OECD Science, Technology and Industry Working Papers
2023/01

Identifying artificial intelligence actors using online data

Hélène Dernis,
Flavio Calvino,
Laurent Moussiegt,
Daisuke Nawa,
Lea Samek,
Mariagrazia Squicciarini

https://dx.doi.org/10.1787/1f5307e7-en

OECD

OECD Science, Technology and Industry Working Papers

This publication contributes to the OECD's Artificial Intelligence in Work, Innovation, Productivity and Skills (AI-WIPS) programme, which provides policymakers with new evidence and analysis to keep abreast of the fast-evolving changes in AI capabilities and diffusion and their implications for the world of work. The programme aims to help ensure that adoption of AI in the world of work is effective, beneficial to all, people-centred and accepted by the population at large. AI-WIPS is supported by the German Federal Ministry of Labour and Social Affairs (BMAS) and will complement the work of the German AI Observatory in the Ministry's Policy Lab Digital, Work & Society. For more information, visit *https://oecd.ai/work-innovation-productivity-skills* and *https://denkfabrik-bmas.de/*.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CIIE/WPIA(2020)5/FINAL

Table of contents

Tables

Figures

Boxes

Acknowledgements

The authors are grateful to Sarah Box, Chiara Criscuolo, Antoine Dechezleprêtre and Dirk Pilat for helpful comments and discussions on the early draft; Sergi Martorell, co-founder and CEO of GlassAI, for support and data provisions; Antonio Ughi for excellent research assistance on the literature review section; Leonidas Aristodemou for his precious help in consolidating the list of AI topics; and participants to the OECD Working Party on Industry Analysis (WPIA) as well as to the Committee on Industry, Innovation and Entrepreneurship (CIIE).

This work contributes to the OECD programme on AI in Work, Innovation, Productivity and Skills (AI-WIPS), supported by the German Federal Ministry of Labour and Social Affairs (BMAS).

Mariagrazia Squicciarini contributed to this work while serving as a Senior Economist and Head of Unit at the OECD Directorate for Science Technology and Innovation. She is now Chief of Executive Office and Director a.i. at UNESCO's Social and Human Sciences Sector. Daisuke Nawa contributed to the analysis while he was detached from the Japan Patent Office (JPO) to the OECD.

Abstract

This paper uses information collected and provided by GlassAI to analyse the characteristics and activities of companies and universities in Canada, Germany, the United Kingdom and the United States that mention keywords related to Artificial Intelligence (AI) on their websites. The analysis finds that those companies tend to be young and small, mainly operate in the information and communication sector, have AI at the core of their business, and aim to provide customer solutions. It is noteworthy that the types of AI-related activities reported by them vary across sectors. Additionally, although universities are concentrated in and around large cities, this is not necessarily reflected in the intensity of AI-related activities. Taken together, this novel and timely evidence informs the debate on the most recent stages of digital transformation of the economy.

Keywords: Artificial Intelligence, Internet websites, Digital transformation.

JEL codes: C81, O3, L2

Résumé

Ce rapport s’appuie sur des données collectées et fournies par GlassAI pour étudier les caractéristiques et activités d’entreprises et universités situées au Canada, en Allemagne, au Royaume-Uni et aux États-Unis qui mentionnent des mots-clés liés à l’Intelligence Artificielle (IA) sur leurs sites internet. L’analyse montre que ces entreprises sont plutôt jeunes et petites, interviennent principalement dans le secteur de l’information et de la communication, placent l'IA au cœur de leur activité, et fournissent des solutions à leurs clients. En outre, le type d'activité liée à l'IA mentionnée par ces entreprises varie selon les secteurs. Par ailleurs, la concentration géographique des universités dans et autour des grandes villes ne se reflète pas nécessairement dans l’intensité des activités liées à l’IA. Ensemble, ces éléments nouveaux éclairent le débat sur les dernières avancées de la transformation numérique de l'économie.

Kurzfassung

In dieser Arbeit werden die von GlassAI gesammelten und zur Verfügung gestellten Informationen verwendet, um die Merkmale und Aktivitäten von Unternehmen und Universitäten in Kanada, Deutschland, dem Vereinigten Königreich und den Vereinigte Staaten zu analysieren, die auf ihren Webseiten Schlüsselwörter im Zusammenhang mit Künstlicher Intelligenz (KI) erwähnen. Die Studie zeigt, dass diese Unternehmen in der Regel jung und klein sind, hauptsächlich im Informations- und Kommunikationssektor tätig sind, KI zum Kern ihres Geschäfts gehört und sie auf die Bereitstellung von Kundenlösungen ausgerichtet sind. Bemerkenswert ist, dass die Art der von ihnen erwähnten KI-bezogenen Aktivitäten je nach Sektor variiert. Obwohl Universitäten in und um große Städte konzentriert sind, spiegelt sich dies nicht unbedingt in der Intensität der KI-bezogenen Aktivitäten wider. Zusammengenommen liefern diese neuen und zeitgemäßen Erkenntnisse Informationen für die Debatte über die jüngsten Phasen der digitalen Transformation der Wirtschaft.

Executive summary

- Artificial Intelligence (AI) has the potential to significantly affect economic growth, and to transform the economic landscape and industries widely. However, still little is known about the patterns of uptake of AI by firms, especially across countries.
- This work provides a comprehensive analysis focusing on organisations with AI-related online presence. For the first time, it analyses the characteristics of companies and universities that mention AI-related keywords on their websites across four countries: Canada, Germany, United Kingdom, and United States. The analysis relies on novel information collected and provided by GlassAI, a private company that reads and interprets open web text at scale.
- Focusing on the characteristics of AI companies identified through web reading, the analysis highlights that those tend to be young and small. The large majority of these AI companies operates in the "Information & Communication" sector.
- These firms tend to have AI at the core of their business, appear often business oriented and aim at providing solutions to their customers. Many of these solutions appear related to *Data Analysis* or the provision of specialised services, such as *Cloud solutions.*
- Beyond the generic *Artificial Intelligence* terms reported on companies' websites, *Machine Learning* emerges as a central topic, often co-occurring with other AI techniques (e.g., *Artificial neural network*, *Natural language processing*, and *Predictive analysis*). Also applications – such as *Robots* or *Computer vision* – are among the terms often appearing on the websites of such companies.
- The types of AI related activities reported by companies vary across sectors. While the leading topics (*Machine learning*, *Artificial neural network*, *Natural language processing*, and *Computer vision*) prevail in the "Information & communication" sector, more significant differences are observed in other sectors.
- Exploring the patenting or trademark activity of companies highlights that a low proportion (less than 10%) of firms that have AI-related online presence also holds these Intellectual Property (IP) rights. Most patenting companies develop technologies related to "Computer technology".
- Zooming in on universities with AI-related online presence highlights the strong concentration of universities in and around large cities. However, this polarisation is not necessarily reflected in the intensity with which AI-related activities take place (proxied by the maximum number of AI-related topics per location).
- This work, which is exploratory in nature, builds upon and complements a range of recent OECD analysis focusing on the diffusion and use of AI based on detailed micro-economic data.

Synthèse

- L'intelligence artificielle (IA) a le potentiel d'agir de manière significative sur la croissance économique, et de transformer largement le paysage économique et les industries. Cependant, les modalités d'adoption de l'IA au sein des entreprises ne sont pas encore bien connues, en particulier suivant les pays.
- Ce travail propose une analyse des organisations affichant des activités liées à l'IA sur leur site internet. Les caractéristiques et activités des entreprises et des universités qui mentionnent des mots-clés liés à l'IA sur leurs sites Web sont analysées pour la première fois dans quatre pays : Canada, Allemagne, Royaume-Uni et États-Unis. L'analyse s'appuie sur les informations collectées et fournies par GlassAI, une société privée qui lit et interprète les textes ouverts du web à grande échelle.
- S'agissant des caractéristiques des entreprises d'IA identifiées par la lecture du Web, l'analyse souligne que celles-ci seraient plutôt jeunes et petites. La grande majorité de ces entreprises d'IA interviennent dans le secteur "Information et communication".
- Ces entreprises ont tendance à placer l'IA au cœur de leur activité, sont souvent tournées vers le commerce et visent à fournir des solutions à leurs clients. Ces solutions semblent pour la plupart liées à *l'analyse de données* ou à la fourniture de services spécialisés, tels que les *solutions de cloud computing*.
- Au-delà des termes génériques d'*intelligence artificielle* indiqués sur le site web des entreprises, l'*apprentissage automatique* apparaît comme un sujet central, souvent associé à d'autres techniques d'IA (par exemple, les *réseaux neuronaux artificiels*, le *traitement du langage naturel* et *l'analyse prédictive*). Les applications - telles que les *robots* ou la *vision par ordinateur* - font également partie des termes qui apparaissent souvent sur les sites web de ces entreprises.
- Les types d'activités liées à l'IA indiquées par les entreprises varient selon les secteurs. Alors que les thèmes principaux (*apprentissage automatique, réseau neuronal artificiel, traitement du langage naturel* et *vision par ordinateur*) dominent le secteur "Information et communication", des différences plus significatives sont observées dans d'autres secteurs.
- L'examen de l'activité en matière de dépôt de brevets ou de marques montre qu'une faible proportion (moins de 10 %) des entreprises affichant des activités liées à l'IA sur leur site internet détient également ces droits de propriété intellectuelle (PI). La plupart de ces entreprises déposant des brevets développent des technologies liées à la "technologie informatique".
- S'agissant des universités signalant des activités liées à l'IA, celles-ci sont massivement dans et autour des grandes villes. Cependant, cette polarisation ne se reflète pas nécessairement dans l'intensité des activités liées à l'IA (évaluée par le nombre maximum de sujets liés à l'IA par localisation).
- Ce travail, de nature exploratoire, s'appuie sur une série d'analyses récentes de l'OCDE axées sur la diffusion et l'utilisation de l'IA et basées sur des données micro-économiques fines et les complète.

Zusammenfassung

- Künstliche Intelligenz (KI) hat das Potenzial, das Wirtschaftswachstum erheblich zu beeinflussen und die Wirtschaftslandschaft sowie die Branchen umfassend zu verändern. Allerdings ist immer noch wenig über die Art und Weise der Einführung von KI sowohl in Unternehmen selbst also auch im Ländervergleich bekannt.
- Die vorliegende Arbeit liefert eine umfassende Analyse, die sich auf Organisationen mit KI-bezogener Online-Präsenz konzentriert. Zum ersten Mal werden die Merkmale von Unternehmen und Universitäten, die KI-bezogene Schlüsselwörter auf ihren Webseiten erwähnen, in vier Ländern analysiert: Kanada, Deutschland, Vereinigtes Königreich und Vereinigte Staaten. Die Analyse stützt sich auf neuartige Informationen, die von GlassAI gesammelt und bereitgestellt wurden, einem privaten Unternehmen, das offene Webtexte in großem Umfang liest und interpretiert.
- Die Analyse konzentriert sich auf die Merkmale der durch das Lesen von Webtexten identifizierten KI-Unternehmen und zeigt, dass diese eher jung und klein sind. Die große Mehrheit dieser KI-Unternehmen ist im Sektor "Information und Kommunikation" tätig.
- Bei diesen Unternehmen steht die KI im Mittelpunkt ihres Geschäfts, sie sind häufig geschäftsorientiert und bieten ihren Kunden Lösungen an. Viele dieser Lösungen scheinen mit der *Datenanalyse* oder der Bereitstellung spezialisierter Dienstleistungen, wie z. B. *Cloud-Lösungen*, zusammenzuhängen.
- Neben den allgemeinen Begriffen der *künstlichen Intelligenz*, die auf Webseiten der Unternehmen genannt werden, erweist sich das *maschinelle Lernen* als zentrales Thema, das häufig zusammen mit anderen KI-Techniken auftritt (z. B. *künstliche neuronale Netze, Verarbeitung natürlicher Sprache* und *prädiktive Analyse*). Auch Anwendungen - wie *Roboter* oder *Computer Vision* - gehören zu den Begriffen, die häufig auf den Webseiten dieser Unternehmen auftauchen.
- Die von den Unternehmen gemeldeten Arten von KI-bezogenen Aktivitäten variieren je nach Sektor. Während die führenden Themen (*maschinelles Lernen, künstliche neuronale Netze, Verarbeitung natürlicher Sprache* und *Bildverarbeitung*) im Sektor "Information und Kommunikation" dominieren, sind in anderen Sektoren größere Unterschiede zu beobachten.
- Die Untersuchung der Frage, inwieweit das Vorhandensein von KI-bezogenen Online-Schlüsselwörtern mit der Patent- oder Markentätigkeit von Unternehmen zusammenhängt, zeigt, dass ein geringer Anteil (weniger als 10 %) der Unternehmen mit KI-bezogener Online-Präsenz auch Rechte an diesem geistigem Eigentum besitzt. Die meisten patentierenden Unternehmen entwickeln Technologien im Zusammenhang mit "Computertechnologie".
- Ein Blick auf die Universitäten mit KI-bezogener Online-Präsenz zeigt die starke Konzentration von Universitäten in großen Städten und deren Umkreis. Diese Polarisierung spiegelt sich jedoch nicht unbedingt in der Intensität wider, mit der KI-bezogene Aktivitäten stattfinden (gemessen an der maximalen Anzahl von KI-bezogenen Themen pro Standort).

- Diese Arbeit, die explorativen Charakter hat, baut auf einer Reihe neuerer OECD-Analysen auf, die sich auf der Grundlage detaillierter mikroökonomischer Daten auf die Verbreitung und Nutzung von KI konzentrieren, und ergänzt diese.

1 Introduction

Artificial intelligence (AI) refers to machine-based systems that are capable of influencing the environment by making recommendations, predictions, or decisions for a given set of objectives (OECD, 2019[1]). In practice, AI consists of machines performing human-like cognitive functions, such as learning, understanding, reasoning and interacting.

AI has the potential to significantly affect economic growth and to transform the economic landscape and industries widely. In fact, it is often referred to as a general-purpose technology, whose applications can potentially bring significant improvements to adopters. However, analyses focusing on the diffusion of AI based on detailed data are still at their early stages, especially across countries.

Identifying which organisations are leveraging AI and studying their characteristics is particularly important from an economic policy perspective. It may shed light on the factors that may enable or prevent this stage of digital transformation and help better understand the implications of AI diffusion for economic outcomes.

In this context, this work provides a comprehensive analysis focusing on organisations with AI-related online presence. For the first time, it analyses the characteristics of companies and universities that mention AI-related keywords on their websites[1] across four countries: Canada, Germany, United Kingdom, and United States. The analysis relies on novel information collected and provided by GlassAI,[2] a private company that reads and interprets open web text at scale.

Focusing on the characteristics of companies provides a number of novel insights. These tend to be young, small, with the large majority operating in the "Information & Communication" sector. They tend to have AI at the core of their business, appear often business oriented and aim at providing solutions to their customers. Beyond generic AI terms, several AI techniques and AI applications often appear on their websites, with some heterogeneity across sectors. Zooming in on universities highlights their strong concentration in and around large cities. However, this polarisation is not necessarily reflected in the intensity with which AI-related activities take place.

This work, which is exploratory in nature, builds upon and complements a range of recent OECD analysis focusing on the diffusion of AI based on detailed micro-economic data (e.g., Calvino and Fontanelli (2022[2]); Calvino et al. (2022[3]); Squicciarini and Nachtigall (2021[4]) and Samek et al (2021[5]); Nakazato and Squicciarini (2021[6]); Baruffaldi et al. (2020[7])), Dernis et al. (2021[8]), as well as to the ongoing OECD-BCG-Insead survey on AI in business.

The rest of the analysis is organised as follows. Section 2 provides a brief overview of the relevant literature. Section 3 describes in detail the data used and the methodology. Section 4 analyses the characteristics of companies with AI-related online presence, while Section 5 further focuses on the AI activities of those companies. Section 6 zooms in on universities with AI-related online presence. Section 7 discusses the main findings, provides some conclusive remarks, and points to some next steps for future analysis.

2 A brief overview of related analyses

This work is related to two main different strands of literature. The first one is an emerging field that analyses the diffusion of AI across firms, the characteristics of AI adopters, and their impact for economic outcomes. This has been reviewed more extensively in other complementary OECD analyses (see for instance Calvino and Fontanelli (2022[2]) and Calvino et al. (2022[3])).Recent contributions to this strand of literature take advantage of information available in firm-level surveys – such as confidential Information and Communications Technology (ICT) surveys –, Intellectual Property (IP) rights (patents and trademarks), or online job postings to identify the characteristics of AI adopters and their role of AI for economic outcomes, such as firm growth or productivity.

More relevantly for the current analysis, information from online job postings has been extensively exploited by recent work to measure firms' demand for AI-related skills. Indeed, in order to develop and adopt AI technologies, firms need specialised human capital. Hence, analysing the demand of jobs featuring AI-related skills may help shed light on the use of AI technologies (Tambe, 2013[9]). Using tools like Machine Learning (ML) algorithms and AI-related keywords, such as the ones found in Baruffaldi et al. (2020[7]), a number of papers including Alekseeva et al. (2020[10]; 2021[11]), Babina et al. (2022[12]), Squicciarini and Nachtigall (2021[4]) and Samek et al. (2021[5]) identify AI-related skills and jobs in online job postings data from Lightcast™ (formerly known as Burning Glass Technologies).

Information for online job postings is not however able to fully capture AI activity by firms (see also Calvino et al. (2022[3]) for further discussion). Although information is comprehensive, it relates by definition only to changes in the demand for AI-related jobs. As firms may be active in AI – or even have AI at the core of their business – without hiring new AI specialists, internet websites are a complementary source of information that is able to capture AI activity from a different perspective.

In line with the current analysis, the second stream of literature is more broadly related to the use of information from internet websites in innovation studies. Indeed, a relatively new body of research leverages the use of data mining algorithms to derive information about firm activities from the vast amount of unstructured web data ("web mining"). Relevantly, data available online – e.g., on firms' websites – may contain detailed information about products, services, and inter-firm linkages which are usually not recorded in standard sources. Moreover, web-mining techniques are effective in timely obtaining granular firm-level data with wide coverage.

A relevant example of web mining in the field of innovation studies is provided by Kinne and Axenbeck (2020[13]). Using a free-to-use web scraping tool, the authors investigate and assess firm websites as a data source for web-based innovation indicators. In particular, they conduct a pilot study on more than 2.4 million firms of the innovation ecosystem in Germany.[3] Closely related to the current analysis, although at a significantly smaller scale, the authors also conduct an exploratory analysis of the AI ecosystem. Using a simple keyword-based approach, they identify Berlin-based companies and scientific institutions mentioning AI on their websites, building indicators of "AI engagement" (both firms directly engaged and firms linked to an AI-engaged partner). Relevantly, for validation purposes, they compare the web-based indicator with the 2019 Community Innovation Survey (CIS) for firms from manufacturing and business-oriented services with at least 5 employees. They find higher shares of adoption for larger firms (both directly or indirectly engaged), which is in line with findings from the CIS, and younger ones.

Relatedly, Daas and van der Doef (2021[14]) develop a model based on web-scraping of firms' websites to detect innovative companies. In particular, they use the web addresses included in the 2016 CIS of the Netherlands (covering the period 2014-16 and firms with more than 10 employees) to test various classification algorithms. Their model is able to reproduce the result from the CIS in terms of firms' innovativeness and to detect innovative companies with less than 10 employees, such as start-ups.

In a similar vein, Krüger et al. (2020[15]) analyse firm-to-firm relationships among 500 thousand German enterprises, using data obtained through web scraping of their websites. The authors create a "digital layer" relating the textual content and the hyperlinks of firms' websites, which results in a geo-located network with over seven million hyperlink relations. On this basis, they focus on proximity among firms and their relationship with innovation. In particular, the authors show that innovative firms have more (hyperlinked) partners and that these are on average more innovative than the partners of non-innovative firms.

With a different focus, Nathan and Rosso (2022[16]) use administrative microdata from the United Kingdom, as well as media and website content about small and medium-sized enterprises (SMEs) to derive measures of firm innovation. They exploit a dataset developed by the data science firm Growth Intelligence (GI) – which uses ML routines on company websites and media content to model firms' lifecycle 'events' – focusing on new products/services reported in the news. They find that past patent activities are related to a firm's current launch activities and that technologically-intensive SMEs are substantially more launch-active than non-tech SMEs. In general, they argue that web-based indicators are a useful complementary measure to existing metrics.

In this context, this study significantly contributes to the existing literature. For the first time, it provides a comprehensive analysis identifying and characterising actors that mention AI-related keywords on their website – a unique source of information that has not yet been extensively exploited so far – across more than one country. This builds upon and extends the work by Calvino et al. (2022[3]), that has characterised AI adopters combining information from online job postings, IP rights, and internet websites, focusing on the United Kingdom only. The next section focuses on describing the data used and the methodology in detail.

3 Data and methodology

This section focuses on describing the data sources and methodology used to identify AI actors using data from online websites. It first presents the data source used – that are based on internet searches carried out by GlassAI – and then provides some descriptive statistics, which highlight the coverage and quality of the data.

Overview of GlassAI data and methodology

GlassAI is a company that reads and interprets open web text at scale, exploiting any machine-readable information contained on companies' or other institutions' websites: sentences, paragraphs, and names. It performs searches that can target specific topics and themes (see Box 3.1 for further details on their methodology).

For the present work, GlassAI was asked to search the web in order to identify any entity with AI-related online presence, i.e. companies as well as other types of institutions (e.g. universities, research centres) that mention AI-related keywords on their websites. To this end, GlassAI was provided with AI-related keywords identified in Baruffaldi et al. (2020[7]), Nakazato and Squicciarini (2021[6]) and Squicciarini and Nachtigall (2021[4]).These keywords identify, for instance, a variety of AI-related methods or applications, such as *artificial neural network*, *speech recognition*, *natural language understanding*, *image/pattern/feature extraction and recognition,* etc. The list of keywords was provided in English and translated into French and German to improve the search on Canadian or German websites.

Searching companies' and other institutions' websites at scale in 2020, GlassAI provided a collection of on-line information related to companies and universities that have an AI-related online presence, likely producing or using AI in their business processes, or having AI-related teaching or research activities, respectively.

Given the very nature of the data source, the GlassAI sample might still include some companies or universities that mention AI-related keyworks on their website without actually developing or using AI. In fact, AI related terms may in some cases be added as ideas or concepts, or mentioned on websites for other reasons (e.g., marketing purposes), while not being actively implemented within a company or university. However, considering that for the large majority of companies and universities more than one AI-related keyword is present on their website, the proportion of such companies or universities is likely to be small.

GlassAI data used in this study include companies and universities in a selection of countries: Canada (1 013 companies and 75 universities), Germany (2 293 companies and 144 universities), the United Kingdom (2 256 companies and 170 universities) and in the United States (8 221 companies and 1 752 universities).

Box 3.1. GlassAI methodology: using AI to read and interpret open web

GlassAI developed AI-research capabilities that deep reads millions of business websites, in a continuous mode, to understand companies and sectors. They rely on machine language "understanding", which aims to give machines the power to understand entire sentences and paragraphs. It relies on various machine learning (ML) and computational linguistics approaches to create an automatic semantic layer for the web. The key features are:

1. Semantic Analysis (Entity Recognition): Natural language processing (NLP) models are implemented to detected entities from text on the open web (e.g. people, addresses, company descriptions, etc.) with a high threshold of precision. Business descriptions, in particular, are identified with language models that consider multiple features such as location on the web page, use of specific keywords and phrases, sentence structure, etc.
2. Resource Crawling: the extraction process was industrialised by developing an intelligent crawling service which discards irrelevant data as above before returning it to storage.
3. Topics ontology: GlassAI derived a 'topic map' for the data, allowing users to query the data by subject classification. For example, a "virtual reality" search will suggest semantically-related simple topics such as "augmented reality" and complex topics such as "Oculus Rift" - formed from a Company and a Product.

Source: Glass AI, https://www.glass.ai/, 2020

GlassAI data: variables and coverage

GlassAI data provision includes some basic information about companies, such as their name, location, and sector, the description of their activities, registered company numbers – when available - as well as records of members of the board (see Table 3.1). In the case of universities, the data extracted were more limited, with fewer variables available (e.g., name, location).

Table 3.1. GlassAI data

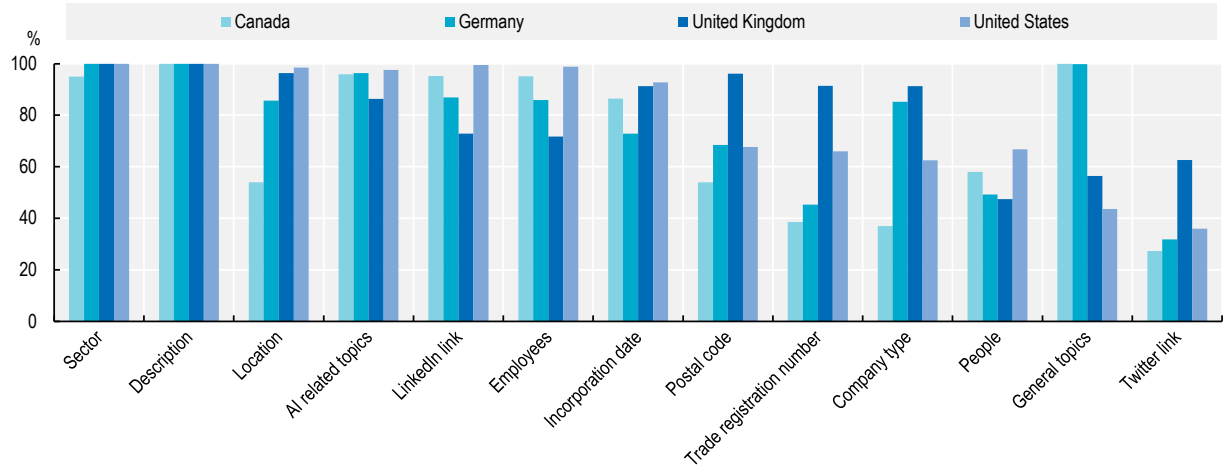
Overview of different source files provided by GlassAI
Orgs: Organisation's main metadata and selected description for currently active AI organisations, including: website details, location, company's description, sector group, list of AI related topics, registered company number; company type; incorporation date; LinkedIn & Twitter links; number of employees…
Sectors: Sector group and sub-sector names predicted for an organisation based on the content read. An organisation may have up to 3 sectors: no predicted sector in case of insufficient content
Org_topics: up to 30 key topics related to each organisation, ranked by order of importance
Specialism_topics: up to 30 topics related to each organisation that are related specifically to statements about skillset, products or services
People: up to 3 key people identified on the company's website or social media (typically leading roles within the organisation)
Partners: mentions of partnerships and collaborations extracted from the company websites
Universities: universities that carry out AI-related activities, based on information on their websites
Ecosystems: sample of evidence from news, announcements and similar content that further describes how AI is being applied across different organisations

Source: GlassAI, 2020.

The extent to which GlassAI was able to retrieve information through reading the web is presented in Figure 3.1. While information related to companies' description, location and sector of activity is well covered, other pieces of information, such as the names of key people or companies' partners, are sometimes lacking, which makes these data not fully exploitable. When screening websites' content, GlassAI was able to allocate a selection of AI-related topics to companies and universities, using an ontology applied on the website's content. AI-related topics may therefore be exploited to better understand the specialisation of a given company or university in specific fields of AI, as well as to identify interconnections across different AI domains. The analysis focuses here on the data related to organisations, sectors, topics and universities. Also considering their more limited coverage, other sets of information provided by GlassAI were not considered in the current analysis.

Figure 3.1. GlassAI fields coverage, company level data, 2020

Share of companies featuring relevant pieces of information, by main variable and country



Source: OECD calculations based on GlassAI data, November 2022.

Several firm-level characteristics, such as the economic sector in which companies operate, the date of firms' incorporation as well as an estimate of the number of employees, were assessed from information featured on the company's webpages. GlassAI allocated the sectors in which companies operate to 19 groups of the 4th revision of the International Standard Industrial Classification of All Economic Activities (ISIC, rev. 4). Such characteristics were in turn compared with other firm-level data available (namely the ORBIS© database licensed by Bureau van Dijk), highlighting a relatively good degree of consistency (see Box 3.2).

Box 3.2. Further analysing information on firm characteristics

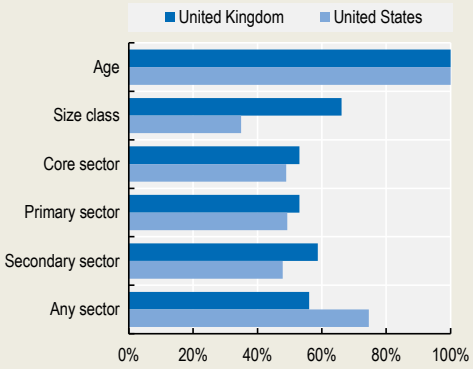
The records provided by GlassAI were compared to commercial data from ORBIS© containing firm-level information (derived from their balance sheets). This was done for companies located in the United Kingdom and in the United States. For the United Kingdom, the two sets of data (GlassAI and ORBIS©) were linked using official registration numbers available in the two sources. For the United States, as the registration number data could not be exploited, the addresses of the companies' webpages were compared to the list of URLs provided in ORBIS© records. For both countries, complementary linkages were established using a series of string-matching algorithms (see the procedure described in Annex A) and were finalised with manual screening to map unallocated records.

Overall, 93% of the GlassAI sample for the United Kingdom could be mapped to company level data in ORBIS©: 2 087 out of 2 256 companies. For the United States, correspondence was found for only 49% of the sample (4 033 out of 8 221 companies). The relatively low matching rate for the United States is explained by the coverage of ORBIS© (see Bajgar et al. (2020[17]) for a thorough discussion on the cross-country coverage of ORBIS©). Selected fields, namely the date of incorporation, the firm size (based on the number of employees), and the industry classes of the firms, were then compared across the two sources. The highest level of consistency is observed for the dates of incorporation, which are unequivocally identified by both sources. The comparison between the firm size bands identified by GlassAI through web reading and that provided in ORBIS© indicate broader variations: 66% of companies in the United Kingdom and only 35% in the United States are allocated to the same size class in both sources. These differences may be driven by the different levels of consolidations available in US company accounts in ORBIS© and possible gaps in the years for which data are available.

Differences are also observed for sectors into which companies are assigned. GlassAI predicted the sector of companies and linked those to the high-level categories of ISIC rev.4. Hence, a company may be allocated to one or many sectors in GlassAI data, and conversely different sectors are provided in ORBIS© (core, primary and secondaries). GlassAI sectors are similar to at least one of the classification codes provided in ORBIS© for 56% of companies located in the United Kingdom, and for 75% of companies in the United States.

Figure 3.2. Comparing sources

Share of companies featuring the same information across sources



Note: Data refer to the number of companies featuring the same information in GlassAI and in ORBIS© over the total number of companies in the matched GlassAI-ORBIS© sample for which the observed variable was available. The economic sector provided by GlassAI was compared to the core sector as well as to primary and secondary sectors listed in ORBIS©.
Source: OECD calculations based on GlassAI data and ORBIS© database, 2019_1, November 2022.

4 Characterising companies with AI-related online presence

This section provides a first descriptive analysis of the characteristics of companies that have an AI-related online presence, based on the data provided by GlassAI described in the previous section. In particular, the section provides insights on the age, size, sector, and activities of such companies across four countries: Canada, Germany, United Kingdom, and United States.

Focusing on those provides novel insights, for the first time across more than one country, about a group of AI actors that highlight engagement in AI on their website. These insights are complementary to those emerging using different methodologies (e.g., based on information from patents, ICT surveys, online job postings). In particular, information collected through web reading provides unique insights about the activity carried out by AI companies beyond sectoral classification. This first characterisation is propaedeutic to digging further into the AI activities of these AI actors, which will be further explored in the next section.

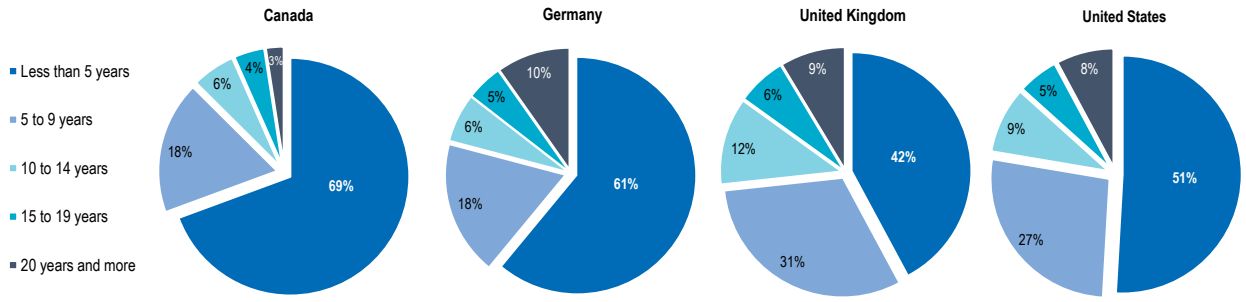
Age and size of companies with AI-related online presence

First, the age and size class of companies with AI online presence are explored. Available data for Canada, Germany, the United Kingdom and the United States indicate that the companies identified are relatively young: more than three quarters of companies were founded after 2010 (see the top panel of Figure 4.1). Indeed, for three out of the four countries analysed, more than half of the companies in the sample have less than 5 years of age. Focusing on the German sample, about 10% of firms active in AI were founded more than 20 years ago. Similar proportions are observed in the United Kingdom (9%) and the United States (8%), while older firms represent only 3% in the Canadian sample.

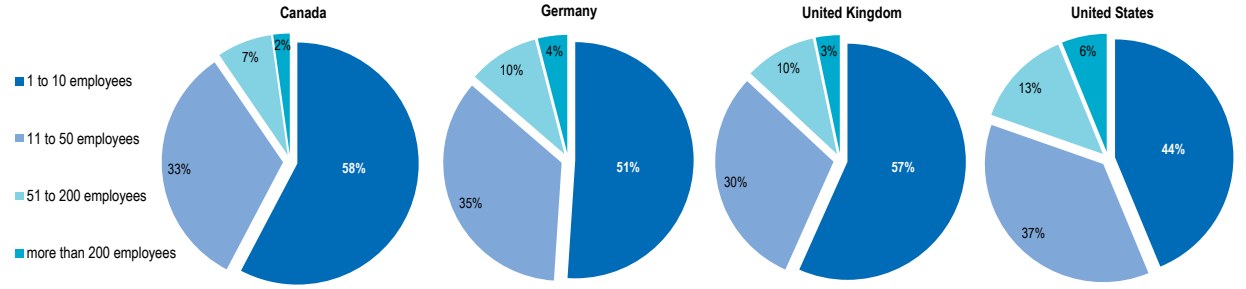
As previously discussed, the sample of firms provided by GlassAI also includes information on the company's size class, based on the number of employees that was retrieved during the website screening. In the four countries considered, more than 80% of the sample refers to micro or small firms, having 50 employees or less (see the bottom panel of Figure 4.1). In 2020, about half of the firms flagging AI activity on their website are identified as micro-firms, with up to 10 employees: this share is the highest in Canada (58%) and lowest in the United States (44%). In turn, about one third of the companies with AI online engagement have between 11 and 50 employees in the four countries.

Figure 4.1. The age and size of companies with AI-related online presence, 2020

Age classes



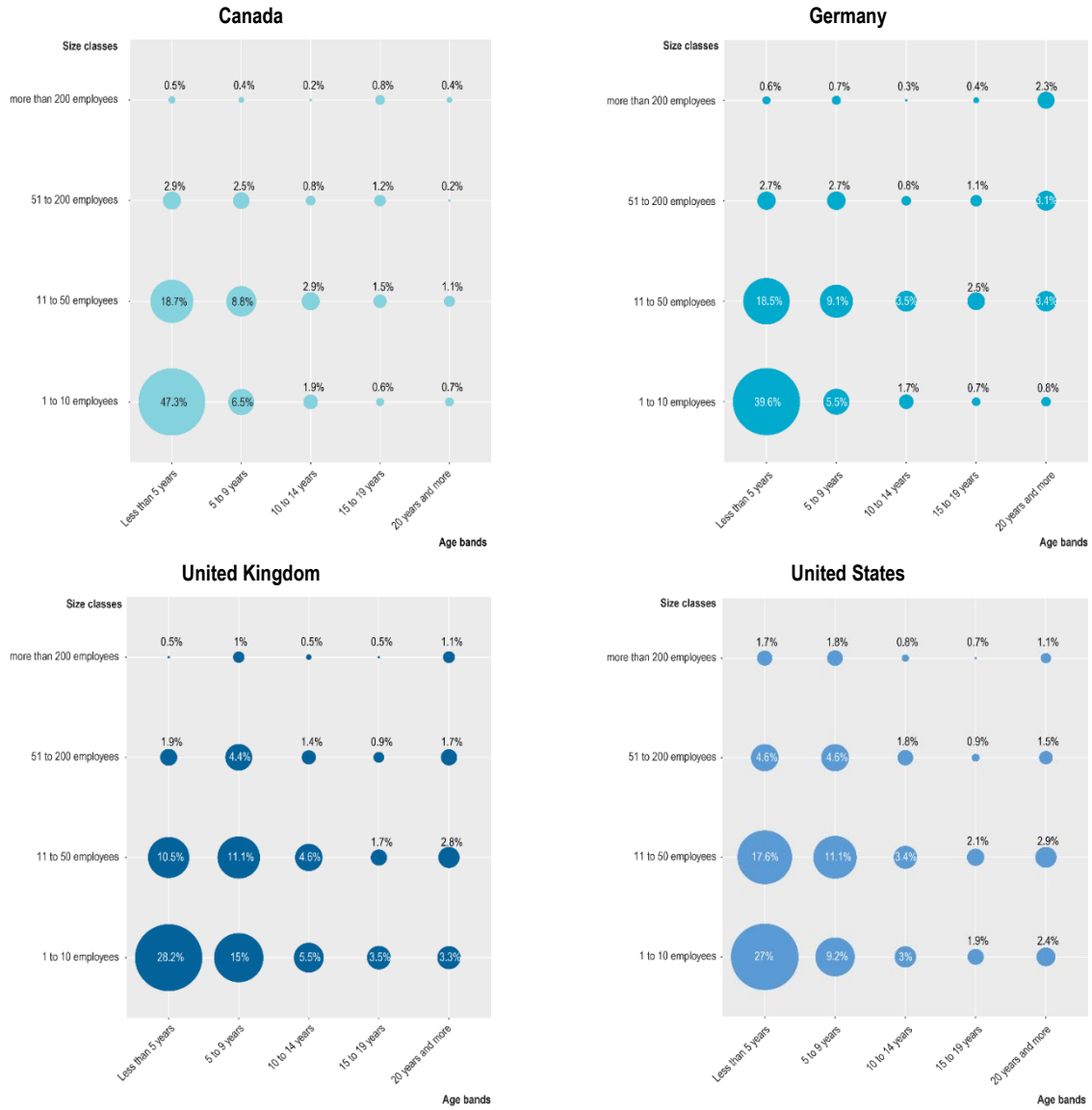
Size classes



Note: The age of companies is calculated as the difference between the year 2020 and the date of incorporation identified by GlassAI. Employees' size classes are provided by GlassAI, when available. The firm age distribution observed for Germany and the firm size distribution for the United Kingdom rely on partial data on age and the number of employees, respectively.
Source: OECD calculations based on GlassAI data, November 2022.

When analysing jointly age and size characteristics, the data highlight that these AI firms are typically young and small. In Canada, nearly half of the companies are identified as micro start-ups, founded after 2015 and featuring 10 or less employees (see Figure 4.2). This proportion amounts to 40% for Germany, 28% for the United Kingdom and 27% for the United States. On the opposite side, older and larger firms represent between 0.4% of the sample of AI firms (in Canada) and 2.3% (in Germany).

Figure 4.2. Companies with AI-related online presence, by size and age classes, 2020



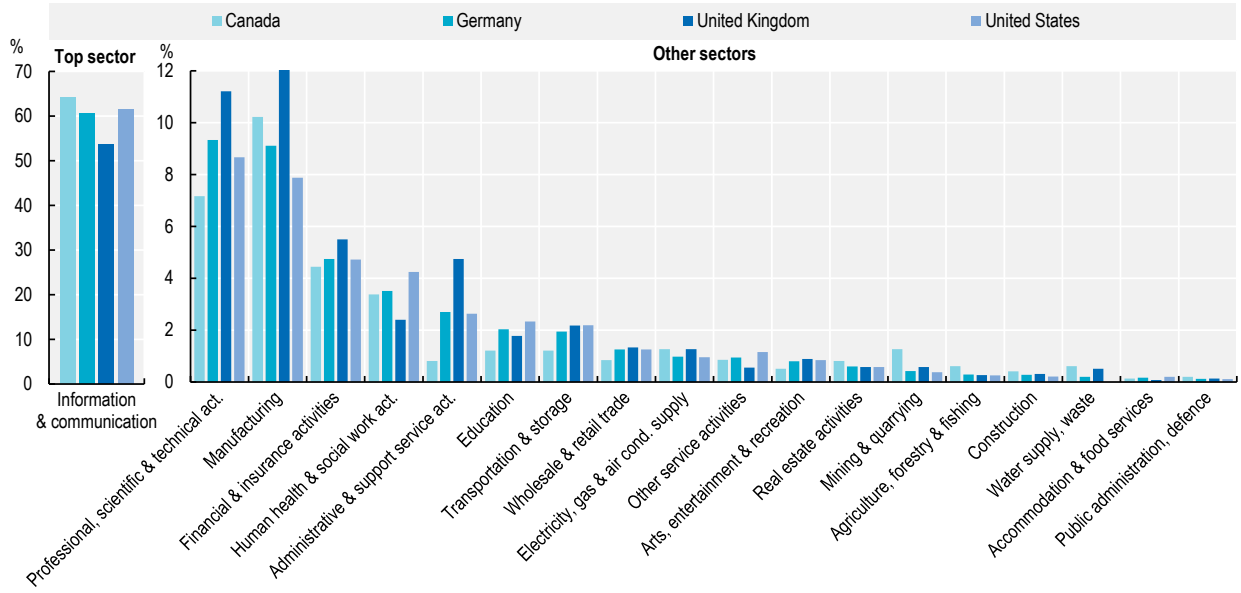
Note: The age of companies is calculated as the difference between the year 2020 and the date of incorporation identified by GlassAI. Employees' size classes are provided by GlassAI, when available. The distribution observed for Germany and the United Kingdom rely on partial data on age and the number of employees.
Source: OECD calculations based on GlassAI data, November 2022.

Sectors of activity of companies with AI-related online presence

Beyond age and size, another key characteristic about companies that have an AI-related online presence is their sector of economic activity. In this respect, the sample of AI firms under scrutiny was allocated to sectors of activity by GlassAI. This was done on the basis of the sector of activities mentioned on the website, when available, or by applying an ontology on textual information provided on websites that proxied the industries in which companies operate. Economic sectors identified were in turn converted into 19 ISIC rev.4 industries to ease visualisation and facilitate comparisons with other data sources.

Figure 4.3. Companies with AI-related online presence by sector, 2020

Distribution of companies by main economic sector, ISIC, rev.4



Source: OECD calculations based on GlassAI data, November 2022.

In all four countries, a majority of AI companies operates in the "Information & Communication" sector (Figure 4.3). This sector alone represents between 53% of AI firms (in the United Kingdom) and 64% (in Canada). In turn, about 9% of AI firms, on average, operate in "Professional, scientific and technical activities", which comprises Scientific R&D as well as Advertising and market research activities, ranging from 7% in Canada to 11% in the United Kingdom. A similar proportion of AI firms operates in the "Manufacturing" sector, where the shares amount to nearly 8% for the United States and up to 12% for the United Kingdom.

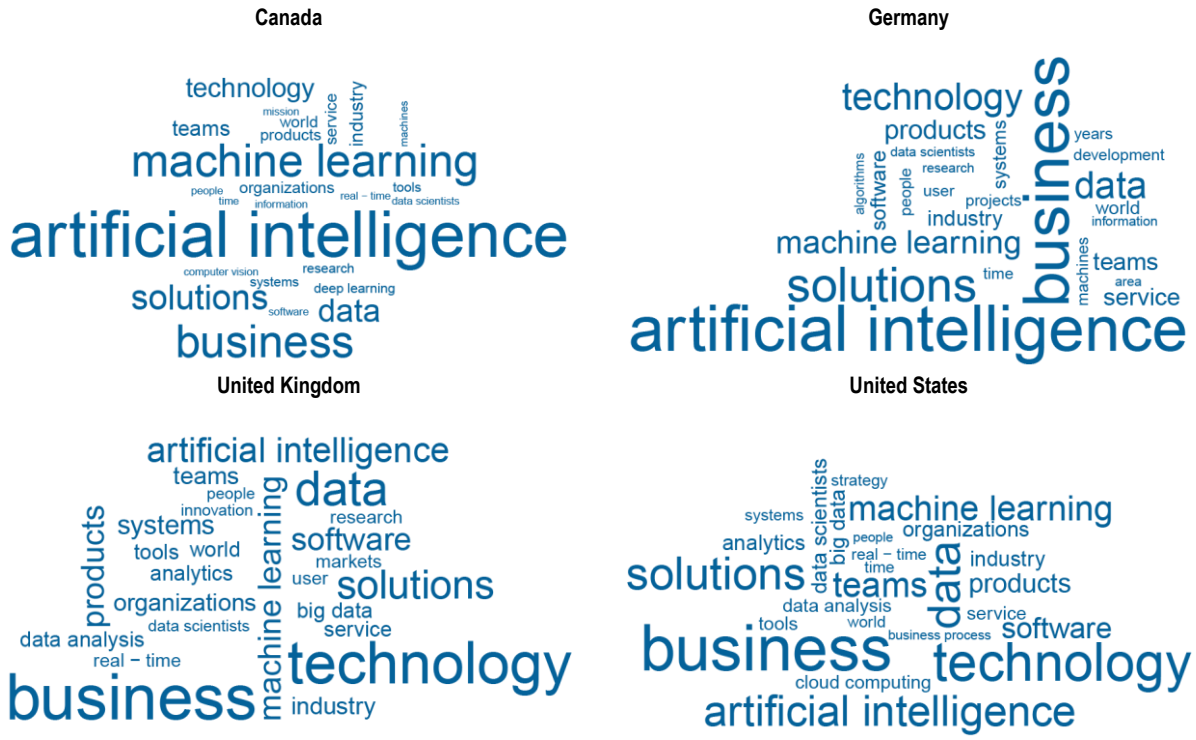
Beyond sectors: additional insights on the main activities of companies

Through reading the companies' webpages, GlassAI captured the descriptions of firms' main activities, and classified those into a non-exhaustive list of general topics. The frequency of general topics identified within each of the four companies' samples provides additional insights on the type of business activities those companies conduct, beyond those provided by the sectoral classification presented above.

The top 25 general topics are listed in the form of word clouds in Figure 4.4. In all countries considered, companies with an AI-related online presence tend to tag AI in their business description, suggesting they tend to have AI at the core of their activity. In fact, topics like *Artificial Intelligence* and *Machine Learning* appear highly important, especially in Canada and Germany.

Companies with AI-related online presence relevantly appear business oriented in all four countries, with a core element related to providing solutions to their customers. Their activities are indeed often oriented towards technology or offering data treatment.

Figure 4.4. Top 25 general topics mentioned on companies' webpages, 2020



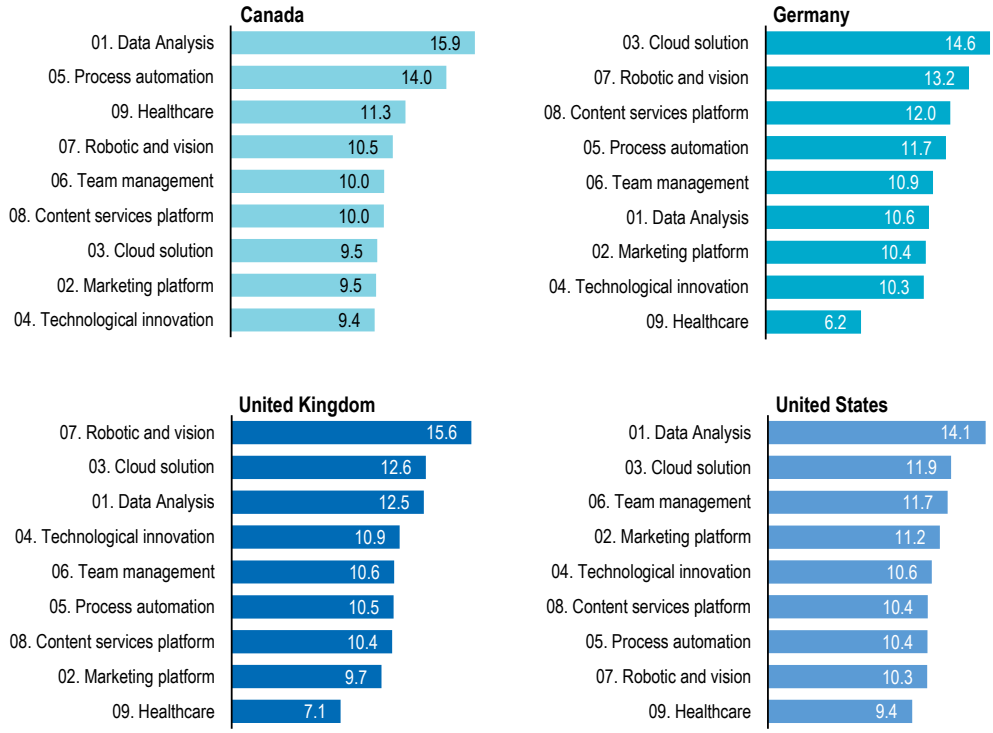
Note: Data relate to the frequency of general topics allocated to companies by GlassAI, amongst AI-related companies identified in each country. The larger the font size, the higher the number of companies featuring that topic.
Source: OECD calculations based on GlassAI data, November 2022.

A text mining approach – topic modelling – was performed on the description of the companies' activities, to complement the picture presented above. This technique enables to uncover different clusters of firms' activities, building on the use of a common terminology. Nine clusters were identified, using the methodology described in Box 4.1 and Annex B, into which companies with AI-related online presence could be classified. Clusters were labelled on the basis of the key terms that tend to appear the most frequently: Data analysis, Marketing platform, Cloud solution, Technological innovation, Process automation, Team management, Robotic and vision, Content services platform, Healthcare.

The categorisation of companies into clusters informs on the extent to which AI technology is applied in broad areas to solve various business problems (Figure 4.5): from the general technological side (e.g. Data analysis, Technological innovation, Process automation), to the provision of specialised services (e.g., Cloud solutions, Content services platform), or to the marketing and human resources side (e.g. Team management, Marketing platform). The two clusters Robotics and vision as well as Healthcare are areas in which AI is now commonly applied. The distribution of companies by clusters tends to vary across countries: Canada and the United States have the largest proportion of companies for which AI is related to Data analysis. In Germany, almost 15% of companies indicate AI activities related to Cloud solutions, followed by Robotics and vision (13%). In the United Kingdom, more than 15% of companies tend to connect AI-related activities with Robotics and vision. Data analysis and Cloud solutions appear among the top three clusters in three out of four countries.

Figure 4.5. Activities of companies with AI-related online presence, 2020

Share of companies in different clusters, by country



Note: Companies are categorised to clusters on the basis of the top 2 topics to which companies have the highest probabilities to belong, using fractional counts based on the probabilities.
Source: OECD calculations based on GlassAI data, November 2022.

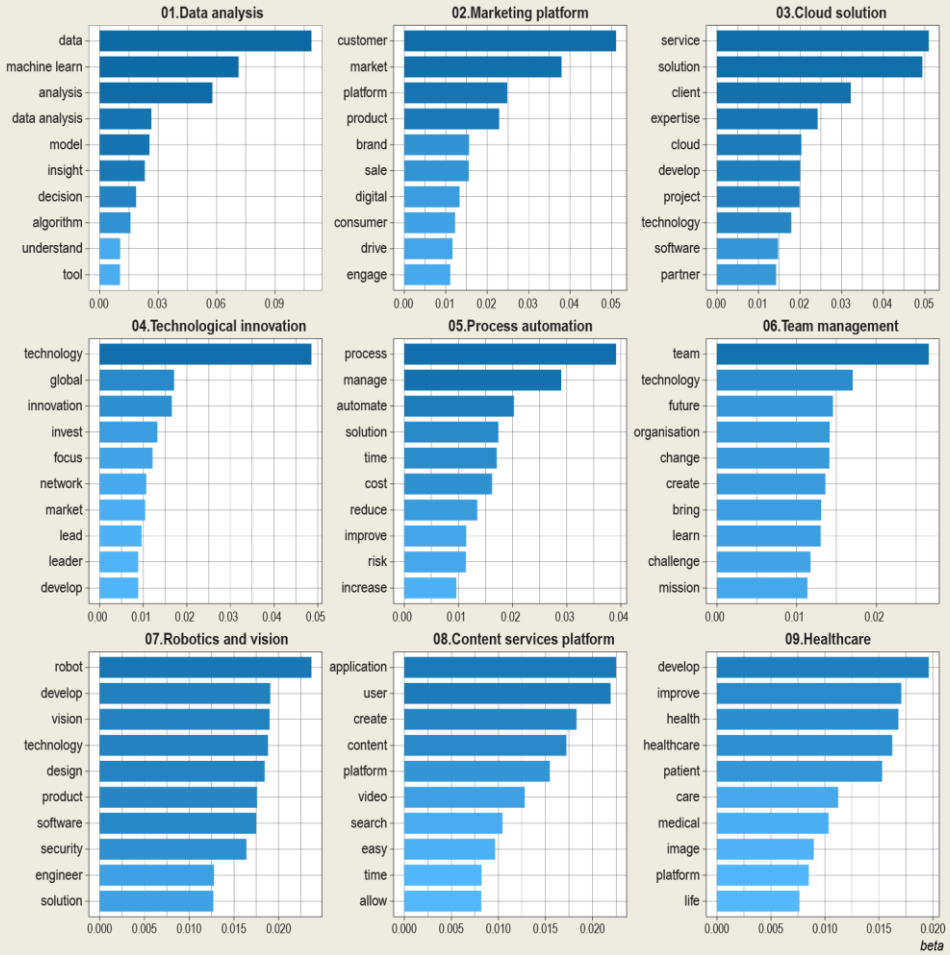
Taken together, the evidence presented so far highlights that the AI companies identified through web reading tend to be small start-ups mainly active in ICT services sectors. These firms tend to have AI at the core of their business, are generally business oriented and aim at providing solutions to their customers. Many of these solutions are related to data analysis or the provision of specialised services. The next section will provide further insights about the AI activity of these companies, taking advantage of the rich information available on websites.

Box 4.1. Characterising companies with ai-related online presence using topic modelling

A topic modelling exercise was performed on the companies' description, for the four countries altogether, using the Latent Dirichlet Allocation (LDA) algorithm and Gibbs sampling methods. "Topics" are identified on the basis of the frequency and combination of words that are included in the business description of the firms, as provided on their websites, and the probability of firms to belong to one or more of the topics. This methodology enables to classify different terms into clusters, and ultimately, to help categorise companies according to the different clusters. The optimal number of topics was identified as described in Annex B.

Different clusters (topics) were labelled on the basis of the terms that are featured together in a given topic. Only the top 10 terms for each cluster are shown in the figure, with beta values representing the probabilities of terms to belong to the clusters. The different shades of blue illustrate the intensity of the probabilities.

Figure 4.6. Top 10 terms per cluster, 2020



Source: OECD calculations based on GlassAI data, November 2022.

5 Analysing the AI activities of companies with AI-related online presence

Information from online websites provides a unique source to analyse the activities of companies that have an AI-related online presence. Initial insights about the company main activity have been provided in the previous section, highlighting that these tend to have AI at the core of their business, are generally business oriented and aim at providing solutions to their customers, e.g., related to data analysis or specialised services.

This section explores further the AI activities of companies with AI-related online presence, taking advantage of the rich set of information available on company websites. The section first focuses on the different AI (rather than general) topics listed on companies' websites, assessing their frequency and co-occurrences. It then analyses the sectoral heterogeneity of those topics. Finally, it provides a glimpse of the extent to which different AI topics are related to IP activity in those companies.

AI topics listed on companies' websites

A key information that is unique of the data source analysed – and is central to the identification of the AI actors under scrutiny – is the list of AI-related topics that emerged from reading their websites. The list was provided by GlassAI and further grouped into broader AI terms or techniques relying on similar concepts (additional details about the grouping are available in Annex C).

Figure 5.1 displays the top terms related to AI identified on companies' websites: the word clouds refer to the top 20 terms based on their frequency; the top 10 words being listed on the right-hand side figures. In the sample of firms under scrutiny, the (more generic) topic *Artificial Intelligence* was flagged in most companies: 51% of firms in the United Kingdom, 65% of firms in the United States, and 70% of companies in Canada and Germany. Owing to the prevalence of this term, it was removed from the word cloud images to better focus on the other AI-related techniques reported on the companies' webpages.

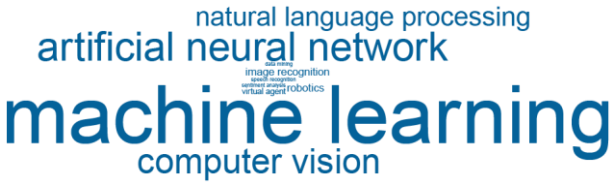
Specific AI-related techniques are among the top terminology mentioned on companies' websites (Figure 5.1): *Machine learning* is the most referred technique that is signalled by 27% of companies (Germany) ranging up to 43% (Canada). *Machine learning* here encompasses a variety of techniques, algorithms (supervised / unsupervised), and models, as presented in Annex C. Other rather "core" techniques for AI, namely *Artificial neural network* and *Natural language processing*, are referred to by 3% to 10% of firms, depending on the country.

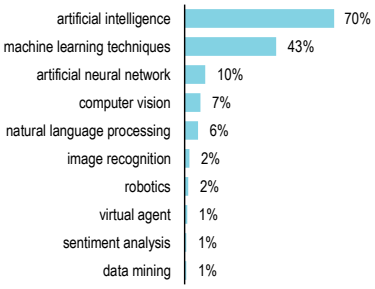
Robotics is also frequently associated with AI on the sample companies' webpages: the term is mentioned by 15% of AI companies located in the United Kingdom, 8% in the United States, 7% in Germany and only 2% in Canada. A related theme, *Computer vision*, is reported by 6% to 7% of AI companies in all four countries.

Figure 5.1. Top AI-related topics listed on companies' webpages, by country, 2020

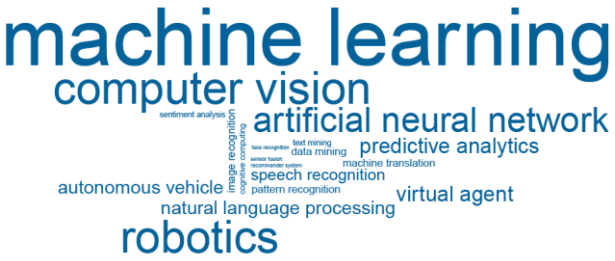
As a share of total AI active firms

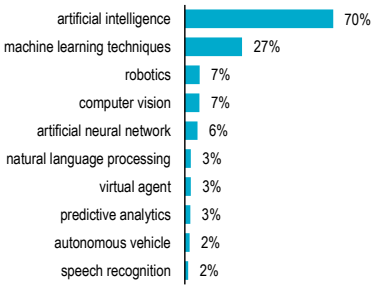
Canada



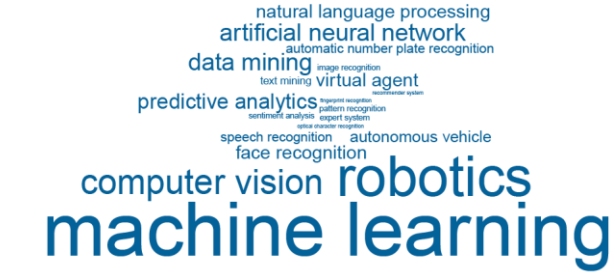


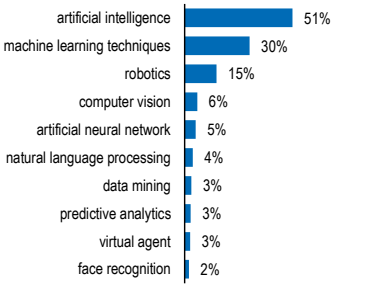
Germany



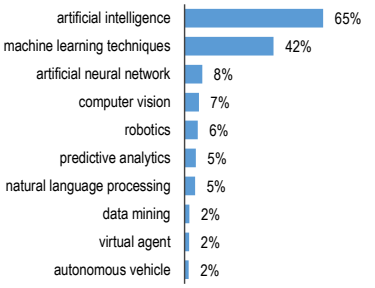


United Kingdom





United States

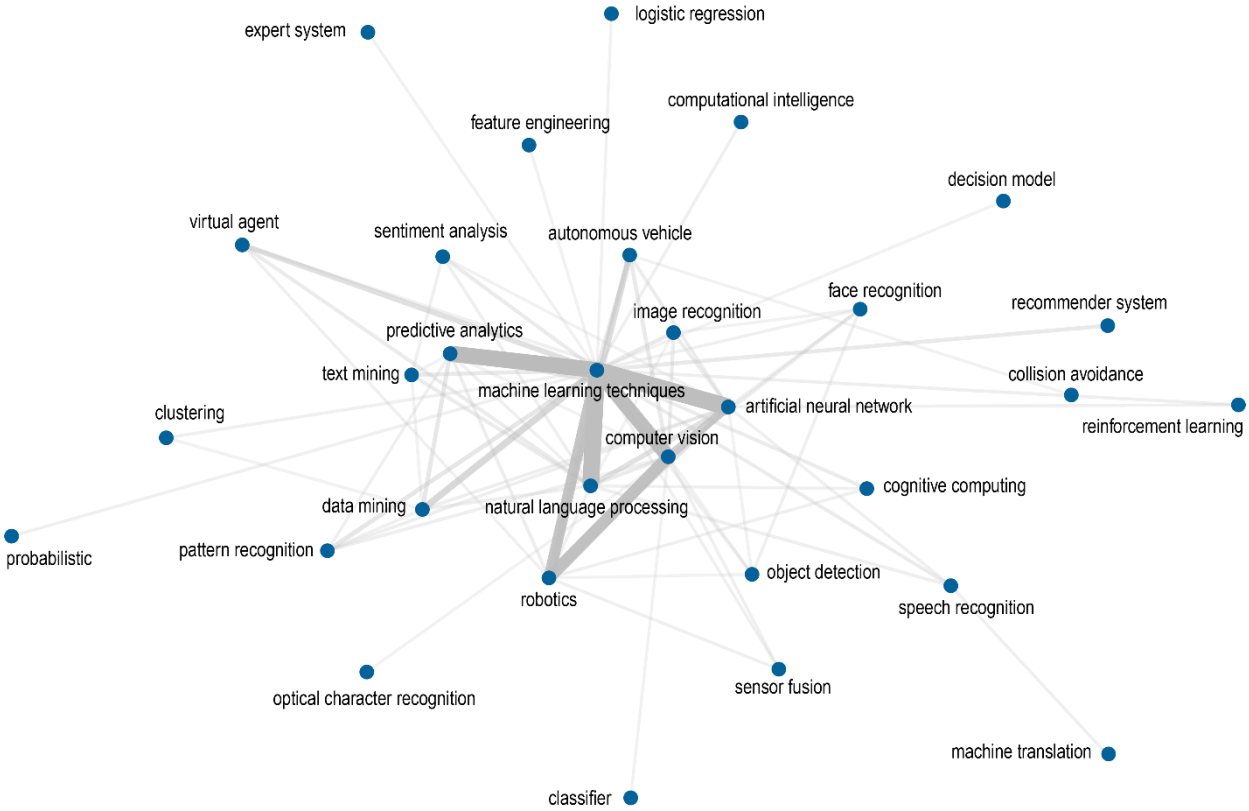


Note: Data relate to the number of companies in the GlassAI sample by specific AI-related topic. The generic topic "Artificial Intelligence" was excluded from the word clouds. The font size of some predominant terminologies (e.g., *machine learning, robotics, artificial neural network*) was intentionally reduced to facilitate the reading of lower frequency terms.
Source: OECD calculations based on GlassAI data, November 2022.

Other AI-topics featuring relatively high in terms of frequency include: *Predictive analytics*, which appears in the top 10 terms for Germany (3%), the United Kingdom (3%) and the United States (5%); *Chatbot*, which is mentioned by 1% to 3% of companies in the sample; and *Autonomous vehicle*s, which is reported by about 2% of AI companies in Germany and in the United States.

Several AI topics can be identified simultaneously on companies' webpages. Analysing how different terms are combined may shed further light about how AI is used in firms, and where interdependencies between techniques are found. *Machine learning techniques* are a central element for AI-active firms located in the four countries, which unsurprisingly appears often with other techniques, such as *Artificial neural network*, *Natural language processing*, and *Predictive analysis* (Figure 5.2). *Computer vision* and *Robotics* are also frequently associated and tend to be cited together and with machine learning techniques. Firms highlighting activities related to *Computer vision* tend also to deal with *Autonomous vehicles*, and *Recognition* of different types (*image*, *face*, *pattern*, *object, optical characters)*. In turn, companies focusing on *Robotics* also engage in *Autonomous vehicles*, *Predictive analysis*, *Cognitive computing*, *Sensor fusion*, *Virtual agent*, to list only a few.

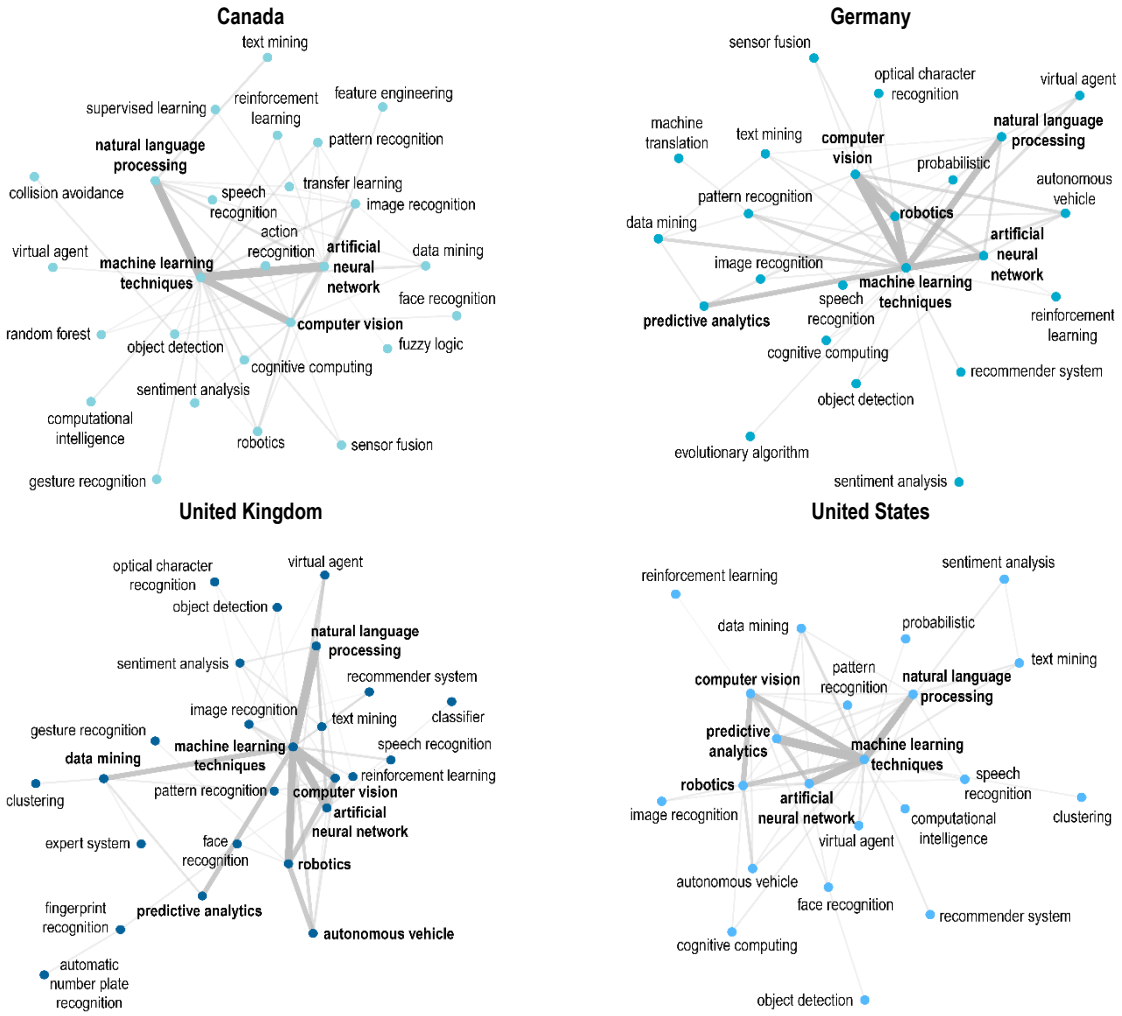
Figure 5.2. Co-occurrence of AI topics listed on companies' webpages, 2020



Note: Data relate to the top co-occurrences of AI-related topics in companies active in AI, using fractional counts. The generic topic *Artificial Intelligence* was excluded from the calculation.
Source: OECD calculations based on GlassAI data, November 2022.

The extent to which AI topics are combined slightly differs across countries. Interestingly, this may be related to their relative specialisation in certain AI techniques, as revealed by the information available on companies' websites. In Canada, the sample of firms display activities mainly oriented towards AI "core" techniques - *Machine learning*, *Artificial neural network*, *Natural language processing* - and to *Computer vision* (see Figure 5.3). For German firms, the top connections are observed between *Computer vision* and *Robotics*, which also strongly link to *Machine learning*, *Artificial neural networks*, *Natural language processing* or *Predictive analysis*. In the United Kingdom and in the United States, the activities of AI firms seem to a certain extent more diversified.

Figure 5.3. Co-occurrence of AI topics, by country, 2020



Note: Data relate to the top 100 co-occurrences of AI-related topics in companies active in AI, by country, using fractional counts. The generic topic *Artificial Intelligence* was excluded from the calculation.
Source: OECD calculations based on GlassAI data, November 2022.

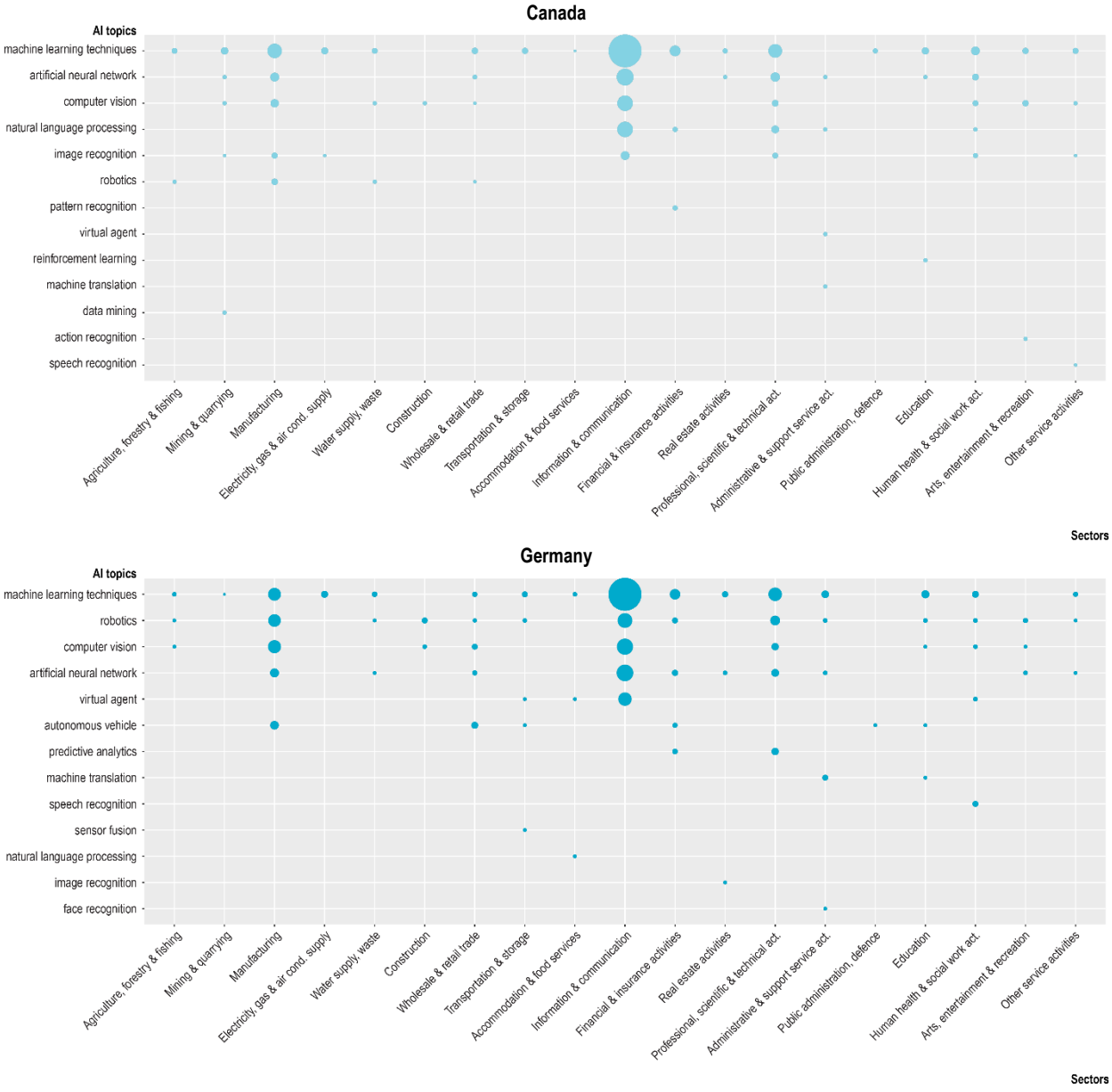
AI activities by economic sector

The types of AI-related activities reported by companies vary across sectors: Figure 5.4 and Figure 5.5 display the top 5 topics identified in different sectors, with the size of the bubble reflecting the proportion of companies in the sample. While the leading topics (*Machine learning*, *Artificial neural network*, *Natural language processing*, and *Computer vision*) prevail in the "Information & communication" sector, differences are observed in other sectors.

Robotics is one of the top 5 subject areas signalled by companies operating in the "Manufacturing" sector, as well as those in "Wholesale & retail trade" sectors across countries. In Germany, the United Kingdom and the United States, companies from these two sectors also contribute to AI developments related to *Autonomous vehicles*. The dispersion of AI topics across sectors is stronger in the United Kingdom and in the United States.

Figure 5.4. Top AI topics, by sector, Canada and Germany, 2020

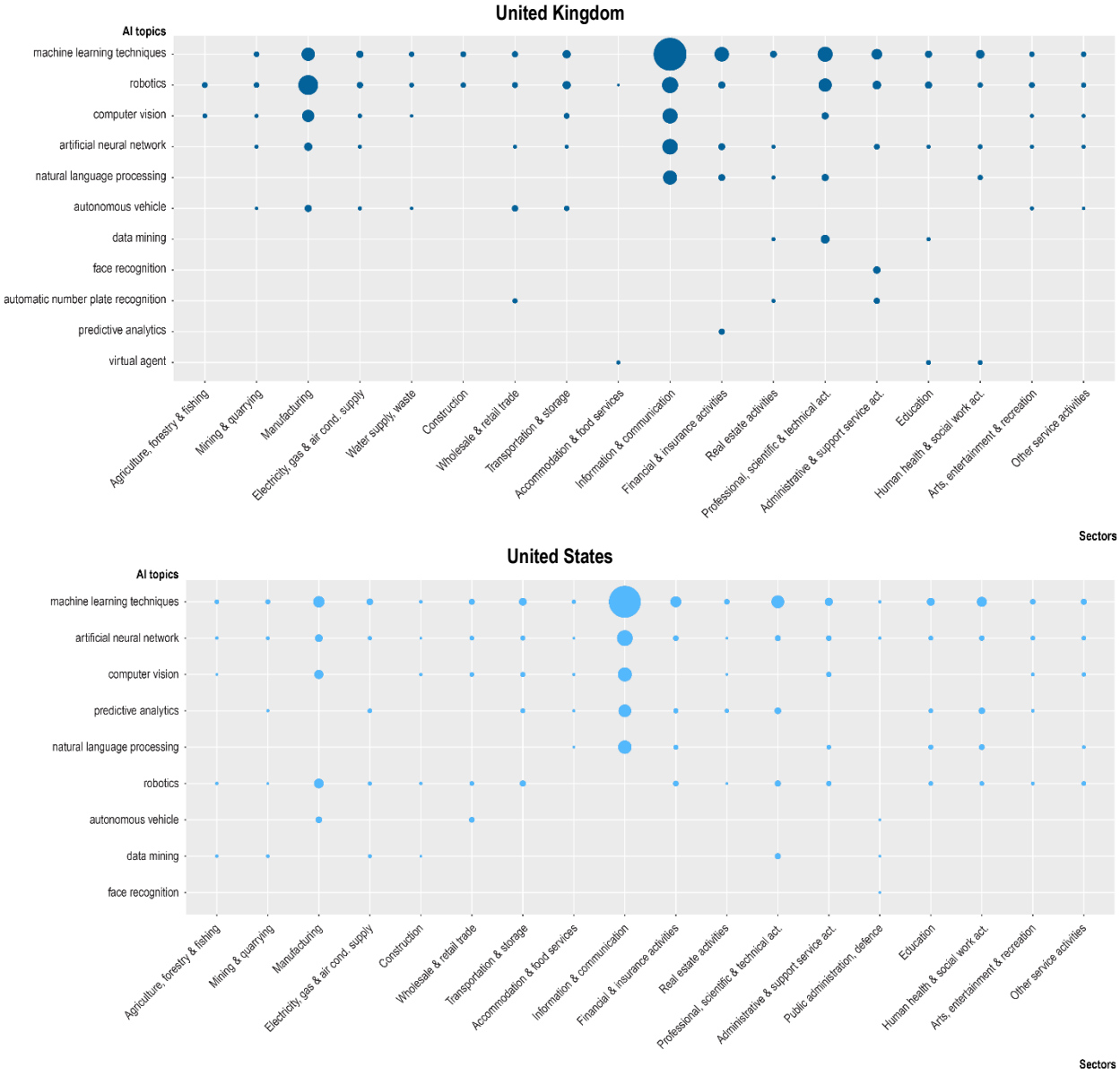
Number of companies featuring specific AI topics on their websites, ISIC, rev.4



Note: The figures display the top 5 AI related topics appearing on companies' websites by sector. The bubble size represents the proportion of firms featuring a given topic.
Source: OECD calculations based on GlassAI data, November 2022.

Figure 5.5. Top AI topics, by sector, United Kingdom and United States, 2020

Number of companies featuring specific AI topics on their websites, ISIC, rev.4



Note: The figures display the top 5 AI related topics appearing on the companies' websites by sector. The bubble size represents the proportion of firms featuring a given topic.
Source: OECD calculations based on GlassAI data, November 2022.

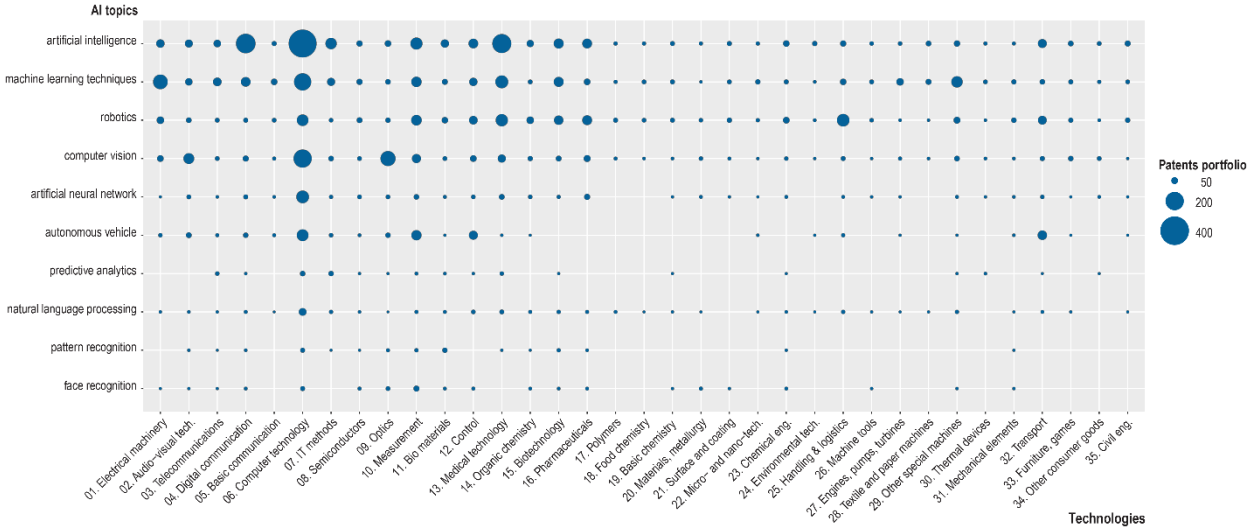
Technologies and products protected by companies with AI-related online presence

The analysis of the IP portfolio of companies with AI-related online presence can provide further insights into the AI activities of such companies, by characterising the type of technologies, goods and services they develop and/or commercialise (see Dernis et al. (2021[8])). With this aim, the sample of companies extracted by GlassAI was linked to administrative data on IP included in the STI Micro-data Lab infrastructure of the OECD (see Box 5.1 and Annex A). IP data constitute a rich source of information to assess technological developments (patents) and identify new products and services introduced on the market (registered trademarks). Noteworthy, firms dealing with AI techniques are likely to use other types of IP rights to get protection on different markets, especially copyrights. However, due to lack of comprehensive data on copyrights, the analysis focuses on patents and trademarks. The result of the matching exercise indicates that the two IP assets have recently been used by less than 10% of companies provided in the GlassAI sample, on average, for the four countries. The coverage of the patents and trademarks portfolios of companies is detailed in Box 5.1.

Technologies protected by patents were mapped to AI topics listed on companies' webpages: the distribution of the top 10 AI topics by technology are presented in Figure 5.6. Patent data are broken down into 35 technology domains, as described in Schmoch (2008[18]). Most companies develop technologies related to "Computer technology", whichever the AI topic considered. Patenting companies mentioning the generic topic *Artificial intelligence* on their websites also frequently protect inventions related to "Digital communication" (16%) or "Medical technologies" (15%). *Machine learning techniques* are associated with technologies related to "Electric machinery" (15%) and "Medical technologies" (11%). In turn, patenting companies featuring the topic *Robotics* are active in technology areas related to "Handling & logistics" (15%), "Medical technologies" (14%), "Measurement" (9%). The topic *Computer vision* is instead associated with patents protecting "Optics" (23%), and *Autonomous vehicles* with patents related to "Measurement" (22%), "Transport" (18%) or "Control" (15%).

Figure 5.6. AI topics and patented technologies

Combinations of AI topics with patented technologies, top 10 topics

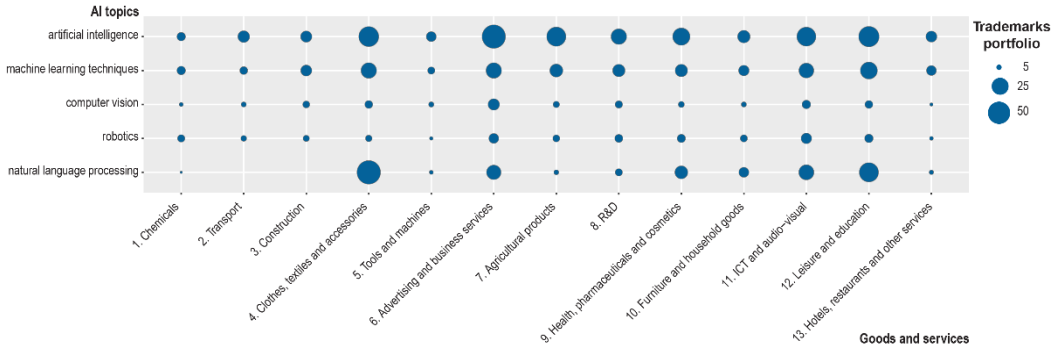


Note: Patents are allocated to technology fields on the basis of their International Patent Classification (IPC) codes, following the concordance provided by World Intellectual Property Organization (WIPO), using fractional counts. See Box 5.1 for further details on the data coverage.
Source: OECD calculations based on GlassAI data and OECD STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, November 2022.

As a final exploratory step, goods and services protected by trademarks were mapped to the AI topics listed on companies' webpages.[4] Trademarks were decomposed into 13 groups of goods and services, as described in Box 5.1 and Annex D. The distribution of AI topics of companies and trademarked products are presented for the top 5 AI topics in Figure 5.7. Companies that feature the generic term *Artificial Intelligence* on their websites and applied for trademarks mainly protected goods and services related to "Advertising and business services" (18%), "Leisure and education" (13%) and "Clothes, textiles and accessories" (13%). The term *Computer vision* is frequently associated with "Advertising and business services" (28%) trademarks, while companies claiming activities related to *Robotics* tend to protect trademarks in "ICT and audio-visual" goods and services (19%).

Figure 5.7. AI topics and trademarked goods and services

Combinations of AI topics with trademarked goods and services, top 5 topics



Note: Trademarks are allocated to groups of goods and services on the basis of their International (NICE) Classification of Goods and Services codes, following the concordance presented in Annex D, using fractional counts. See Box 5.1 for further details on the data coverage.
Source: OECD calculations based on GlassAI data and OECD STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, November 2022.

Box 5.1. Uncovering the IP portfolio of companies in the GlassAI sample

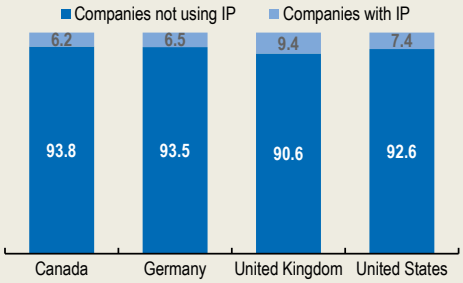
The names of companies provided in the GlassAI sample have been matched to the names of patent and trademark applicants, using the procedure described in Annex A. Patent data derives from the Worldwide Patent Statistical Database maintained by the European Patent Office (EPO)'s (also known as PATSTAT), in its Spring 2022 edition, as available in the STI Micro-data Lab infrastructure. Trademarks relate to those registered at the following three offices: the EU Intellectual Property Office (EUIPO), the Japan Patent Office (JPO) and the US Patent and Trademark Office (USPTO). The focus on patents and trademarks provides a partial picture of the IP portfolio of companies, as other types of IP protection, such as copyrights or trade secrets, might be preferred by AI companies. However, because of the lack of available data, these additional assets are not included in the analysis.

The name matching process enabled to identify IP data filed during the period 2010-18 for 63 companies located in Canada (6.2% of companies), 148 in Germany (6.5%), 213 in the United Kingdom (9.4%) and 607 in the United States (7.4%). The matching rates are lower for Germany, Canada and the United Kingdom, as trademark filings at the national IP office are not included in the STI Micro-data Lab infrastructure. Therefore, the trademark activity of AI firms presented for these three countries are likely to be partial, as, for the European area, only trademarks registered at the EUIPO are covered.

Patent-based indicators rely on families of patents filed within the Five IP offices (IP5 patent families), namely the EPO, the JPO, the Korean Intellectual Property Office (KIPO) and the China National Intellectual Property Administration (CNIPA). IP5 patent families are defined as sets of patent applications filed in several IP offices to protect the same invention, covering at least one of the IP5, provided that another family member has been filed in any other office worldwide, by earliest filing date observed in the family (see Dernis et al. (2015[19]) and Daiko et al. (2017[20]) for a discussion on IP5 patent families). Technology fields are defined using the concordance developed by the World Intellectual Property Organization (WIPO) between the International Patent Classification (IPC) and 35 technology domains (Schmoch, 2008[18]). Trademark indicators are based on registrations made at the EUIPO, JPO and USPTO only. Trademarks are allocated to products using the international classification of goods and services applied for the registration of marks, the NICE classification, aggregated into 13 groups (see Annex D).

Figure 5.8. IP use by companies with AI-related online presence, 2010-18

Share of companies using patents or trademarks



Source: OECD calculations based on GlassAI data and OECD STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, November 2022.

6 Zooming in on universities with AI-related online presence

While the previous sections focused on companies, data from online websites offer the unique opportunity to focus on other organisations that are central in the AI ecosystem. In this context, universities with AI-related online presence were identified using the same technique as that employed for companies, by searching for AI keywords on their websites. GlassAI data cover 75 universities located in Canada, 144 in Germany, 170 in the United Kingdom and 1 752 in the United States. However, the information extracted from websites is more limited compared to the variables available for companies, and include the name, address and a list of AI-related terms found on each university's website.

This information is described more in detail and visualised graphically in the following subsections, which provide a unique outlook about universities with AI-related online presence, which are likely to play a key role for AI-related tertiary education or research.

Location of universities with AI-related online presence

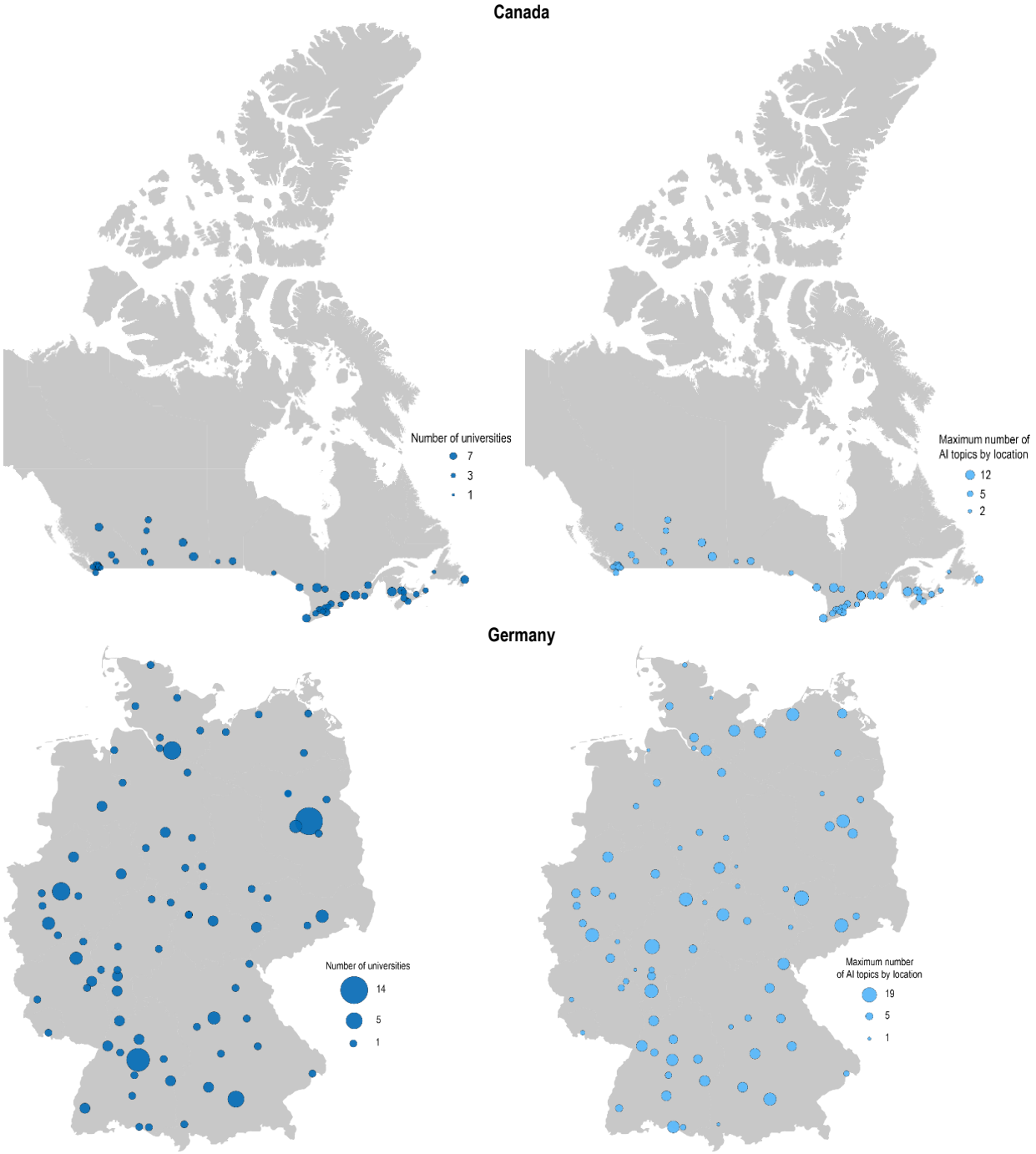
Comparisons between the location of universities and the number of AI terms identified in universities per location indicate the concentration of AI-related (likely education or research) activities within a given area. Therefore, Figure 6.1 and Figure 6.2 display the number of universities active in the AI space (left-hand side) and the maximum number of AI-related topics identified on universities' websites (right-hand side) for Canada and Germany, as well as the United Kingdom and the United States (only the conterminous states are presented on the map). It becomes evident that, although universities dealing with AI are spread all over the countries, they are concentrated in and around large cities.

In Canada (Figure 6.1), the regions of Ontario, British Columbia and Quebec host most of the AI-related universities, which tackle all a similar range of AI-related topics although the actual topics may differ across universities. In the case of Germany (Figure 6.1), the largest number of AI-related universities is located in Berlin, Stuttgart, Hamburg and Munich. These cities are also important hubs in which start-ups as well as car manufacturers tend to agglomerate,[5] which may explain the wide range of AI terms provided. Nevertheless, AI activities are also evident more widely in other federal states (Bundesländer), such as Hesse or North Rhine-Westphalia, where concentration (as evident by the maximum number of terms) is similarly high, e.g., Darmstadt and Giessen in Hesse, or Bochum in North Rhine-Westphalia.

Compared to Canada or Germany, AI-related universities are much more polarised with respect to their location in the United Kingdom (Figure 6.2). Although they are spread across the country and some smaller hubs are evident in and around Manchester as well as Birmingham, more than 30 universities active in the AI space are in London. Interestingly, this polarisation is not necessarily reflected in the AI intensity.

Figure 6.1. Universities active in AI, Canada and Germany, 2020

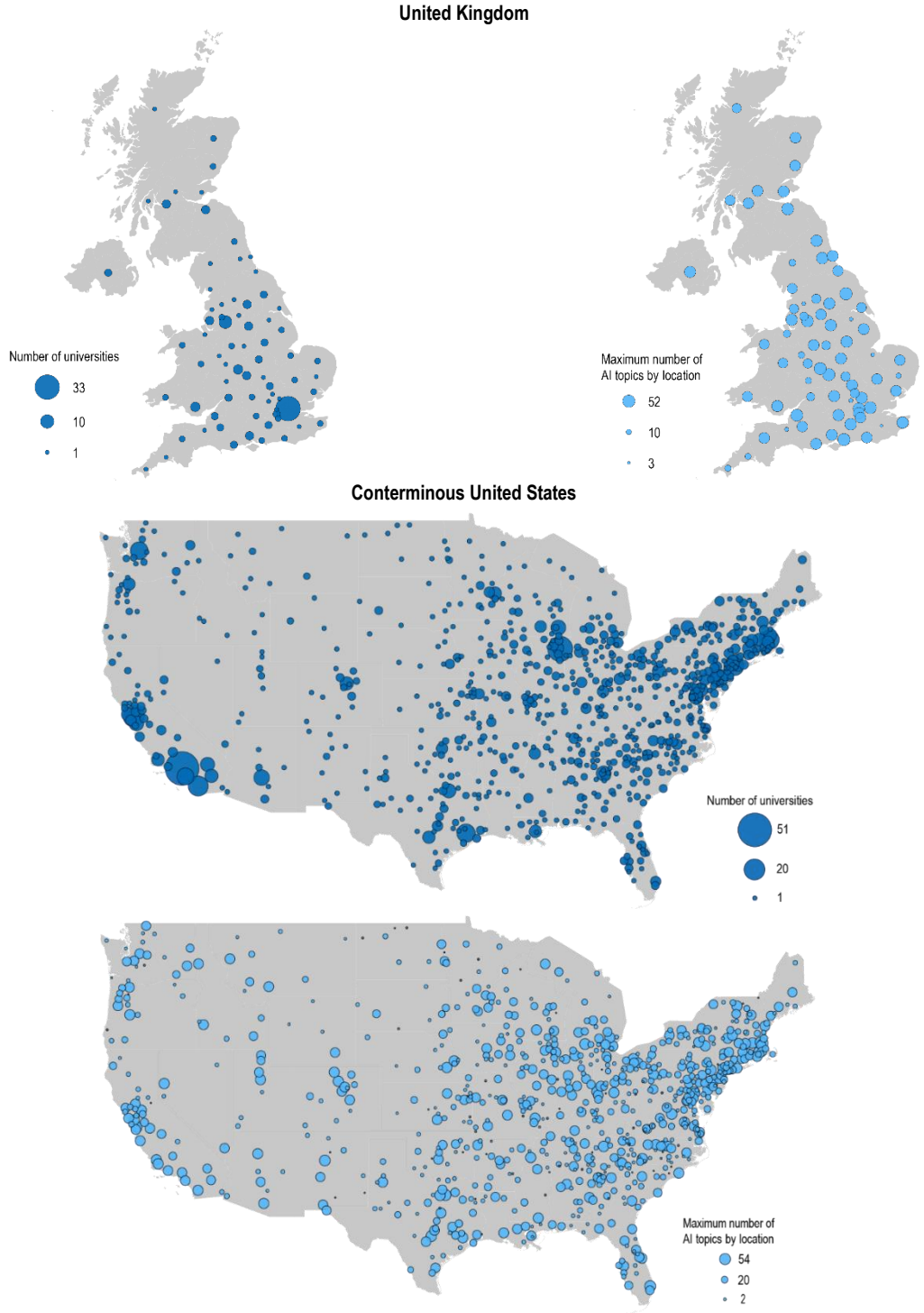
Location of universities signalling AI-related activities, number of universities and number of AI topics



Note: Maps display the number of universities in a given location, and the maximum number of AI topics identified on the website of universities in a given location.
Source: OECD calculations based on GlassAI data, November 2022. Maps designed in Cartes & Données © Articque

Figure 6.2. Universities active in AI, United Kingdom and United States, 2020

Location of universities signalling AI-related activities, number of universities and number of AI topics



Note: Maps display the number of universities in a given location, and the maximum number of AI topics identified on the website of universities in a given location.
Source: OECD calculations based on GlassAI data, November 2022. Maps designed in Cartes & Données © Articque

A large variety of AI terms is in fact evident across all AI universities in the United Kingdom, suggesting that many different AI topics are likely studied or investigated in all parts of the country. The landscape is similar in the United States (Figure 6.2), with a strong concentration of universities featuring AI on their websites in California (about 12% of US AI universities). This is followed by New York State, Texas, Pennsylvania and Illinois. However, the intensity with which AI-related (education or research) activities take place is – similarly to the United Kingdom – much more widespread across the United States. The number of AI terms also tends to be much higher in those two countries compared to Canada and Germany. However, this does not necessarily imply that AI universities in those countries are more AI intensive as it could also partly reflect keyword translation issues.

AI topics listed on universities' websites

Interestingly, the analysis of AI activities that was carried out for companies can also be applied to universities, uncovering their main AI-related topics. For consistency, the lists of AI-related topics provided by GlassAI are also further grouped for this exercise aggregating AI terms into AI groups relying on similar concepts (Annex C).

Figure 6.3 presents the top AI groups retrieved on universities' websites, with the word clouds referring to the top 20 topics based on their frequency of occurrence (left-hand side) and the bar charts presenting the top 10 topics based on the topic share in all AI-related universities in the respective country (right-hand side).

Although also very prevalent on companies' websites, the generic term *Artificial Intelligence* was featured in almost all universities. In Canada, 100% of AI-related universities and in the United Kingdom and the United States (note that this term was not included in the sample extracted for Germany) more than 90% flagged this term on their website. Therefore, it was removed from the word cloud images to allow for other AI-related topics to appear.

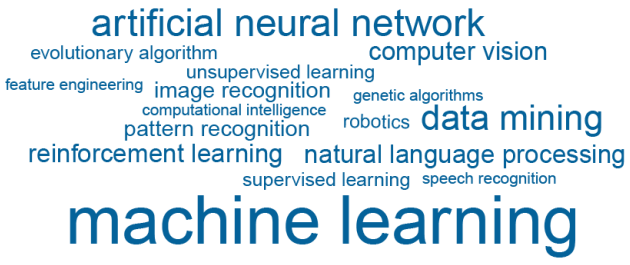
Machine learning techniques and *artificial neural network* are not only predominant on companies' but also on universities' websites in all countries for which data is available. However, there are some country specificities worth highlighting. For instance, in Canada and in the United States, applications of AI related to *computer vision* and *pattern/image recognition* appear among the top listed AI topics, which potentially signify a wide development of related applications. Computer vision is also prevalent among German universities but in this case a strong focus is also on *autonomous driving*, which is plausible given the strong automotive industry in Germany and the fact that autonomous cars often leverage computer vision technologies. Interestingly, *robotics* features very highly in all the AI-related universities but the Canadian ones.

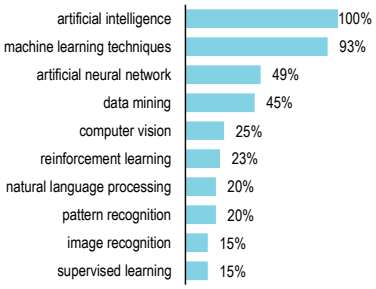
Overall, AI technologies (proxied by AI terms mentioned on websites) appear to be rather similar when comparing AI activities of companies and universities. While business and academic website profiles differ, they commonly refer to *machine learning techniques*, *computer vision* and *robotics*.

Figure 6.3. Top AI-related topics tackled by universities, by country, 2020

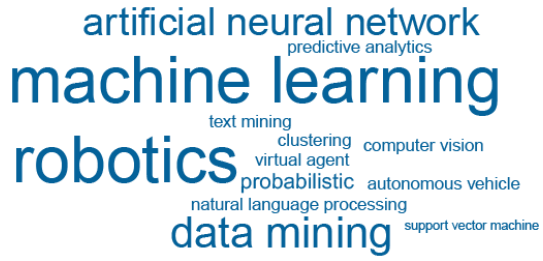
As share of total AI active universities

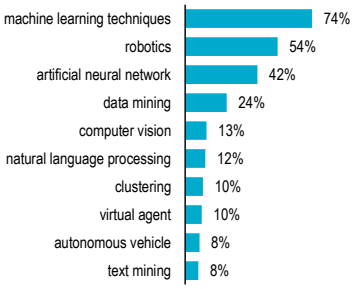
Canada



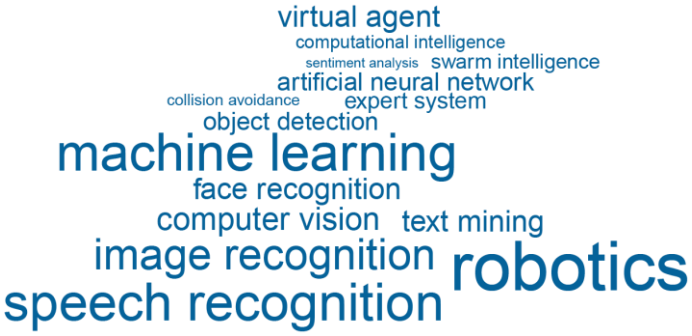


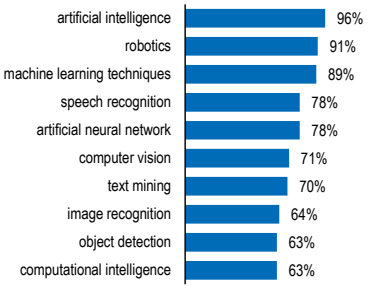
Germany



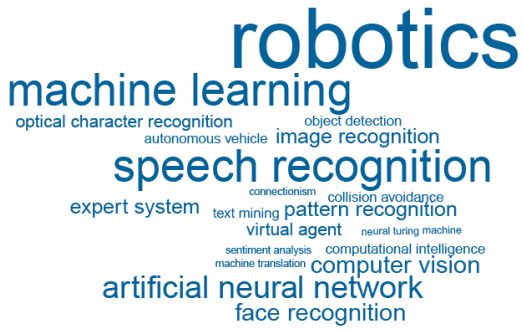


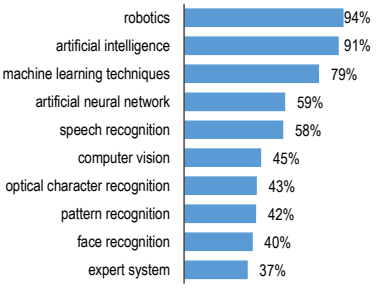
United Kingdom





United States



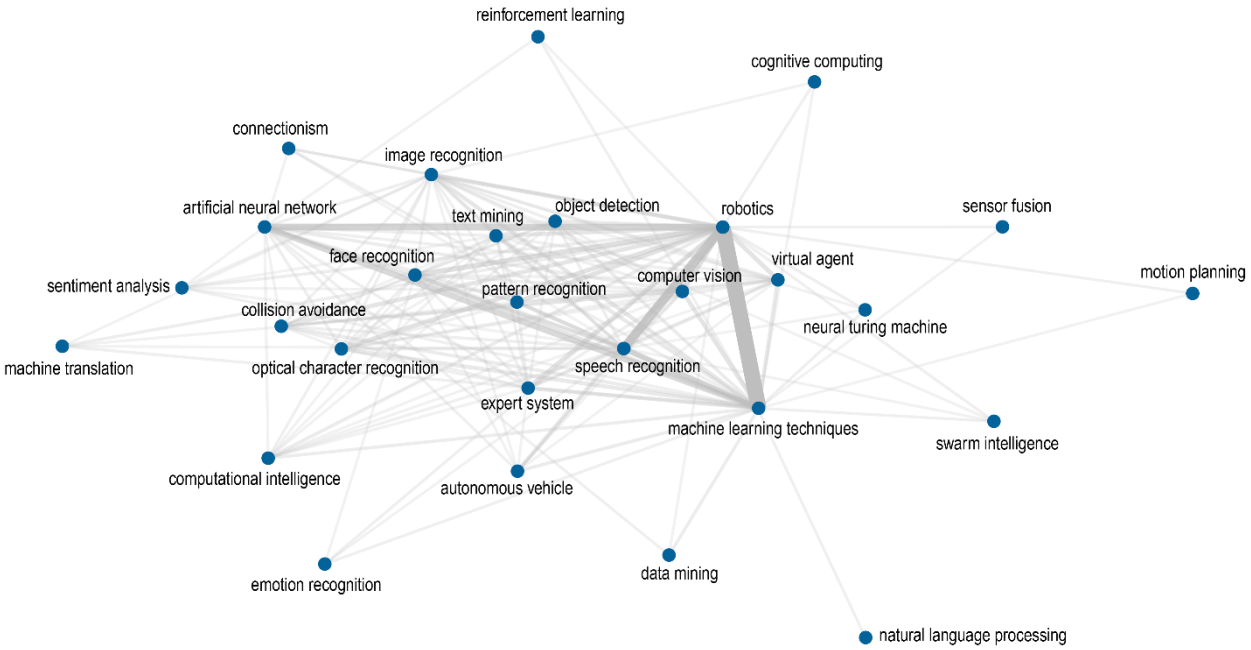


Note: Data relate to the number of universities in the GlassAI sample by specific AI-related topic. The generic topic "Artificial Intelligence" was excluded from the word cloud. The font size of certain terminology (e.g., machine learning, robotics, artificial neural networks) was intentionally reduced to facilitate the reading of the terms.
Source: OECD calculations based on GlassAI data, November 2022.

To get a better understanding of the way in which different AI technologies are studied, researched or used at universities, it is important to also examine the co-occurrences of these terms (Figure 6.4).

Across the four countries, *machine learning techniques* and *robotics* are often mentioned together on university websites; this pattern was also evident among companies. Those terms also appear often in connection with *artificial neural networks* and *speech recognition*. Topics related to other recognition or detection technologies are also central to the network but appear on university websites together with a large number of different technologies, rather than with the same ones most of the time.

Figure 6.4. Co-occurrence of AI topics in universities, 2020



Note: Data relate to the top co-occurrences of AI-related topics in universities active in AI, using fractional counts. The generic topic *Artificial Intelligence* was excluded from the calculation.
Source: OECD calculations based on GlassAI data, November 2022.

In general, when comparing co-occurrences of AI topics on company and university websites, it appears that companies tend to focus on relatively fewer topics, which are also often mentioned together, possibly implying a higher degree of specialisation. The network of topics identified on universities' websites, in contrast, is much more interconnected albeit exhibiting weaker nodes among each term, suggesting that university websites may present a wider range of topics. Note that those are not necessarily the same across universities within or across countries.

7 Discussion and concluding remarks

This work provides a comprehensive analysis focusing on organisations with AI-related online presence. For the first time, it analyses the characteristics of companies and universities that mention AI-related keywords on their websites across four countries: Canada, Germany, United Kingdom, and United States. The analysis relies on novel information collected and provided by GlassAI, a private company that reads and interprets open web text at scale.

Focusing on the characteristics of AI companies identified through web reading, the analysis highlights that those tend to be young and small. In particular, for three out of the four countries analysed, more than half of the companies in the sample have less than 5 years of age. In all countries analysed, about half of the firms flagging AI activity on their webpages are instead micro-firms, with less than 10 employees.

The large majority of these AI companies operates in the "Information & Communication" sector, which accounts alone for between 53% (the United Kingdom) and 64% (Canada) of AI firms. The sector ranking second is the "Professional, scientific and technical activities", which accounts for about 9% of companies that have an AI-related online presence.

Analysing their core activity, the analysis highlights that these firms tend to have AI at the core of their business. Applying a text mining approach to online company descriptions highlights that these companies appear often business oriented and aim at providing solutions to their customers. Many of these solutions tend to relate to *Data Analysis* or the provision of specialised services, such as *Cloud solutions.*

Relevantly, online websites provide a unique source to analyse the AI activities of companies that have an AI-related online presence. Beyond the generic *Artificial Intelligence* terms, *Machine Learning* emerges as a central topic, often co-occurring together with other AI techniques (e.g., *Artificial neural network*, *Natural language processing*, and *Predictive analysis*). Also applications – such as *Robots* or *Computer vision* – are among the terms often appearing on the websites of such companies. These topics appear strongly connected especially in Germany.

The types of AI related activities reported by companies vary across sectors. While the leading topics (*Machine learning*, *Artificial neural network*, *Natural language processing*, *Computer vision*) prevail in the "Information & communication" sector, differences are observed in other sectors. For instance, *Robotics* is one of the top 5 subject areas signalled by companies performing in the "Manufacturing" sector across countries, while in Germany, the United Kingdom and the United States, companies from this sector also contribute to AI developments related to *Autonomous vehicles*.

Exploring the patenting or trademark activity of companies highlights that a low proportion (less than 10%) of firms that have AI-related online presence also holds these IP rights. Most patenting companies develop technologies related to "Computer technology". Zooming in on universities with AI-related online presence highlights the strong concentration of universities in and around large cities. However, this polarisation, which is especially evident in the United Kingdom (London) and the United States (California) is not necessarily reflected in the intensity with which AI-related activities take place (proxied by the maximum number of AI-related topics). While business and academic website profiles differ, the latter also commonly refer to machine learning techniques, computer vision and robotics.

Although comprehensive along several dimensions, the current analysis could be further extended in different directions. First, future work could further analyse the patterns of technology adoption over time, e.g., by exploiting website information available in the Internet Archive. This could be particularly relevant in the context of the COVID-19 pandemic, to analyse the extent to which different companies started mentioning specific technologies on their websites at different points in time. Future analysis could also aim at combining and comparing the current data with additional information on AI start-ups, for instance available in other commercial databases. Finally, future work may broaden the scope of analysis based on data from online websites, for instance extending the country coverage or focusing on different technological areas that may be complementary to the one examined in this paper (e.g., green technologies, in the context of the twin transition).

Endnotes

[1] Given the very nature of the data source, the sample may include few organisations that feature AI-related terms on their webpages while not implementing those in their main activities. This is further discussed in Section 3, together with additional details on the data and methodology.

[2] See https://www.glass.ai/.

[3] The authors first study the web address (URL) coverage across firms, finding that it varies with firm characteristics (as derived from the Mannheim Enterprise Panel and the European Patent Office). Patenting firms and almost all medium to large sized firms have websites, while very young and very small firms tend to exhibit a lower URL coverage.

[4] Trademark activity is however likely not to be fully comprehensive, so findings from this exploratory exercise should be taken with caution, see Box 5.1 for further details.

[5] See https://startupverband.de/fileadmin/startupverband/mediaarchiv/research/dsm/DSM_2022.pdf and https://www.gtai.de/resource/blob/64100/817a53ea3398a88b83173d5b800123f9/industry-overview-automotive-industry-en-data.pdf

References

Alekseeva, L. et al. (2021), "The demand for AI skills in the labor market", *Labour Economics*, Vol. 71, p. 102002, https://doi.org/10.1016/j.labeco.2021.102002. [11]

Alekseeva, L. et al. (2020), "AI Adoption and Firm Performance: Management versus IT", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3677237. [10]

Babina, T. et al. (2022), "Artificial Intelligence, Firm Growth, and Product Innovation", *SSRN Electronic Journal*, https://doi.org/10.2139/SSRN.3651052. [12]

Bajgar, M. et al. (2020), "Coverage and representativeness of Orbis data", *OECD Science, Technology and Industry Working Papers*, No. 2020/06, OECD Publishing, Paris, https://doi.org/10.1787/c7bdaa03-en. [17]

Baruffaldi, S. et al. (2020), "Identifying and measuring developments in artificial intelligence: Making the impossible possible", *OECD Science, Technology and Industry Working Papers*, No. 2020/05, OECD Publishing, Paris, https://doi.org/10.1787/5f65ff7e-en. [7]

Calvino, F. and L. Fontanelli (2022), "A portrait of AI adopters across countries: firm characteristics, assets' complementarity, and productivity", OECD Internal document. [2]

Calvino, F. et al. (2022), "Identifying and characterising AI adopters: A novel approach based on big data", *OECD Science, Technology and Industry Working Papers*, No. 2022/06, OECD Publishing, Paris, https://doi.org/10.1787/154981d7-en. [3]

Daas, P. and S. van der Doef (2021), "Using Website texts to detect Innovative Companies", No. 01-21, CBDS Working paper, https://www.cbs.nl/en-gb/about-us/innovation/project/using-website-texts-to-detect-innovative-companies (accessed on 2 December 2022). [14]

Daiko Taro et al. (2017), "World Top R&D Investors: Industrial Property Strategies in the Digital Economy", *Publications Office of the European Union*, https://doi.org/10.2760/861062. [20]

Dernis, H. et al. (2015), "World Corporate Top R&D Investors: Innovation and IP bundles", *JRC Working Papers*, https://ideas.repec.org/p/ipt/iptwpa/jrc94932.html (accessed on 22 June 2020). [19]

Dernis, H. et al. (2021), "Who develops AI-related innovations, goods and services? : A firm-level analysis", *OECD Science, Technology and Industry Policy Papers*, No. 121, OECD Publishing, Paris, https://doi.org/10.1787/3e4aedd4-en. [8]

Kinne, J. and J. Axenbeck (2020), "Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study", *Scientometrics*, Vol. 125/3, pp. 2011-2041, https://doi.org/10.1007/S11192-020-03726-9/TABLES/6. [13]

Krüger, M. et al. (2020), "The Digital Layer: How innovative firms relate on the Web", *Discussion Paper*, No. 20-003, ZEW - Centre for European Economic Research, https://ssrn.com/abstract=3530807 (accessed on 2 December 2022). [15]

Levenshtein, V. (1965), "Binary codes capable of correcting deletions, insertions and reversals.", *Soviet Physics Doklady 10 (8): 707--710 (February 1966) translated from Doklady Akademii Nauk SSSR, V163 No4 845-848*, https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf (accessed on 23 July 2020). [21]

Murzintcev, N. (2020), *Select number of topics for LDA model*, https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html (accessed on 11 January 2023). [23]

Nakazato, S. and M. Squicciarini (2021), "Artificial intelligence companies, goods and services: A trademark-based analysis", *OECD Science, Technology and Industry Working Papers*, No. 2021/06, OECD Publishing, Paris, https://doi.org/10.1787/2db2d7f4-en. [6]

Nathan, M. and A. Rosso (2022), "Innovative events: product launches, innovation and firm performance", *Research Policy*, Vol. 51/1, p. 104373, https://doi.org/10.1016/J.RESPOL.2021.104373. [16]

OECD (2019), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)", *OECD Digital Economy Papers*, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en. [1]

Samek, L., M. Squicciarini and E. Cammeraat (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. [5]

Schmoch, U. (2008), *Concept of a Technology Classification for Country Comparisons - Final Report to the World Intellectual Property Organisation (WIPO)*. [18]

Squicciarini, M. and H. Nachtigall (2021), "Demand for AI skills in jobs: Evidence from online job postings", *OECD Science, Technology and Industry Working Papers*, No. 2021/03, OECD Publishing, Paris, https://doi.org/10.1787/3ed32d94-en. [4]

Tambe, P. (2013), "Big Data Investment, Skills, and Firm Value", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.2294077. [9]

Winkler, W. (1999), "The State of Record Linkage and Current Research Problems", U.S. Bureau of the Census, https://www.census.gov/srd/papers/pdf/rr99-04.pdf (accessed on 23 July 2020). [22]

Annex A. Linking data using string matching algorithms

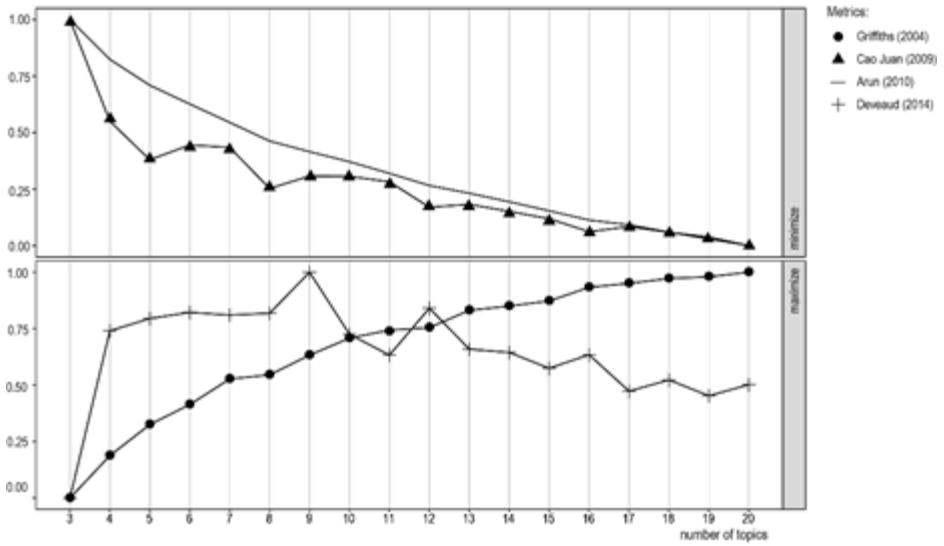
The matching of company names between two different sources is carried out on a by-country basis using a series of algorithms contained in the Imalinker (Idener Multi Algorithm Linker) system developed by IDENER (http://www.idener.es/). The matching exercise was notably implemented on the whole population of ORBIS© companies, and patent and trademark applicants included in the STI Micro-Data-Lab over a number of key steps:

1. For each data source, the names of companies are harmonised using country-specific 'dictionaries' developed internally. These aim to deal with legal entity denomination (e.g. 'Limited' and 'Ltd'), common names and expressions, as well as phonetic and linguistic rules that might affect the way in which company names are written. Failing to account for such features of the data might mistakenly lead to excluding a company (not considering only because its name had been misspelt or shortened in some places), or double counting a company (because different spellings of its name made it appear to be different entities).
2. In a second step, a series of string-matching algorithms – mainly token-based and string-metric-based, such as token frequency matching and Levenshtein (1965[21]) and Jaro-Winkler (1999[22]) distances – are applied to compare the harmonised names from the different datasets and to provide a matching accuracy score for each pair of names. The precision of the match, which depends on minimising the number of false positive matches, is ensured through a selection of pairs of names made on the basis of high-score thresholds imposed on the algorithm.
3. A manual post-processing stage is then applied, consisting in:
   a. reviewing the results of the matches;
   b. assessing the proportion of non-matched companies (possibly false negatives, that is, names that the algorithm had failed to recognise as part of the sample); and
   c. identifying new matches on a case-by-case basis (e.g. allowing for lower thresholds for a given algorithm), by correcting and augmenting dictionaries and through manual searches.

Annex B. Characterising AI-related activities using topic modelling

While running topic modelling, the choice of the optimal number of clusters (or topics) is non-trivial: it is typically the one after which only marginal improvements are observed in the data distribution, that will be best suited for human interpretation of the clusters. Four different metrics were compiled on the data to identify the optimal number of topics, as plotted below. For further elaboration and links to the different metrics, see Murzintcev (2020[23]) accessible here: https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html

Figure A B.1. Number of topics by LDA metrics



Source: OECD calculations based on GlassAI data, November 2022

Annex C. Dictionary of AI-related topics

AI topics	Related AI terms
action recognition	*activity recognition; human action recognition; human activity recognition*
artificial intelligence	*ambient intelligence; artificial general intelligence; artificial intelligence solutions; machine intelligence; intelligence augmentation; intelligent agent; intelligent infrastructure; machine perception*
artificial neural network	*deep learning; deep neural network; neural network; apprentissage automatique; apprentissage profond; deep convolutional neural network; recurrent neural network; convolutional neural*
association rule	
autoencoder	
automatic number plate recognition	
autonomic computing	
autonomous vehicle	*autonomous car; autonomous driving; driverless car; driverless vehicle; self-driving car*
bayesian networks	*bayesian learning; naive bayes classifier; bayesian statistics*
brain computer interface	
classifier	*classification tree*
clustering	*cluster analysis; hierarchical clustering*
cognitive computing	*cognitive modelling; cognitive automation*
collaborative filtering	
collision avoidance	*pedestrian detection; obstacle avoidance*
computational intelligence	*computational learning*
computational pathology	
computer vision	*machine vision*
connectionism	*connectionist*
conversational interface	
cyber physical system	*visual servoing; layered control system*
data mining	*mapreduce; information retrieval; dimensionality reduction; rough set; dynamic time warping*
decision model	*decision tree*
emotion recognition	*facial expression recognition*
evolutionary algorithm	*evolutionary computation*
expert system	*expert systems*
face recognition	*facial recognition*
feature engineering	*feature extraction; feature selection; high-dimensional data; feature learning*
fingerprint recognition	
fuzzy logic	*fuzzy set*
generative adversarial network	
genetic algorithms	*genetic algorithm; genetic programming*
gesture recognition	
gradient boosting	
image recognition	*image classification; image segmentation; image processing; image retrieval*
independent component analysis	
inductive logic programming	
k-means	
logistic regression	
machine learning techniques	*advanced machine learning; machine learning; machine learning platform; machine learning models; learning model; learning algorithm; learning automata; learning classifier system; q-learning; relational learning; semi-supervised learning; similarity learning; kernel learning; latent variable*
machine translation	
meta learning	*rule learning; rule-based learning*
motion planning	
multi-agent system	
multi-objective optimization	
natural language processing	*natural language generation; natural language understanding*
neural turing machine	*turing test*
neuromorphic computing	
object detection	*object recognition*
optical character recognition	
pattern recognition	
predictive analytics	*regression tree*
probabilistic	*gaussian process*
random forest	*random field*
recommender system	*recommendation engine*
reinforcement learning	
robotics	*advanced robotics; industrial robot; mobile robot; quadruped robot; robot; service robot; legged robot; humanoid robot; social robot; surgical robot; biped robot; wheeled mobile robot*
sensor fusion	*sensor data fusion; multi-sensor fusion*
sentiment analysis	
speech recognition	*speech to text; text to speech*
supervised learning	*backpropagation; instance-based learning*
support vector machine	*support vector regression; vector machine*
swarm intelligence	*particle swarm optimisation; swarm behavior; swarm optimisation*
symbolic computation	
text mining	*text analytics; word embedding; semantic web; latent semantic analysis*
topic model	
transfer learning	
unsupervised learning	*hebbian learning; self-organizing map*
virtual agent	*human-robot interaction; conversational interface; chatbot*
virtual agent	
word2vec	
xgboost	

Source: OECD, 2022

Annex D. Aggregation of NICE classes by fields

1. Chemicals
1. Chemical goods
2. Paints and colorants
4. Oils and fuels

2. Transport
12. Vehicles
39. Transport and packaging

3. Construction
6. Metals
17. Rubber and plastics
19. Building material
27. Carpets and floor covers
37. Building services

4. Clothes, textiles and accessories
14. Precious goods
18. Leather and complements
22. Fibrous products
23. Yarns and threads
24. Textiles
25. Clothing and footwear
26. Decorations

5. Tools and machines
7. Machineries
8. Hand tools

6. Advertising and business services
35. Business and advertising
36. Insurance and finance
45. Legal and personal services

7. Agricultural products
29. Food
30. Condiments and cereals
31. Animals and grains
32. Low and non alcohol drinks
33. Alcoholic drinks
34. Tobaccos.

8. R&D
42. R&D and software

9. Health, pharmaceuticals and cosmetics
3. Cleaning products
5. Pharmaceutical products
10. Medical instruments
44. Medical and hygiene services

10. Furniture and household goods
11. Lightening and heating
20. Furniture
21. House utensils

11. ICT and audio-visual
9. Instruments & computers
38. Telecommunications

12. Leisure and education
13. Firearms
15. Musical instruments
16. Papers and packaging
28. Games
41. Education and sport

13. Hotels, restaurants and other services
40. Treatment of materials
43. Food, drink and accommodation

Source: OECD, groupings based on WIPO, Nice classification, http://www.wipo.int/classifications/nice/en/