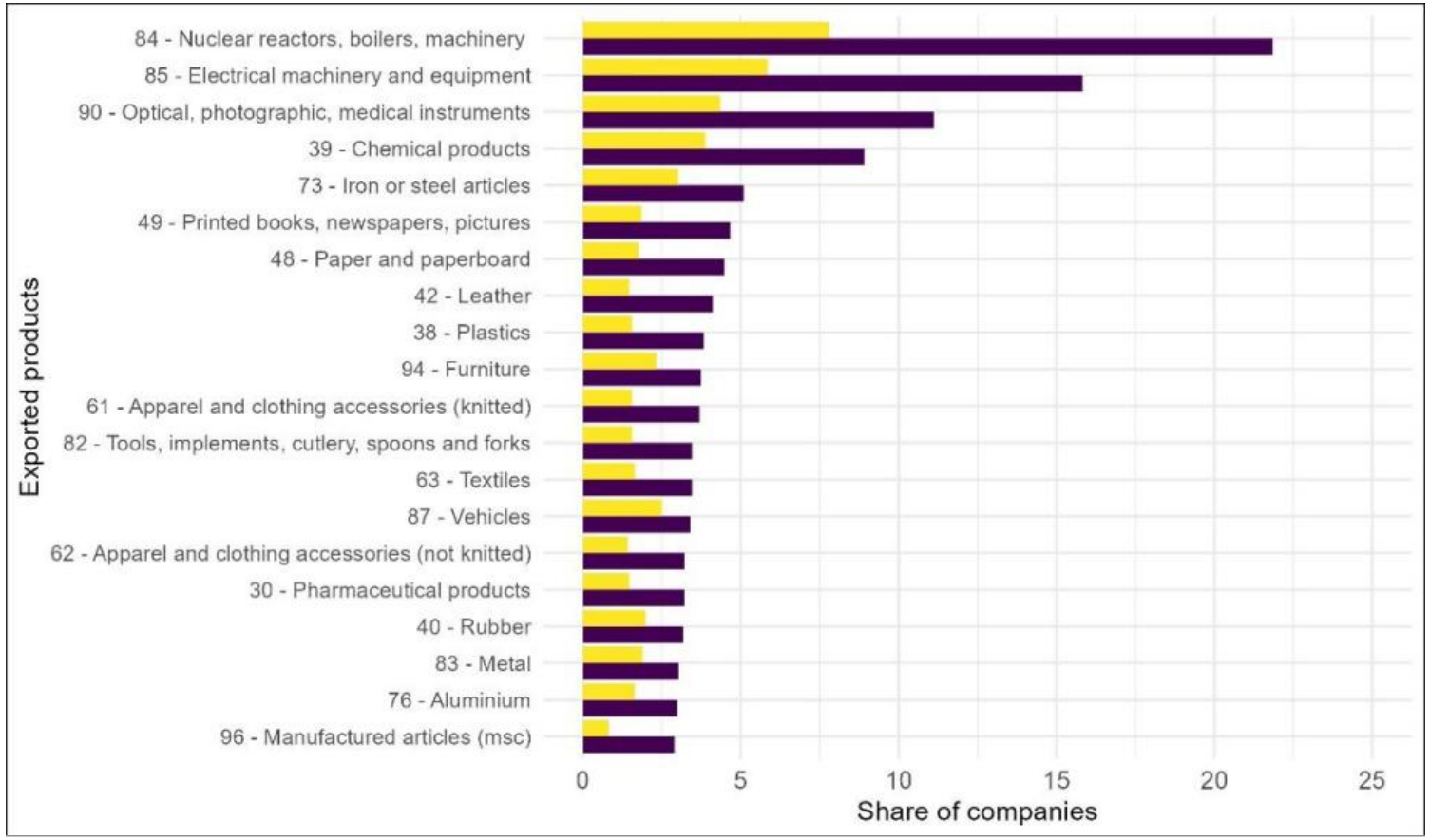


Share of exports/total revenue, per cent, average 2015-2022



Source: Authors' calculations based on Lightcast and Orbis data.

Data and non-data firms differ in the types of products traded



Source: Authors' calculations based on Lightcast/Orbis data and the UK trader dataset.

Take aways and new ideas

Contribution of our work

- Flexible NLP algorithm that can be extended across countries and time series; to 66 languages, as well as beyond digital skills (e.g. green skills)
- Disaggregated insights into digital skills on labour markets (data as well as AI skills)
- Combine several data sources (online job advertisements, and two-firm-level databases)
- Provide a new methodology on how to match ORBIS with Lightcast data

Future work

- Extend the work to trade in services
- Expand the matching exercise to additional countries (ongoing at the OECD)

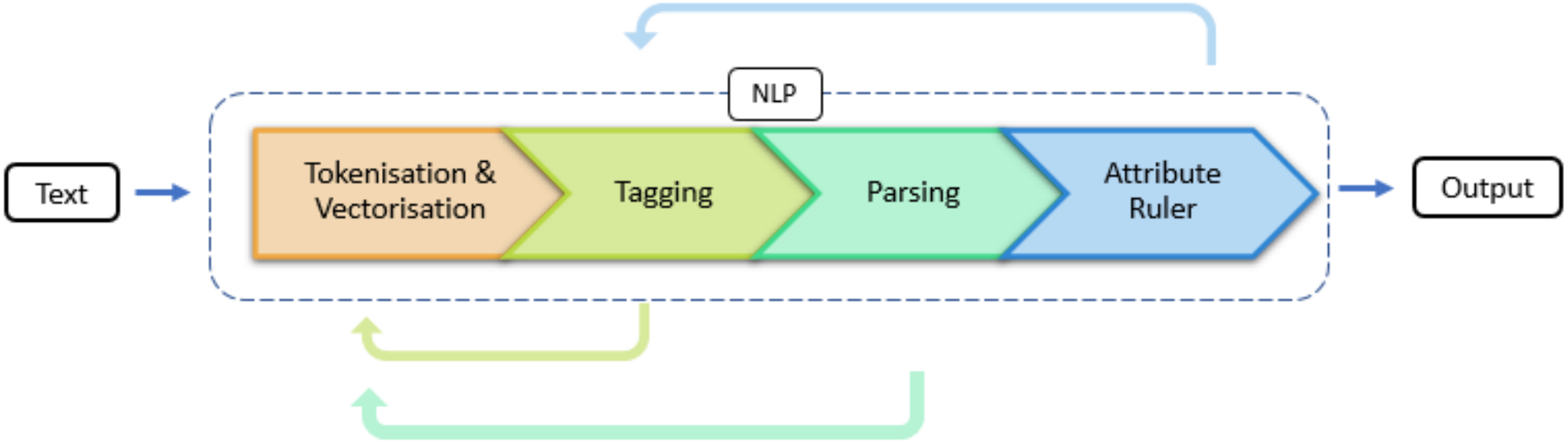
Questions?

References

- ADB (2022), Digital jobs and digital skills. A shifting landscape in Asia and the Pacific, http://dx.doi.org/10.22617/SPR220348.
- Bellatin, A. and G. Galassi (2022), *What COVID-19 May Leave Behind: Technology-Related Job Postings in Canada*, https://www.bankofcanada.ca/2022/04/staff-working-paper-2022-17/.
- 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.
- 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.
- Cameraat, E. and M. Squicchiarini (2021), Burning Glass Technologies' data use in policy-relevant analysis: An occupation-level assessment, OECD Publishing, https://dx.doi.org/10.1787/cd75c3e7-en.
- 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.
- 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.
- Calvino, F. et al. (2018), *A taxonomy of digital intensive sectors*, OECD Publishing, https://doi.org/10.1787/f404736a-en.
- Garasto, S. et al. (2021), *Developing experimental estimates of regional skill demand*, Economic Statistics Centre of Excellence, https://www.escoe.ac.uk/publications/developing-experimental-estimates-of-regional-skill-demand/.spaCy (2022), Language Processing Pipelines, https://spacy.io/usage/processing-pipelines.
- Muro, M. et al. (2017), Digitalization and the American Workforce, https://www.brookings.edu/research/digitalization-and-the-american-workforce/.
- 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, 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.
- 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.
- 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

Process the online job advertisements

- NLP captures the **meaning and structure of** a word/sentence in different contexts



Source: Authors' illustration based on (spaCy, 2022[43])

Tokenising and vectorising online job advertisements

Tokenisation

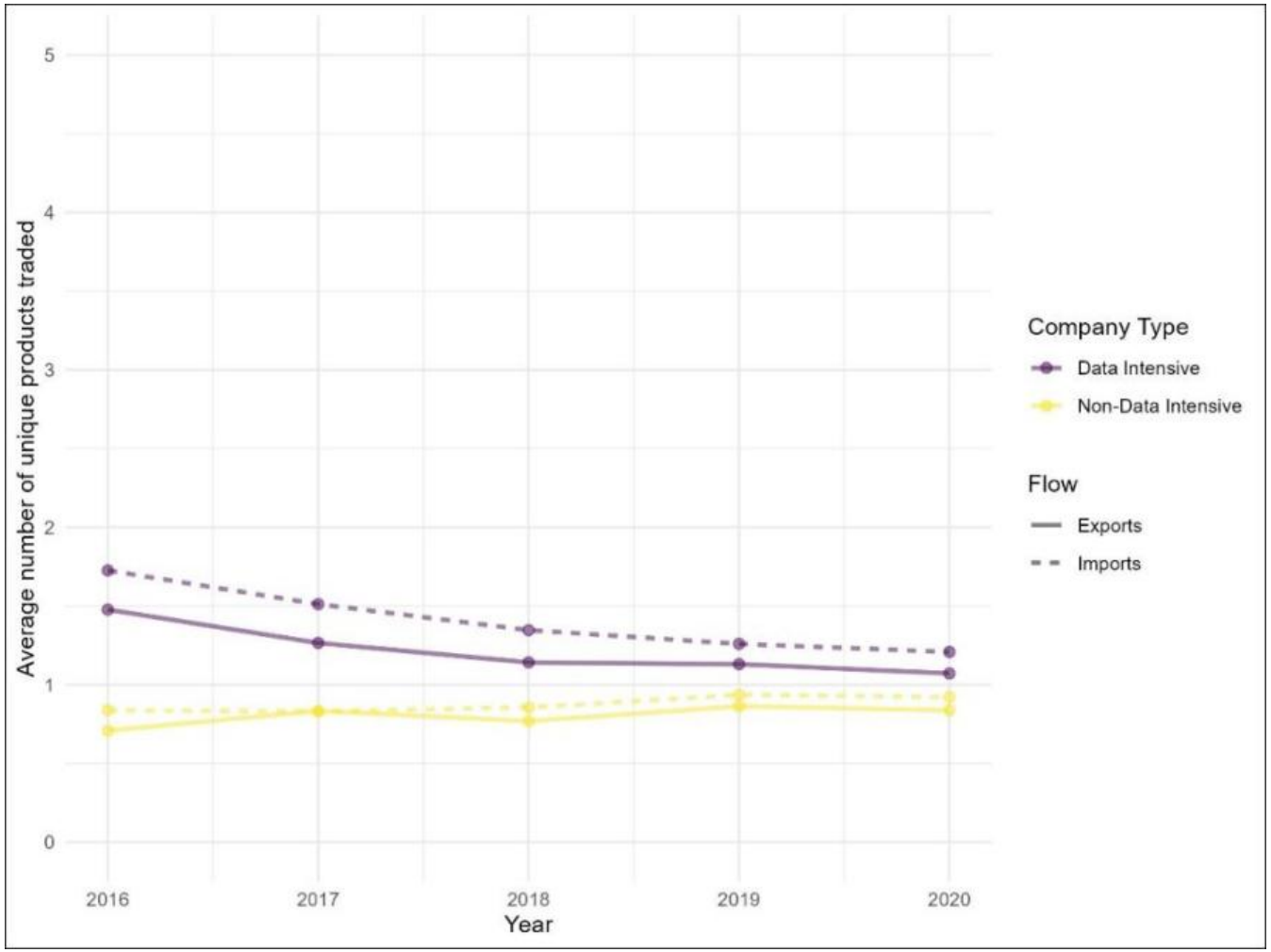
"A data scientist is a high-skilled professional who uses analytical, statistical and programming knowledge skills to analyse large datasets."

- data scientist
- high-skilled professional
- analytical statistical programming knowledge skills
- analyse large datasets

Vectorisation

```
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]
```

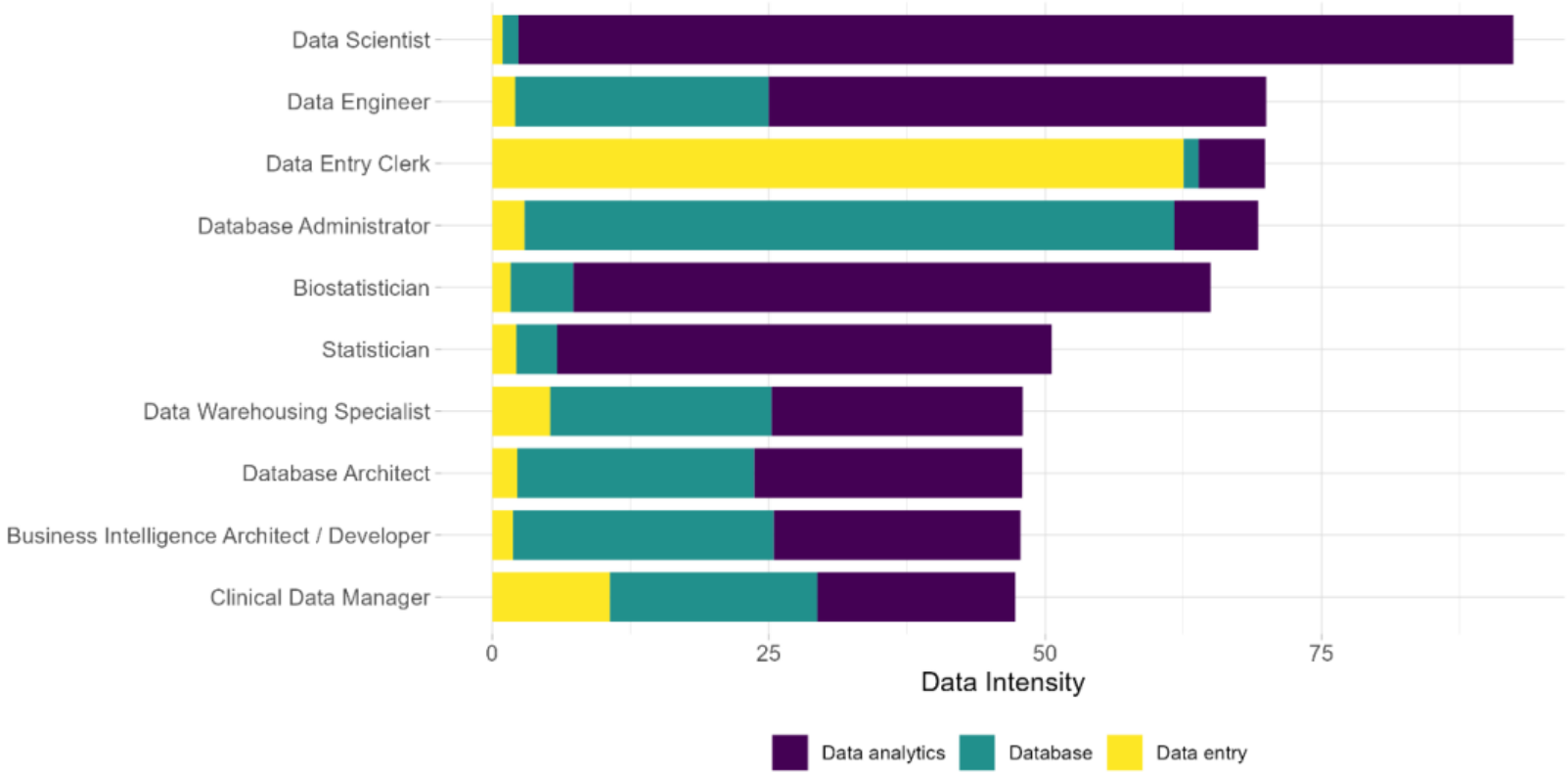
The average number of products traded are similar across firms



Source: Authors' calculations based on Lightcast/Orbis data and the UK trader dataset.

High data-intensive professions are linked to data analytics

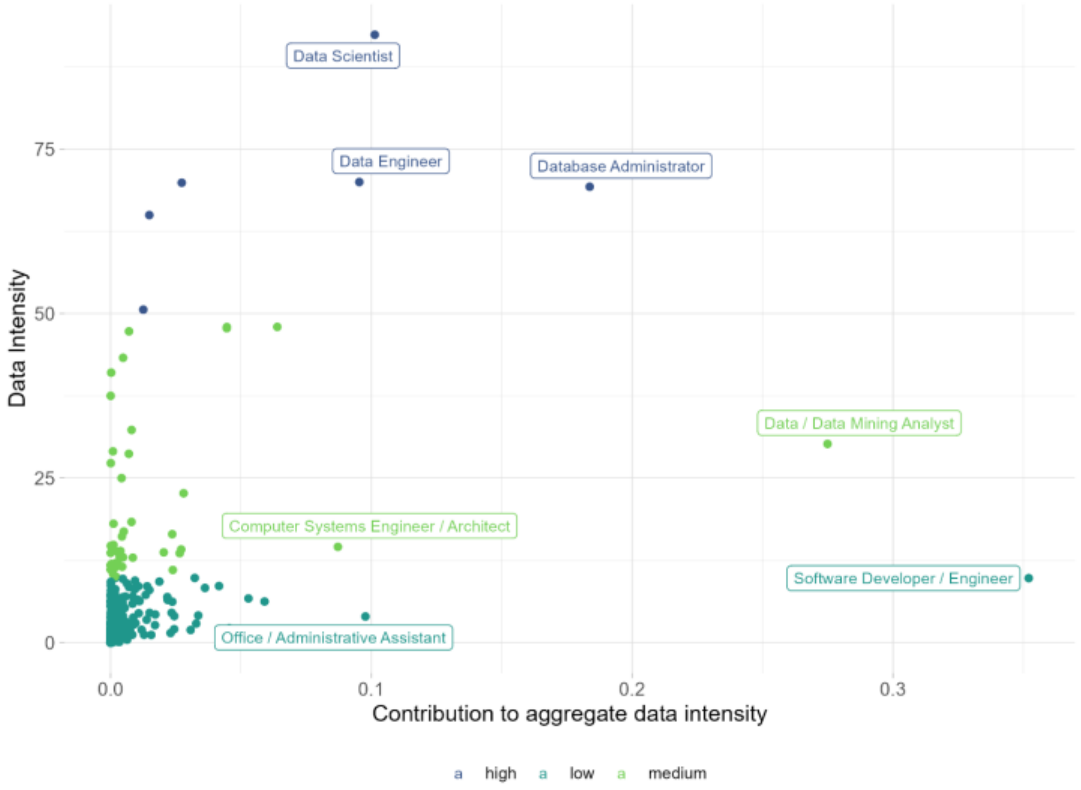
Top 10 data-intensive occupations in the United Kingdom, per cent, 2020



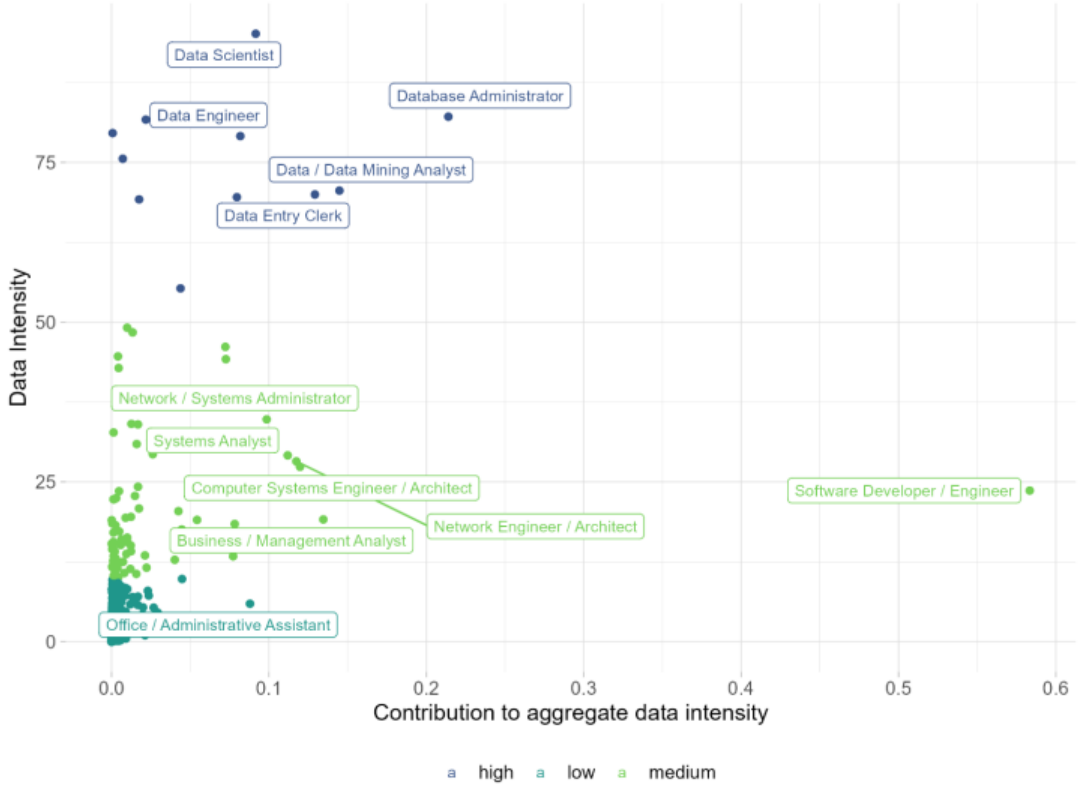
Source: Authors' calculation based on LightCast data.

The distribution of data professions is highly unequal

A - United Kingdom, per cent, 2020



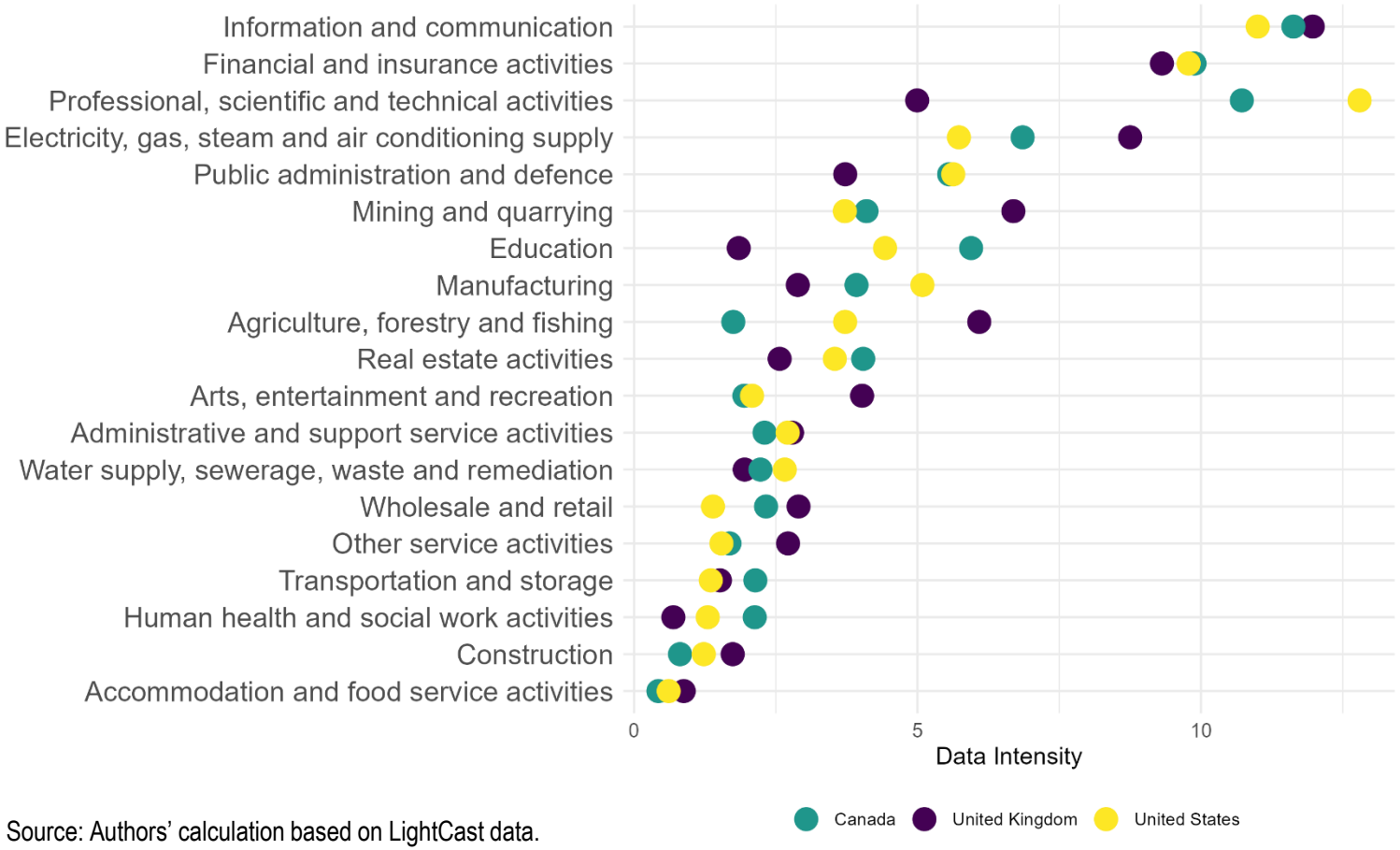
B- United States, per cent, 2020



Source: Authors' calculation based on LightCast data.

Differences across countries are bigger at industry level

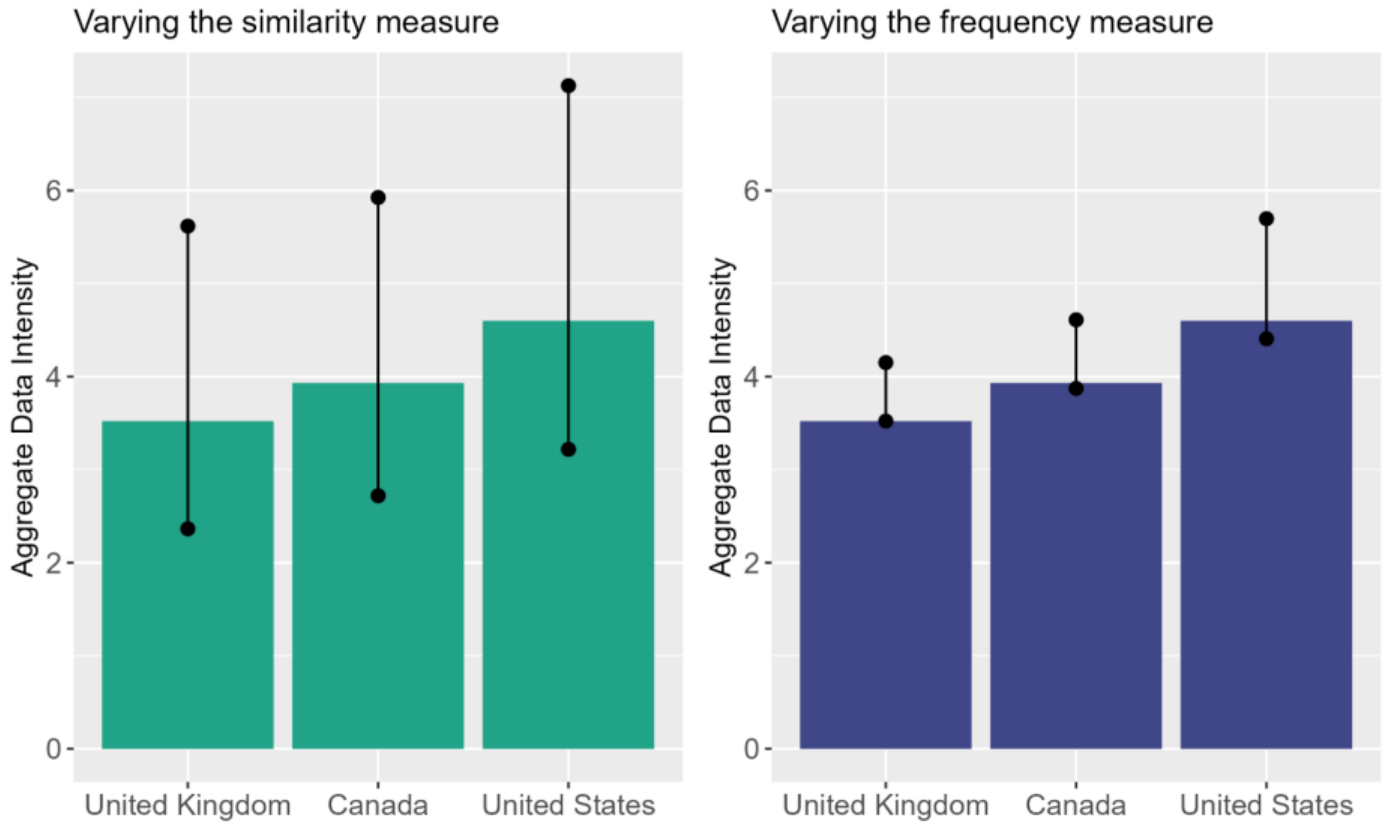
Data intensity at industry level, per cent, 2020



Source: Authors' calculation based on LightCast data.

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

- Careful calibration of classification rule
- Order of magnitude of results remains stable
- Results vary larger for changes in similarity measure, but with the same magnitude across countries



Source: Authors' calculation based on LightCast data.