OECD Science, Technology and Industry Working Papers
2020/05

Identifying and measuring developments in artificial intelligence: Making the impossible possible

Stefano Baruffaldi,

Brigitte van Beuzekom,

Hélène Dernis,

Dietmar Harhoff,

Nandan Rao,

David Rosenfeld,

Mariagrazia Squicciarini

https://dx.doi.org/10.1787/5f65ff7e-en

OECD Science, Technology and Industry Working Papers

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the authors. Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to OECD Directorate for Science, Technology and Innovation, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France; e-mail: sti.contact@oecd.org .

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

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Acknowledgments

The authors are grateful to participants in the workshop “Patent-based definition of Artificial Intelligence” for their valuable inputs, especially members of the OECD-led Intellectual Property Statistics Task Force, namely: Erica Fraser, Sean Martineau and Anne-Marie Monteith from the Canadian Intellectual Property Office (CIPO); Markus Arndt, Yann Ménière and Ilja Rudyk from the European Patent Office (EPO); Alvaro Gonzalez Lopez of Instituto Nacional de Propiedad Industrial (INAPI, Chile); Rich Corken and Christopher Harrisson from the United Kingdom Intellectual Property Office (UK IPO); Catriona Bruce and Pushpika Wijesinghe from IP Australia; Uri Coppenhagen of the Israel Patent Office (ILPO); Yoichi Kaneki and Shohei Nakazato of the Japan Patent Office (JPO); Alistair Nolan and Karine Perset (OECD); Simone Bruschi Polidori and Lino Fanella from Ufficio Italiano Brevetti e Marchi (UIBM); and, James Forman, Nicholas Pairolero and Andrew Toole from the United States Patent and Trademark Office (USPTO).

The authors are grateful to Alessandra Colecchia, Antoine Dechezleprêtre, Fernando Galindo-Rueda, Dominique Guellec, Dirk Pilat and Andrew Wyckoff (OECD) for comments on earlier versions of this paper. Jonathan Frankle and Tad Reynolds (Massachusetts Institute of Technology) as well as Ken Goldberg (University of California, Berkeley) also provided invaluable comments, especially with respect to the keywords related to Artificial Intelligence. Thanks also go to the OECD Economics Department for the feedback provided.

Nandan Rao and David Rosenfeld were both at the OECD at the time the work was performed.

A previous version of this document was presented and discussed by, among others, the OECD Working Party on Industry Analysis (WPIA), the OECD Working Party on Measurement and Analysis of the Digital Economy (WPMADE), the Committee on Industry, Innovation and Entrepreneurship (CIIE) and is available on O.N.E under the reference code: DSTI/CIIE/WPIA(2018)4.

Identifying and Measuring Developments in Artificial Intelligence: making the impossible possible

Stefano Baruffaldi*, Brigitte van Beuzekom, Hélène Dernis, Dietmar Harhoff*,
Nandan Rao, David Rosenfeld and Mariagrazia Squicciarini

This paper identifies and measures AI-related developments in science, algorithms and technologies using information from scientific publications, open source software (OSS) and patents. A three-pronged approach relying on established bibliometric and patent-based methods, and machine learning (ML) implemented on purposely collected OSS data, unveils a marked increase of AI-related developments over time. Since 2015, AI-related publications increased at 23% a year; in 2014-18, OSS commits related to AI grew about three times as other OSS contributions; in 2017 the share of AI-related IP5 patent families averaged more than 2.3%. The growing role of China in the AI space emerges throughout.

* Max Planck Institute for Innovation and Competition (MPI Munich)

Table of contents

Tables

Figures

Executive Summary

Artificial Intelligence (AI) is a term commonly used to describe machines performing human-like cognitive functions (e.g. learning, understanding, reasoning and interacting). AI is expected to have far-ranging economic repercussions, as it has the potential to revolutionise production, to influence the behaviour economic actors and to transform economies and societies.

The vast potential of this (now considered) general purpose technology has led OECD countries and G20 economies to agree on key principles aimed at fostering the development of ethical and trustworthy AI (OECD 2019). The practical implementation of such principles nevertheless requires a common understanding of what AI is and is made of, in terms of both scientific and technological developments, as well as possible applications.

Addressing the challenges inherent in delineating the boundaries of such a complex subject matter, this paper proposes an operational definition of AI based on the identification and measurement of AI-related developments in science, algorithms and technologies. The analysis is based on information contained in scientific publications, open source software and patents and results from collaboration with the Max Planck Institute for Innovation and Competition (MPI Munich). The work has further benefitted from advice from leading AI scientists and from patent experts, including patent examiners, members to the OECD-led Intellectual Property (IP) Statistics Task Force.

The three pronged approach developed here relies on an array of established bibliometric and patent-based methods, and is complemented by an experimental machine learning (ML) approach implemented on purposely collected open source software data. We do so as traditional approaches, such as those based on keywords identification and/or classifications, are relatively "easy" to implement and have a demonstrated ability to deliver sound results, whereas ML-based techniques, while non-trivial to design and implement, are still in a development phase and at times deliver results that may be difficult to assess or interpret.

While the search strategy detailed in the paper and outlined below aims to produce an encompassing operational definition of AI, such definition can only account for past and present developments, and will need to be periodically revised and refined, as AI evolves:

- The identification of the science behind AI developments is based on a bibliometric two-step approach, whereby a first set of AI-relevant keywords is extracted from scientific publications classified as AI in the Elsevier's Scopus® database. This set is then augmented and refined using text mining techniques and expert validation. This two-step approach leads to identifying 168 groupings of AI-related terms (and variations thereof, e.g. convolutional neural networks and neural networks). Scientific publications and conference proceeding articles are finally tagged as being AI-related if they contain in their abstract at least two AI keywords related to different groupings. This is done to contain the number of false positives and minimise over identification.
- As AI is ultimately implemented in the form of algorithms, and in the impossibility to access data related to private firms' AI software, we use open-source software's information about software commits (i.e. contributions) posted on GitHub (an online hosting platform) to track AI-related software developments and applications. Such data are combined with information from papers presented at key AI conferences to identify "core" AI repositories. Machine learning techniques

trained using information for the thus identified core set are used to explore the whole set of software contributions in GitHub to identify all AI-related repositories.

- Information contained in patent data is used to identify and map AI-related inventions and new technological developments embedding AI-related components, independently of the technological domain in which they occur. Text mining techniques are used to search abstracts and patent documents referring to AI-related papers. This leads to identifying the International Patent Classification (IPC) codes most frequently allocated to AI-related inventions. Such list of IPC codes, upon validation by patent examiners and experts in the field, is refined so that some IPC codes are supposed to be considered in full as being AI-related, whereas identification of other patent codes needs to rely on keyword searches on patent. Finally, experts agreed to implement refined keyword-only searches to identify AI developments happening in other technology areas.

A number of stylised facts emerge upon implementation of the approach detailed above:

- An acceleration in the number of publications in AI in the early 2000s, followed by a steady growth of 10% a year on average until 2015, before accelerating again at a pace of 23% a year since then. The share of AI-related publications in total publications increased to over 2.2% of all publications in 2018.
- 28% of the world AI-related papers published in 2016-18 belongs to authors with affiliations in China. Over time, the share of AI publications originating from EU28, the United States and Japan has been decreasing, as compared to the levels observed ten years earlier.
- Since 2014, the number of open-source software repositories related to AI has grown about three times as much as the rest of open-source software.
- Topic modelling implemented on the content of AI-related commits offers interesting insights about the specific fields and applications embedding AI. Among them, text mining, image recognition and biology.
- It can also be appreciated that many AI-related software contributions are general in nature, and at the basis of several of topic areas identified.
- Figures based on IP5 patent families[1] exhibit a marked increase in the proportion of AI-related inventions over the total number of inventions after 2015. This ratio averaged to more than 2.3% in 2017.
- *Neural networks* and *image processing* are the most frequent terms appearing in the abstracts of AI-related patents.
- In 2014-16, Japan, the United States and China represent the top three countries in which the inventors of AI-patents are located.
- In AI-related patents, the contribution of China-based inventors multiplied more than six fold since the mid-2000s, reaching nearly 13% in the mid 2010s.

[1] Inventions protected in at least two jurisdictions, at least one of which needs being one of the Five IP Office (the European Patent Office, the Japan Patent Office, the Korean Intellectual Property Office, the US Patent and Trademark Office and the National Intellectual Property Administration of People's Republic of China).

1. Introduction

Artificial Intelligence (AI) is a term used to describe machines performing human-like cognitive functions (e.g. learning, understanding, reasoning and interacting). It has the potential to revolutionise production as well as contributing to tackling global challenges related health, transport and the environment (OECD, 2017).

AI is high on the agenda of businesses and policy makers alike: many observers expect AI to have far-ranging economic repercussions in the near future. The boundaries of this complex subject matter, which has attracted the imagination of writers and scientists for generations, are difficult to identify and to delineate in a neat way. Also, due to its popularity, the locution AI is at times overused or misused, making it harder for analysts to clearly decide what is AI and what is not AI.

Important developments in AI began in the 1950s, when pioneers in mathematics, psychology, and statistics set out to work on a number of concrete problems to measure progress towards goals of general intelligence. These included playing games, classifying images, and understanding natural language. Since then AI has evolved significantly, and while the interest in, and optimism regarding, "general" AI has waxed and waned over the years, progress and success in solving some specific problems has led to the development of subfields. Some of them, such as machine vision, speech recognition, and machine translation (often referred to as “weak” AI or “Artificial Narrow Intelligence”), have become commercially viable in recent years.

Along with recent success in specific tasks, came renewed interest in the 2010s of pursuing general AI again[2]. Technologies developed by AI researchers became extremely valuable in and of themselves, as well as for many other purposes and developments. Machine learning is one such technology. Arthur Samuel (1959) is credited as the father of this technology. Machine learning, which involves parameterising a decision making process and letting the machine learn the correct parameters by “training” on examples, has since been developed into what is the dominant technique in AI research today and a useful tool in many other areas[3]. This combination of interdisciplinary origins, wavering trajectories, and recent commercial success make "artificial intelligence" a difficult concept to define and measure. The term itself is used interchangeably both as the still-faraway goal of true machine intelligence and as the currently available technology powering today’s hottest startups.

[2] OpenAI, for example, is a celebrated research organisation working towards machine intelligence which can "reach human performance on virtually every intellectual task."

[3] A seminal moment for machine learning occurred with the publication of a paper on the applications of deep learning techniques to image classification, by Krizhevsky et al. (2012), at the Twenty-sixth Conference on Neural Information Processing Systems (NIPS).

Box 1. Recommendation of the OECD Council on Artificial Intelligence

In May 2019, the OECD Council adopted principles on Artificial Intelligence proposed by the AI Group of Experts at the OECD (AIGO), agreeing on the understanding of AI terminology: AI systems are thus understood as "(…) *machine based system that can, for a given set of human defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy*" (OECD, 2019). This agreement on a common understanding of AI is an important milestone for developing an operational definition of AI.

1.1. Identifying and measuring AI is difficult…

Efforts to measure AI and to map AI developments are fraught with difficulty, and the scope of what could be considered AI-related covers a wide range:

The academic field of AI research itself. This might involve the pursuit of "general AI", the act of pursuing fundamental research, as well as solving AI-general problems that can be translated into commercial applications:

- Techniques, originally invented within the field of general AI, that now constitute their own field of research. Currently this is dominated by machine learning and related sub-subfields, such as statistical learning theory or neural networks;
- Individual problems, such as image recognition, machine translation, or voice recognition, that have become commercial applications in and of themselves.

The application of techniques developed in AI research to other domains. Examples are the use of machine learning in genetics, ecology, economics) and other domains.

Distinct, non-AI fields of study that have provided direct influence on AI. These include statistics, optimal control theory, mathematics and optimisation, parallel/distributed computing, microprocessors). Improvements in these fields continue to fuel AI's progress, and can be broken down into:

- Developments that coincidentally contribute to AI, e.g. advances in distributed computing;
- Developments that are researched and developed specifically for techniques used in AI research. Examples are neuromorphic chips designed for neural networks.

In addition, difficulties arise in that some fields evoke AI, as is the case of e.g. robotics, but not always rely on AI as such. Optimal control theory, for example, is heavily used in industrial robots and autonomous vehicles. However, such field has been developing within engineering as a very distinct academic pursuit from that of AI and represents one of the fields that may work in combination with AI, but are not the same.

Finally, as any science and technology field, AI spans the range of basic research to applied innovation. In addition, as a field related to computer science, it is implemented through software code. Hence, in order to capture AI developments and to shed light on this new technological paradigm it is important to gather and use data proxying or mirroring developments in basic science, technological innovations and software.

1.2. … but not impossible

Despite the difficulties mentioned above, operationally defining and mapping AI developments remains a must: in the absence of measurement and empirical evidence, policies may be become ineffective if not distortive. To this end, and to substantiate the policy discussion on AI, in what follows we pursue a three-pronged approach measuring AI developments in:

1) *science*, as captured in scientific publications;
2) *technological developments*, as proxied by patents; and
3) *software*, and in particular open source software[4].

The rationale behind using these three sources of information is to provide as complete a view as possible of AI. In each case, the same overall steps were used to identify AI, using text mining methods. To this end, the approach implemented entailed:

I. Identifying documents (publications, patents and software) which are unambiguously AI-related, using expert advice (both direct and indirect);
II. Picking a method to measure document similarity.
III. Finding “similar” documents to those identified in step 1, and labelling those as AI-related.

This approach was operationalised in slightly different ways, depending on the source of information considered, in order to maximally exploit the information contained in the different corpuses and to account for their peculiarities, as detailed in the methodological descriptions that follow.

The strength of this approach is twofold. On the one hand, it provides a more complete view of the AI phenomenon than the one that would be obtained by looking at academic publications, patents and software individually. On the other hand, the approach proposed here and the sub-approaches implemented (e.g. keyword search, word embedding, etc., as detailed later) may represent the elements of a more general approach, to be used to define and measure any new technological trajectory or paradigm[5] emerging in the future.

Evidently, the fact that the approaches pursued differ somewhat for each corpus considered means that the results may not be fully comparable across the different sources of information used. Also, the outcomes and the statistics reported in the present report are to be considered as first results to be further refined: they hinge heavily on the set of criteria used to select the most similar elements (i.e. documents, patents, software packages) and the way “core” AI elements are identified.

Additionally, it is well know that consensus does not represent a scientific argument. The fact that different methodologies lead to identify certain scientific papers, patents and software packages as being AI-related does not mean that other developments are not related to AI. It only means that those advancements seem to be identified as being AI-related, no matter the approach used, and that, very likely, these AI-developments are mostly related to the past and to “more established” AI. Newer developments and experimentations can only seldom (if at all) be captured through consensa-based approaches, and future work will try to design ways to identify the different technological trajectories that may be characterising recent AI developments and experimentations.

Alternative approaches, as well as sensitivity analysis on the extent to which different approaches lead to different mapping and measurement outcomes, will be carried out as a next step. Also, methods will need

[4] Accessing data related to proprietary software at present remains unfeasible.

[5] See Dosi (1982) for a discussion about technological paradigms and technological trajectories.

to be developed to quantify the success and error rate of any individual technique in classifying AI vs non-AI, thus allowing the relative efficiency of techniques to be compared.

1.3. Measuring AI developments: A burgeoning field

In 2018, the OECD initiated work aimed at identifying and measuring Artificial Intelligence-related developments building on data on scientific publications (using Elsevier's Scopus® database), on open-source software (OSS) and on patent filings. The aim was to inform the policy discussion on AI and, more generally, to pave the way for the analysis of frontier technologies in support of evidence-based policy making. A preliminary version of the paper was presented in October 2018 and discussed at several OECD Working Parties and with key partners from Intellectual Property Offices member to the OECD-led Intellectual Property Statistics Task Force[6]. In parallel to the OECD efforts, in 2018 and 2019, several institutions and research groups proposed alternative approaches to measuring AI using data on scientific publications and/or on patents. In what follows, we propose an overview of these contributions, highlighting key features and possible differences with the OECD work.

In July 2018, the China Institute for Science and Technology Policy at Tsinghua University (CISTP) published the "2018 China AI Development Report" (CISTP, 2018). This report analyses the development of AI using scientific articles, patents and "talents"[7] in AI. Scientific papers are selected using Clarivate Analytics' Web of Science data, with a list of keywords provided by experts in the field. These same keywords are used to identify AI-related patent families using the Derwent World Patent Index™ (DWPI) database, complemented by additional keywords derived from Derwent Manual Codes. Differently from the present work, CISTP's study overlooks the technological classes allocated to patent documents during the examination period. Moreover, considering that data compiled in the report belong to private sources and that the complete list of keywords on which the study replies is not made available to the broader public, the analysis cannot be replicated or extended to cover different time periods or geographical domains.

In December 2018, Stanford University released the "AI Index Annual Report 2018" (Shonam et al, 2018). Among other indicators, the report provides measures based on scientific publications' data, using Scopus®, open-access archive services (ArXiv), as well as conference papers of the Association for the Advancement of Artificial Intelligence (AAAI); AI and Machine Learning (ML) courses enrolment data; participation in AI conference; and patent data. Stanford's team relied on a limited list of keywords[8] to extract AI-related articles from Scopus®, and used several complementary approaches, based on classifications and on keywords. They analysed data from a patent searching service (amplified.ai) to identify AI-related patents.[9] Shonam et al (2018) patent statistics are presented according to DocDB[10]

[6] The OECD-led Intellectual Property Statistics Task Force (IP Task Force) aims to improve IP data availability and quality and to foster methodological work related to IP rights. Projects and activities are conducted in close co-operation with representatives from a number of OECD IP offices and statistical institutions, as well international IP offices and Organisations, including the European Patent Office (EPO), the European Union Intellectual Property Office (EU IPO) and the World Intellectual property Organization (WIPO).

[7] AI talents are defined as researchers possessed of creative research ability and technical expertise in their research area and active in AI research with innovative outcomes (CISTP 2018).

[8] *Artificial Intelligence* and *Computer science* were the two keywords used to identify AI publications in Scopus®.

[9] See http://aiindex.org/2018/patent-report.html

[10] EPO's master documentation database (DocDB) identifies families of patents sharing an identical technological content (patents with the exact same priorities) : https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families/docdb.html

patent families identified by the European Patent Office (EPO). In synthesis, the approach pursued by the Stanford team differs from the OECD one in terms of both the type of data sources exploited for the purpose, as well as the breadth of AI-related keywords considered.

More or less at the same time as Stanford, the publishing company Elsevier published a report on Artificial Intelligence (Elsevier, 2018). Elsevier used supervised machine learning to mine and extract keywords from several bodies of text. Among them, the text and structure of representative books, using their Scopus® database; the syllabi of massive open online courses (MOOCs); patents; and news items. Elsevier further relied on Natural Language Processing (NLP) techniques to reduce the list of keywords thus obtained, and get a list of 20 000 concepts that were manually reviewed. All this process led to the identification of about 800 unique keywords, which were then used to identify AI scholarly publications.

The end of the 2018 was a very busy period in the AI measurement space, with the Joint Research Centre of the European Commission (EC-JRC) also publishing a flagship report proposing a European view of AI (see Craglia et al., 2018). This work relies on patents and scientific publications' data, which were analysed following the so-called Techno-Economic Segment (TES) approach[11] to identify AI players using a list of keywords. Besides offering a general overview of the approach pursued, the report unfortunately does not contain sufficient information about the scope of patents or of the publications data used and the way the analysis has been implemented, to be able to replicate or expand it.

Early in 2019, the World Intellectual Property Organization (WIPO) also produced a technology trends' report dedicated to AI (WIPO, 2019). The report presented key findings and recent trends in innovation in AI, using the patent data and other information sources, such as scientific publications, litigation records, and firm acquisition activities. The analysis was refined on the basis of expert advice from AI leading companies.

In WIPO's report, AI is subdivided into three main dimensions: techniques, functional applications and application fields. The patent-based statistics used for the analysis are based on data from a commercial provider, Questel[12], and are consolidated into patent families using the "PatFam" definition provided by Questel (a patent family concept close to the notion of equivalent families in Martinez 2011). In the analysis, no geographic restrictions are applied and, differently from the present study, WIPO also includes singletons, i.e. unique patent documents filed at any patent office, in the study). The study relies on a threefold patent search strategy: search for purposely identified Cooperative Patent Classes (CPC) in patents; search of AI-specific keywords on full text data; and mix of the two, i.e. search for a list of CPC or International Patent Classification (IPC) classes containing certain AI-related keywords. The outcomes of these searches are then manually curated to remove false positive AI patents. With respect to scientific publications, articles are extracted from Elsevier's Scopus® database using the previously identified sets of keywords augmented with additional keywords contained in the CPC and IPC classes formerly selected. Country-based measures are reported according to the location of the IP office at which the patent was filed, while scientific publications statistics refer to the affiliation country of the author(s).

The three sections that follow outline the methodology developed by the OECD to identify AI developments using scientific publications, open source software data and patents. The last section of the three presents the outcomes of the consultation had with experts from IP offices of the OECD-led IP Statistics Task Force, including a proposed search strategy for the identification of AI-related patent, the list of keywords to be used for the purpose, as well as the latest trends in AI development.

[11] The Techno-Economics Segment (TES) analytical approach aims to offer a timely representation of an integrated and dynamic technological domains not captured by official statistics or standard classifications. See https://ec.europa.eu/jrc/en/publication/ai-techno-economic-segment-analysis for more details.

[12] Questel is a private provider of IP, science and business data (see https://www.questel.com/)

2. Finding AI-related scientific documents

Scientific publications have long been used to proxy the outcome of research efforts and of advancements in science[13]. Despite their limitations as an indicator of research output – as not all research outcomes and science are disclosed or described in a written scientific piece nor are peer-reviewed -, scientific publications nevertheless provide extremely valuable and reliable information about advancements occurring in all fields of science and technology.

2.1. AI-related scientific documents in Scopus® (bibliometric database)

The bibliometric analysis that follows is based on data from Elsevier's Scopus®, the largest abstract and citation database of peer-reviewed literature, which includes scientific journals, books and conference proceedings. Scopus® is the most comprehensive bibliometric database for 1996 onwards. Especially in the case of fast-developing fields, conference proceedings help get a good sense of latest developments, as conference papers and presentations take less time to be issued and become public than papers published in peer-review journals. Also, in the case of "hard sciences", participation in key conferences is subject to highly competitive selection processes as conference proceedings often count among the publications considered for career purposes (e.g. academic tenure).

2.1.1. The search strategy

Scopus® covers a number of different subject areas, which are denoted by an "All Science Journals Classification" (ASJC) name and code. The ASJC classifies scientific publications helping readers find publications in an area of interest. Among ASJC codes, an AI-related tag groups the journals and conference proceedings related to Artificial Intelligence.

As a first step, only the ASJC AI-tagged journals were considered. This entailed that if, for example, someone wrote an article on AI and it was published in *Nature* (which is not tagged as being AI by the ASJC), the article would not enter the count of AI-tagged documents.

This simple approach helped establish a sort of a lower-bound estimate of AI-related documents, with the caveat of assuming that all articles published in the ASJC AI-tagged journals actually relate to AI. The ASJC AI collection of journals and conference proceedings was further exploited for the compilation of a list of AI-related keywords, to be used to search on abstracts of all the documents in Scopus®. To this end, the approach followed consisted in:

- considering the keywords listed in ASJC AI subject documents;

[13] See OECD and SCImago Research Group (CSIC) (2016) for a discussion.

- performing a co-occurrence analysis based on AI-tagged documents, using the documents' titles and abstracts.[14] This analysis, which aimed to uncover the extent to which more than one keyword appeared in the very same document, was done using VOSviewer, a software tool used to visualise bibliometric networks, allowing users to set thresholds to control the visualisation process. This visualisation and text mining software enabled the creation of a corpus file based on relevance scores, frequency and link strength.[15] Some exploratory statistics helped identify some key distribution-related thresholds as follows: terms were considered insofar as they appeared at least 100 times and belonged to the top 60% in terms of relevance of the terms. This approach allowed for additional words to be included in the initial list.

This work resulted in a list of 193 AI-related keywords, which were then used to search all abstracts in Scopus®. As a first step, documents were considered as being AI-related in so far as they contained at least one of the identified keywords in the abstract. Articles were counted only once irrespective of the number of keywords contained. For instance, an article featuring both "deep learning" and "machine learning" would only be counted once. The types of documents in Scopus® considered for the purpose were: articles, books, business articles, chapters, conference papers, articles in press and reports. The following document types were excluded: abstract reports, book reviews, conference reviews, dissertations, editorials, errata, letters, notes, press reviews, reviews, short surveys and working papers. Overall, the search resulted in identifying more than 2.4 million distinct AI-related documents, for the 1996 to 2016 period, all source types considered.

As a second step, a co-occurrence analysis was performed on the documents selected as a result of the first step, so that only those documents with two or more keywords would be considered as being related to AI. This was done to avoid including among AI documents those articles whose titles or abstracts try to convey the idea of contributing to AI developments - given the popularity of the topic - but actually only vaguely (if at all) relate to it. Such a conservative approach aimed at avoiding overestimating the phenomenon and at reducing type 1 errors, whereby articles not related to AI may be wrongly considered as such. This led to a reduced list of documents (in what follows referred to as the "AI-193 list") featuring only one third of the initial sample for the 1996-2016 period, i.e. about 720 thousand documents.

To enhance the accuracy of the keyword exercise, the AI-193 list was brought to the attention of AI experts from academia and business, who also helped identifying those terms that belonged to past AI developments. Despite the encouraging result of this first robustness test, a more extensive validation will be undertaken in the near future, to ascertain that the list is not missing certain keywords or that terms that not relevant have been included.

Future work will further aim at minimising type 2 errors, i.e. avoiding that relevant AI-related keywords are missed out. This will be done by e.g. scanning 2017-18 articles in leading newspapers and magazines (e.g. *Financial Times, Economist*) in search for keywords contained in AI-related articles. Newspapers and magazines will be identified on the basis of their being "authoritative" and able to influence readers and leaders around the world, without being considered peer-reviewed outlets.

[14] In the case of scientific publications, titles and abstracts tend to be as suggestive as possible as the actual content of an article, and are purposely crafted to attract readers. Performing a similar analysis on full articles' text may only marginally improve the exercise, while representing a computationally demanding task that may likely create a lot of noise.

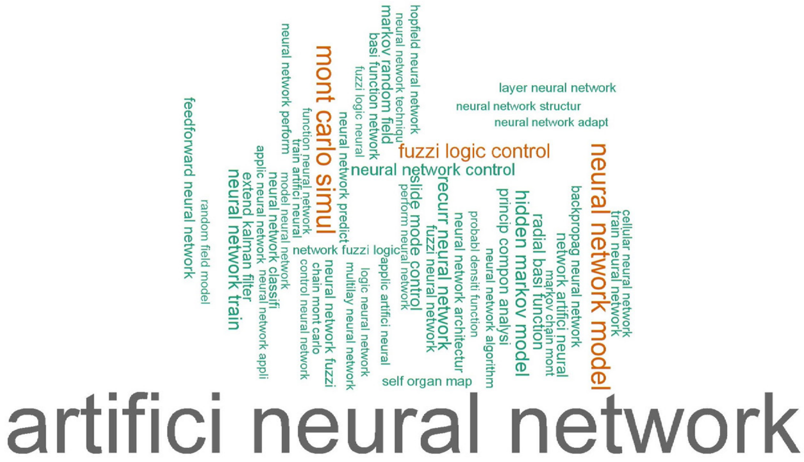
[15] See the van Eck and Waltman (2018) for more on relevance scores and link strength. Terms with high relevance scores generally represent specific topics contained in the text data, whereas when terms feature low relevance score it means they are of a more general nature and are not representative of any specific topic. By excluding low relevance terms, general terms get filtered out and the analysis can focus on more specific and informative terms. The strength of a link is defined as the number of links an item has with other items.

2.1.2. First keywords-related statistics

Figure 2.1. and Figure 2.2 show some word clouds displaying the top 50 trigram, i.e. bundles of three words, which correspond to the top 50 word combinations - based on frequency counts -, found in the documents' abstracts. Data relate to the years 1996 and 2016, respectively.

Figure 2.1. AI-related keywords word cloud, top 50 trigram word combinations based on frequency, 1996

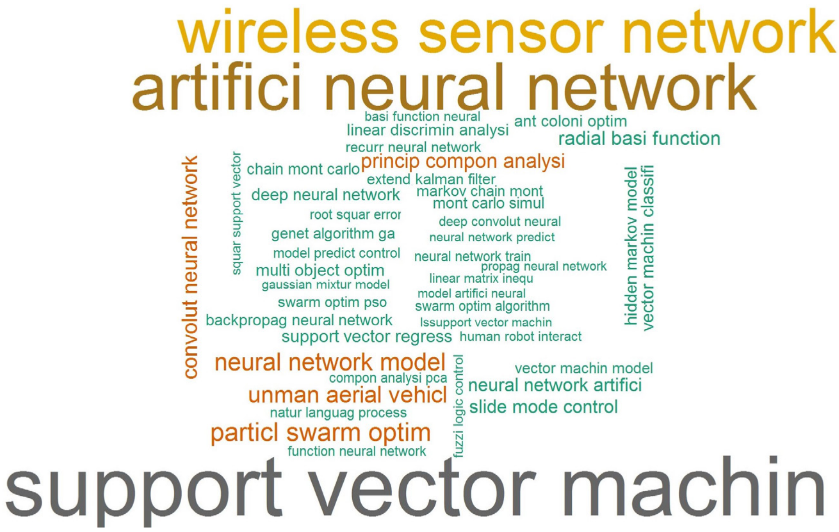
Based on the abstracts of AI_193 documents



Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

Figure 2.2. AI-related keywords word cloud, top 50 trigram word combinations based on frequency, 2016

Based on the abstracts of AI_193 documents

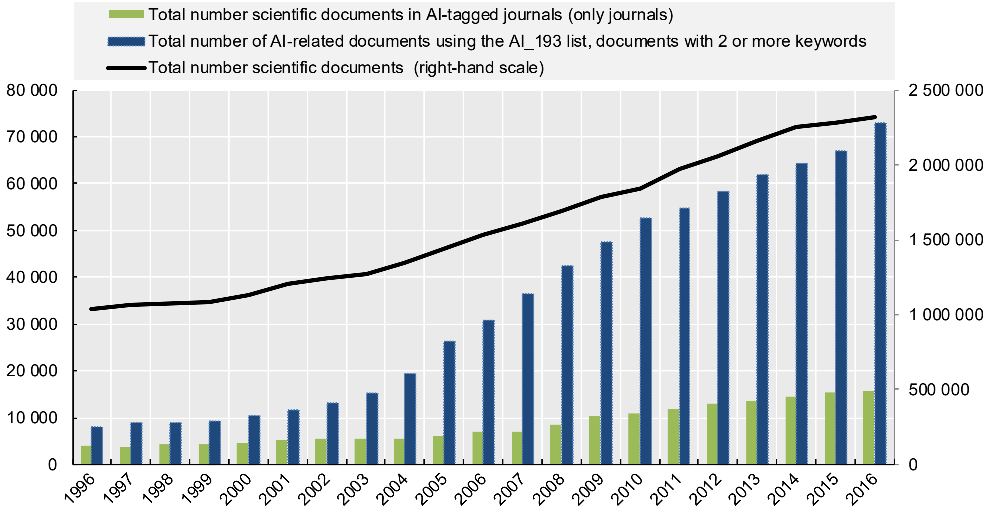


Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

By comparing the word clouds related to the 1996 and the 2016, one can see how much the focus of AI developments changed in the two decades considered. While "Artificial" and "Neural Networks" continued to represent areas where numerous scientific contributions occurred, although to a relatively lesser extent compared to 1996, in 2016 one can observe "machin" and "support vector" terms taking central stage.

Figure 2.3 conversely shows the yearly distribution of AI-tagged ASJC documents and of AI-193 list documents, and compares it with the growth in the total number of scientific documents in Scopus®. As can be seen, using the AI-193 list of keywords allows identifying a greater number of AI related documents, as compared to counting the AI ASJC documents only. Over the 1996-2016 period, AI-193 documents experienced an average annual growth rate of 12% against only 4% on average across of all scientific documents in Scopus® (7% for the documents in AI-tagged journals).

Figure 2.3. Total number of scientific documents, documents in AI-tagged journals (ASJC) and AI-related documents using the AI_193 keyword list, 1996-2016



Note: AI-tagged journals only includes documents in journals. There is a possible year misclassification issue for conference proceedings in Scopus® which requires further investigation

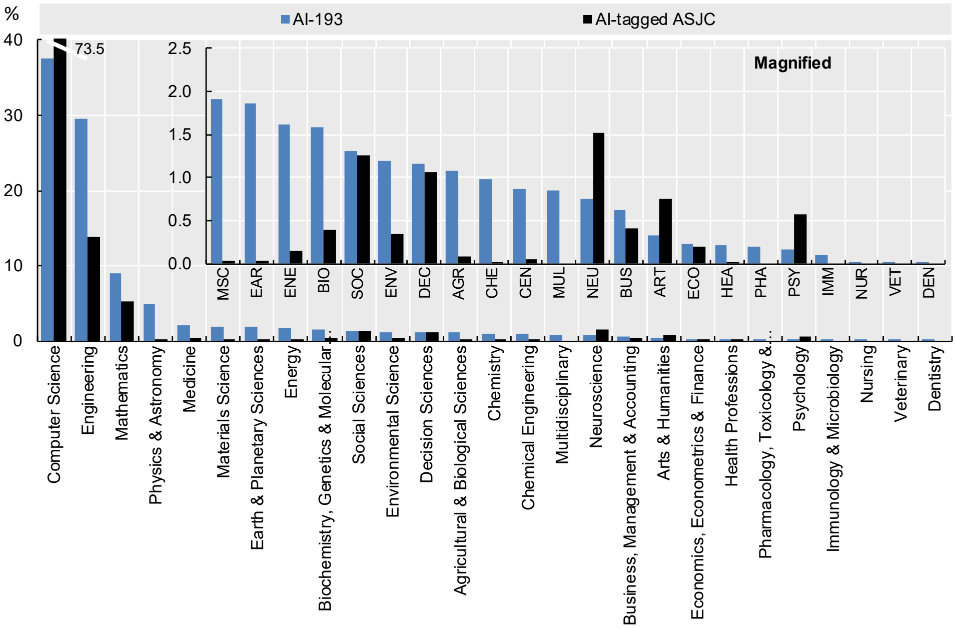
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

Figure 2.4 further shows and compares the ASJC fields to which the AI-193 and the AI-tagged ASJC documents belong, with shares based on fractional counts. This means that if e.g. a document appears in both *Computer Science* and *Engineering* classified journals, an equal weight of a half will be assigned to both. When the AI-tagged ASJC classification is considered, almost three quarters of all AI-documents belong to *Computer Science*. The AI-193 classification conversely spreads the documents over more fields and brings the number of publications in *Computer Science* to a bit over one third. For AI-193, more than a quarter of all documents are in *Engineering* outlets and close to 10% in *Mathematics*. It is also interesting to note that AI-advancements disclosed in scientific publications occur in many other fields, including *Materials Science*, *Medicine* or even *Chemistry*.

This first evidence argues in favour of approaches, like the present one, aimed at capturing not only developments of AI technologies per se, but also AI-related developments and applications occurring in other scientific domains.

Figure 2.4. Scientific fields for AI-tagged and 'AI-193 list' scientific documents, 1996-2016

Shares of the "AI" (AI-193) documents by ASJC field, fractional counts



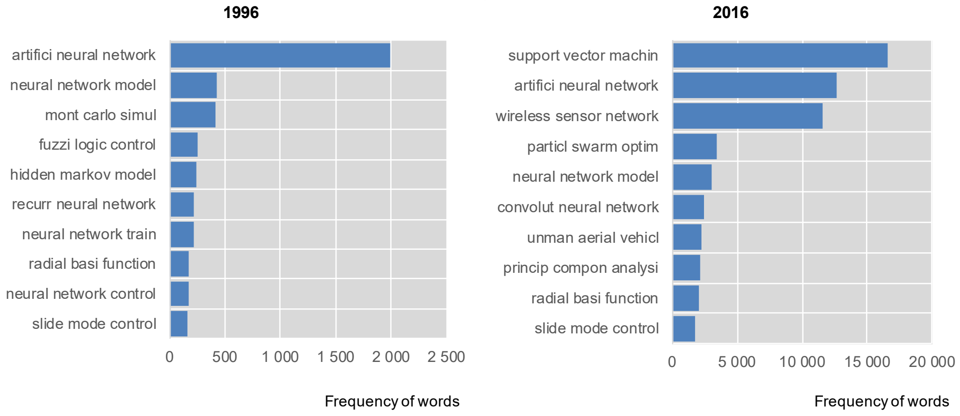
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

Figure 2.5 illustrates the top frequency counts for 2016, as compared to those observed in 1996. In the latest year observed, key AI-related keywords occur with higher frequencies relative to the 1996, and mirror the expansion of this technological paradigm. Also, by comparing 1996 and 2016 keyword frequencies it is possible to see how much AI developments' focus changed, and that in 2016 several key areas seem to be developing to a similar extent, as compared to the very narrow focus of the 1996.

Interestingly, four combinations of keywords, namely "artificial neural network", "neural network model", "radial basis function" and "neural network train", appear among the top ten AI keywords in both 1996 and 2016. Conversely, most of the terms appearing more frequently in 2016 do not appear in the 1996 list - principal component analysis being the only exception. This suggests that new developments occurred over time. Also, it can be observed that all the 1996 top terms still appear in 2016 but rank relatively lower in terms of frequency, thus signalling that they represent relatively more mature areas.

Figure 2.5. Top 10"AI" keyword combinations based on frequency, 1996 and 2016

Based on the abstracts of AI-193 documents



Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

When looking at new technologies or new technological paradigms, as is the case of Artificial Intelligence, it is important not only to shed light on the growth occurred in the different domains contributing to AI but also on the accelerated development of some of them. This helps identify technological trajectories and may inform about the possible developments that may occur in the (near) future, as recent-in-time accelerations allow detecting domains that are likely to continue experiencing sustained developments in future years.

2.1.3. Detecting emerging AI-related technologies in science

The “DETECTS” text mining approach allows the identification of technologies whose development increases sharply (i.e. “bursts”), compared to previous levels and to the development of other technologies. It further helps mappings the time it takes for technological trajectories to unfold (see Dernis et al., 2016).

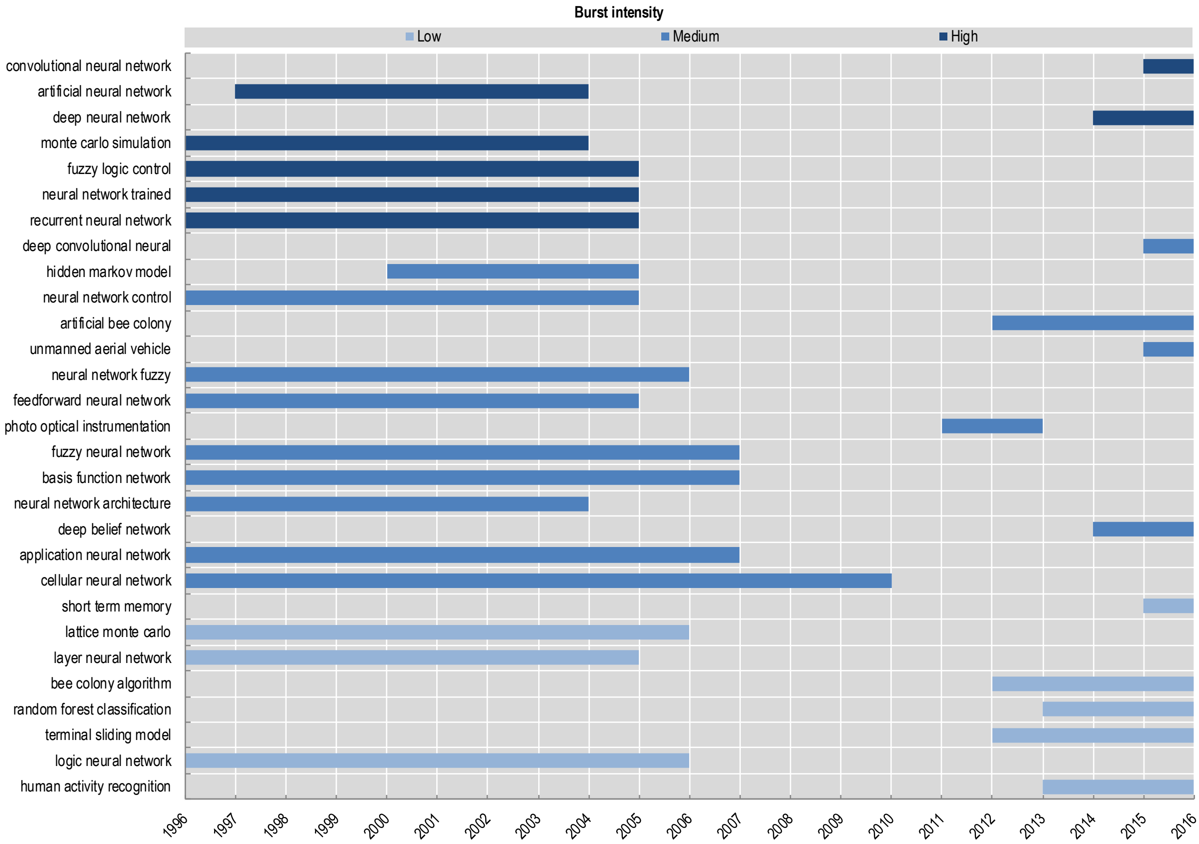
The DETECTS approach was applied here on information contained in the abstracts of AI-193 list of documents. A technology field is said to burst –or to accelerate- when there is a substantial increase in the frequency of developments (in this case, publication of scientific articles) of the technology observed. Accelerations are monitored in relative terms, i.e. compared to past development patterns in the technology and relative to the pace of development of other technologies within AI itself, in the present case.

Monitoring technologies in which accelerations occur is important for policy making, as developments tend to persist in these areas, over the short and medium term. Furthermore, information contained in the documents about the technologies themselves and the geographical location of authors[16] or inventors enables the identification of economies leading in these fields, and can help shed light on the generation of new fields arising from the cross-fertilisation of different technologies.

[16] This approach can be applied on any dataset, e.g. patents, software, for which dates of appearance or frequencies are available.

Figure 2.6. Burst analysis based on "AI-193" documents, 1996-2016

Based on the abstracts of AI-193 documents



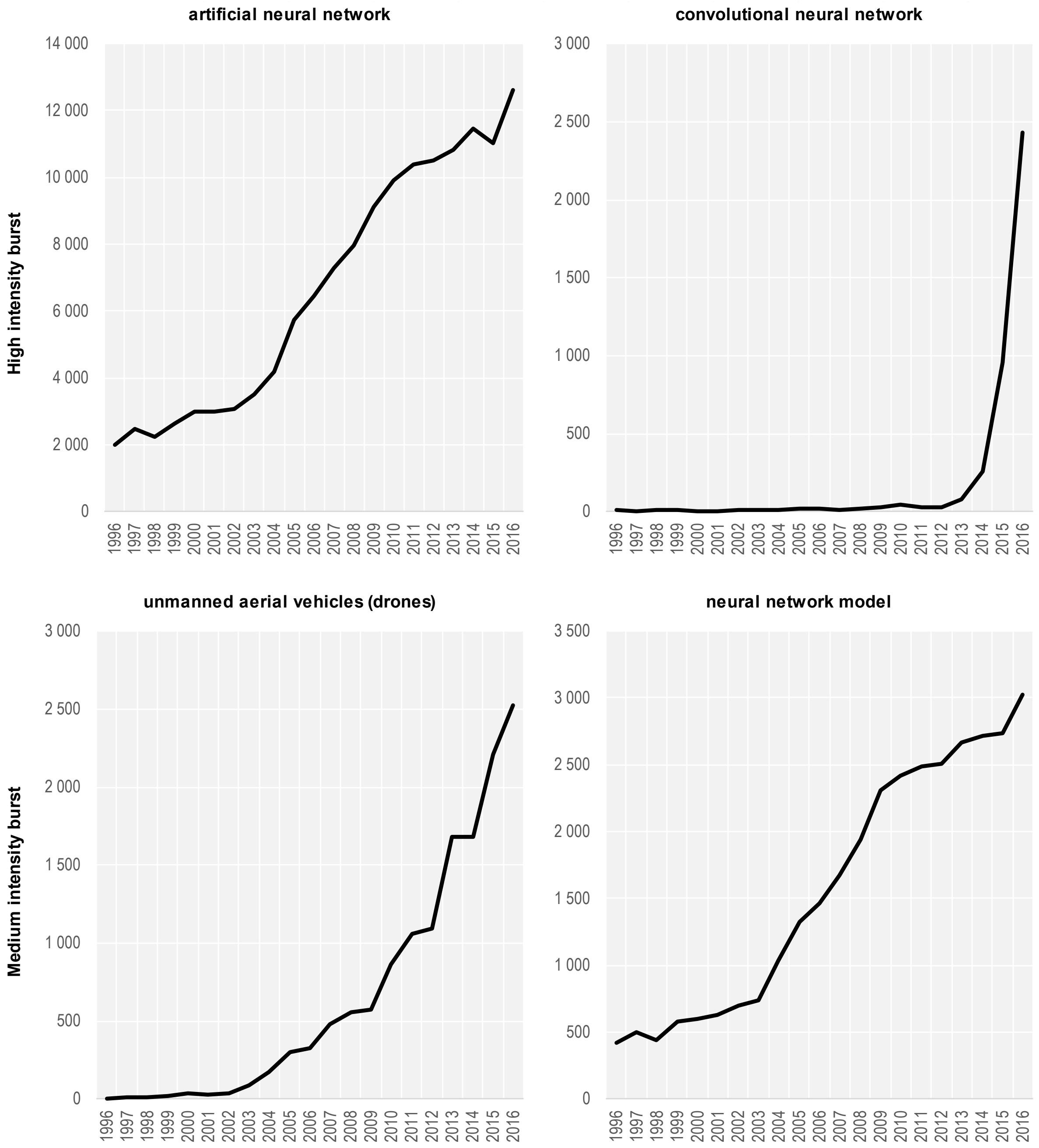
Note: The period 2007-12 does see a number of accelerations, but they are not displayed since they were of less intensity than those observed at the beginning or the end of the period considered.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

The results of the burst analysis depicted in Figure 2.6 allow to uncover the AI subfields that developed at an accelerated pace, i.e. that "burst", and the number of years during which such sustained developments occurred, within the two decades considered. The darker the like, the stronger the acceleration of the subfield. The "open bursts", i.e. the lines that can be observed on the right hand side of Figure 2.6 show the subfields that at the end of the period, i.e. in 2016, appeared to be developing at an accelerated pace. These are likely to have continued bursting over the years that followed the 2016, but data availability constraints at present do not allow checking that this has actually been the case.

To clarify the burst concept, Figure 2.7 graphically shows what bursts entail, by means of displaying the frequencies of selected subfields that experienced an acceleration during the period considered. As can be seen, in some cases the accelerated development is so marked that lines almost get vertical (e.g. convolutional neural network).

Figure 2.7. Frequency charts for four bursts, 1996-2016

Based on the frequency counts of the term in the abstracts of AI_193 documents

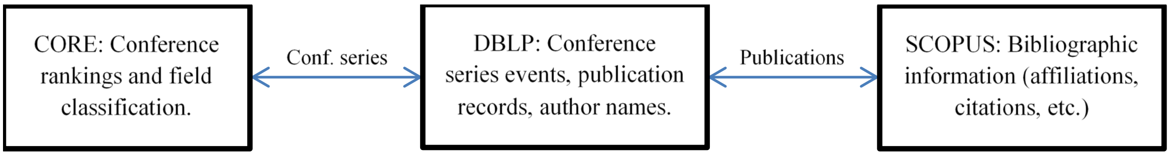


Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

2.2. AI conference proceedings data

An alternative approach to capture the body of scientific literature relevant for AI is to identify publications from conference proceedings in conferences focused on AI. Conferences are often the main loci where cutting-edge findings are presented. In computer science, publications in conference proceedings have also remained the outlet of highest importance for researchers, constituting often the ultimate research output of a research project. This implies that: 1) conferences are of primary importance to trace the evolution of AI and computer science in general; b) conference proceedings in computer science receive more attention and they are covered in bibliographic data, making them better observable. Moreover, conference proceedings constitute a timely snapshot of the characteristics, origin and composition of different scientific communities.

Figure 2.8. Sources of AI conference proceedings data



CORE: Conference rankings and field classification.

Conf. series

DBLP: Conference series events, publication records, author names.

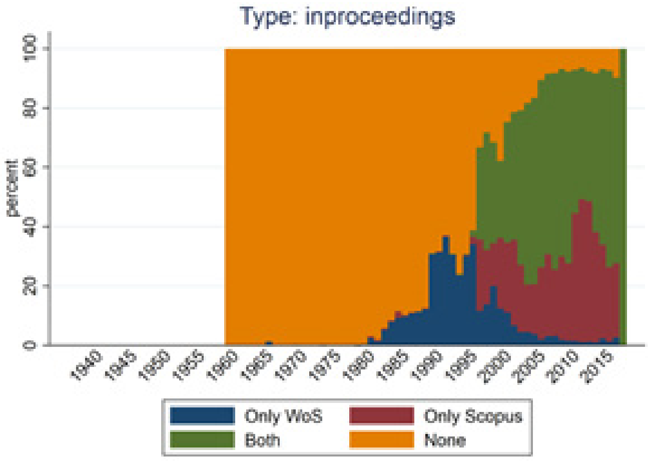
Publications

SCOPUS: Bibliographic information (affiliations, citations, etc.)

Source: Max Planck Institute for Innovation and Competition, 2018.

A dataset of conference proceedings was constructed with information on conferences quality and bibliographic information, combining different sources (Figure 2.8). Conference series and conference events main information is obtained from DBLP (Digital Bibliographic Library Browser)[17]: a bibliographic database specialised in Computer Science. The analysis relies on rankings provided by the Computing Research and Education Association of Australasia (CORE)[18] to rank conferences according to their quality. CORE provides expert based assessments of all major conferences in the computing disciplines, with information on their subfield of research. The conference ranking has been further validated by interviews with experts. Finally, conference proceedings from DBLP were merged with additional bibliographic information (affiliations and citations) from Scopus®.

Figure 2.9. Coverage of DBLP publications in Scopus® and Web of Science



[17] https://dblp.org/

[18] http://www.core.edu.au/index.php/

Source: Max Planck Institute for Innovation and Competition, 2018.

The database was restricted to conferences classified as "Artificial Intelligence and Image Processing". The resulting data contain information for 262 conference series in AI. Conference series are ranked into five categories:

1. *A - flagship conference, a leading venue in a discipline area;
2. A - excellent conference, and highly respected in a discipline area;
3. B - good conference, and well regarded in a discipline area;
4. C - other ranked conference venues that meet minimum standards;
5. Other: regional and minor conferences.

Notably, out of 64 of top-ranked conferences (*A or A) in CORE, full information (from DBLP and Scopus®) was obtained for 59, for most years (Table 2.1).

Table 2.1. Top ranked conference series

Acronym	Title	CORE Rank	Expert opinion
UAI	Conference in Uncertainty in Artificial Intelligence	*A	top
ICML	International Conference on Machine Learning	*A	top
COLT	Annual Conference on Computational Learning Theory	*A	top
NIPS	Advances in Neural Information Processing Systems	*A	top
SIGKDD	ACM International Conference on Knowledge Discovery and Data Mining	*A	top
WWW	International World Wide Web Conference	*A	important
SIGIR	ACM International Conference on Research and Development in Information Retrieval	*A	important
ACL	Association of Computational Linguistics	*A	important
WSDM	ACM International Conference on Web Search and Data Mining	*A	important
ICCV	IEEE International Conference on Computer Vision	*A	important
CVPR	IEEE Conference on Computer Vision and Pattern Recognition	*A	important
ICDM	IEEE International Conference on Data Mining	*A	important
KR	International Conference on the Principles of Knowledge Representation and Reasoning	*A	
EC	ACM Conference on Economics and Computation	*A	
SIGGRAPH	ACM SIG International Conference on Computer Graphics and Interactive Techniques	*A	
FOGA	Foundations of Genetic Algorithms	*A	
IJCAI	International Joint Conference on Artificial Intelligence	*A	
IJCAR	International Joint Conference on Automated Reasoning	*A	
RSS	Robotics: Systems and Science	*A	
ISMAR	IEEE/ACM International Symposium on Mixed and Augmented Reality	*A	
AAAI	National Conference of the American Association for Artificial Intelligence	*A	
ICAPS	International Conference on Automated Planning and Scheduling	*A	
IEEE InfoVis	IEEE Information Visualization Conference	*A	
AAMAS	International Joint Conference on Autonomous Agents and Multiagent Systems	*A	
AISTATS	International Conference on Artificial Intelligence and Statistics	A	top
ECCV	European Conference on Computer Vision	A	important
EMNLP	Empirical Methods in Natural Language Processing	A	important
SDM	SIAM International Conference on Data Mining	A	important
CIKM	ACM International Conference on Information and Knowledge Management	A	important
ECML	European Conference on Machine Learning	A	important
FSR	International Conference on Field and Service Robotics	A	
VRST	ACM Virtual Reality Software and Technology	A	
ITS	International Conference on Intelligent Tutoring Systems	A	
ECAI	European Conference on Artificial Intelligence	A	
WACV	IEEE Workshop on Applications of Computer Vision	A	
ESWC	Extended Semantic Web Conference	A	
PPSN	Parallel Problem Solving from Nature	A	
VR	IEEE Virtual Reality Conference	A	
IROS	IEEE/RSJ International Conference on Intelligent Robots and Systems	A	
PG	Pacific Conference on Computer Graphics and Applications	A	
NAACL	North American Association for Computational Linguistics	A	
AIED	International Conference on Artificial Intelligence in Education	A	
ICARCV	International Conference on Control, Automation, Robotics and Vision	A	
ICONIP	International Conference on Neural Information Processing	A	
MICCAI	Medical Image Computing and Computer-Assisted Intervention	A	
IJCNN	IEEE International Joint Conference on Neural Networks	A	
Interspeech	Annual Conference of the International Speech Communication Association	A	
ICCAD	IEEE/ACM International Conference on Computer-Aided Design	A	
ISR	International Symposium on Robotics	A	
FUZZ-IEEE	IEEE International Conference on Fuzzy Systems	A	
ALIFE	International Conference on the Simulation and Synthesis of Living Systems	A	
IEEE Alife	IEEE International Symposium on Artificial Life	A	
IEEE VIS	IEEE Visualization	A	

Source: CORE-2018, http://www.core.edu.au/conference-portal/2018-conference-rankings-1.

3. Identifying AI in open source software

3.1. Using machine learning on open-source software (OSS) data to identify AI-related software

As all practical implementations of artificial intelligence rely on software, software becomes an important medium to analyse for the purpose of tracking AI developments and applications. In what follows, the report focuses specifically on open source software hosted on GitHub, an online hosting platform for version control using Git (i.e. for tracking changes to software code). While GitHub is not the only online platform for Git version control, it is by far the largest worldwide, with 24 million registered users in 2017, versus 6 million on Bitbucket, the next largest platform.

Performing analysis on software is important for many reasons, including to have to better understanding of the type of knowledge on which AI-related developments rely, and to monitor the different types of and the directions along which AI developments are occurring. Also, getting a solid grasp of AI software development helps to better identify the fields of application and the speed at which the AI is advancing. Such an understanding would be even deeper if data related to proprietary software were available, but this is not the case. However, qualitative evidence suggests that proprietary software often builds upon and combines open source software components[19]. Hence, using open source software-related data help shed light on technological developments otherwise impossible to apprehend.

GitHub users host projects on *repositories*. Within each repository, they may upload code files, makes changes to them, receive changes from other users, and so on. Additionally, repositories also typically include a *Readme* file, in which users describe the content of the repository.

The identification strategy of AI-related software developments occurring in the open source software world has been as follows:

- Gather publications from a list of key AI conferences, as identified by experts in the field.
- Identify GitHub repositories that cite these conference publications in their Readme files – this will be an indicator that the repository is in some way AI-related (in fact, some of the repositories are the coding implementation of a publication by the same researcher). These will constitute the "core" of AI repositories, which can be unambiguously labelled as AI.
- Use machine learning techniques to identify repositories whose Readmes are similar enough to the "core" to also be labelled as AI repositories.

[19] This is the case, for instance of Google's TensorFlow, an open-source software that is widely used for programming neural networks from Google Translate to Mozilla's speech recognition. TensorFlow is programmed collaboratively on GitHub, receiving over 41,000 commits (modifications) from over 1,600 distinct contributors. It is widely used in industry, with over 110,000 "stars" (a way to bookmark on GitHub, signalling interest) and over 68,000 "forks" (copies of the code for further modification). See https://github.com/tensorflow/tensorflow and https://www.tensorflow.org/about/uses .

3.2. Gathering data about AI publications to identify "core" AI-related software repositories

The AI conferences were identified using the DBLP (Digital Bibliographic Library Browser), provided by the University of Trier, Germany. The list as ranked by CORE (Computing Research and Education Association of Australasia) and AI experts included conferences that were too general to accurately identify "core" AI publication (see section 2.2). For instance, the International World Wide Web Conference, which includes many publications unrelated to AI, and the SIGKDD which focuses partly on data mining and hence encompasses more than just AI. For this reason, the conferences were restricted to those ranked as "top" by experts, namely:

- The Conference in Uncertainty in Artificial Intelligence (UAI);
- The International Conference on Machine Learning (ICML);
- The Annual Conference on Computational Learning Theory (COLT);
- Advances in Neural Information Processing Systems (NIPS);
- International Conference on Artificial Intelligence and Statistics (AISTATS).

This allowed obtaining a list of around 7 thousand AI-related publications, to be used in the "core" AI identification exercise. The latter was performed by searching for exact string matches (including upper case letters) of publication titles within the Readme files, using GitHub's search API. Using exact string matching increases the likelihood that the search would return actual citations of a publication rather than just generic terms or phrases.

Additional refinements were also performed to ensure as much as possible that the search returned repositories which both included actual software and a reference to an actual paper, given that:

- Some repositories seemed to include only a bibliography of AI publications, but did not include actual software. To avoid overestimating AI developments, repositories referencing more than 12 publications were removed from the initial sample[20].
- Additionally, some publication titles featured common phrases that seemed to match many repositories[21] without being references to the papers were removed[22]. This reduced the number of titles kept to about 1.8 thousand and the number of GitHub "core" repositories obtained to about 3 thousands.

Given the classification technique used, diminishing returns on the number of papers and repositories identified apply, and do so very quickly. In other words, using the text from a handful (dozens) of *Readmes*, assuming they are representative, will yield results almost as good as with thousands. Conversely, while

[20] The actual threshold was chosen in an ad-hoc manner based on viewing samples of repositories excluded, such that these appeared to be just bibliographies (rather than actual description of a project).

[21] Titles such as "Management of Uncertainty" or "Direct and Indirect Effects", but also "Bayesian PCA" which is an AI-related technique but is a keyword as well as a publication title. Examples of titles which were kept include "Monte-Carlo Planning in Large POMDPs" or "Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks", which are more specific and are thus more likely to refer to actual papers.

[22] These were removed as follows: a score was assigned to each title based on its word length minus a penalisation term, learned from the probability density function of a Gamma distribution[22] based on the number of GitHub repositories it returned in the search. The penalty parameter and the overall score threshold over which to keep the title were learned by tuning the model on a small number (around 20) of titles. This aimed to ensure that no false positives would remain in that set, i.e. that all publication titles included would actually be publication titles and not common phrases. This resulted in, for example, titles made of only two of words featuring in only one repository being kept, whereas titles with more words but many search hits being discarded.

the classifier is by design somewhat robust to random outliers, we show in the sensitivity analysis that it is still influenced by the choice of these documents. This is why it is preferable to have a smaller number of *Readme* files which are unambiguously AI, than a larger number containing non-AI elements.

3.3. Embedding: mapping words to vectors of numbers

Similarity measures between the *Readme* files were then computed by embedding their text in a “metric space”. First, a continuous representation of words was learned[23], which allowed performing mathematical operations such as: "King - Man = Queen". This shows that not only was the similarity between "King" and "Queen" learned, which is useful, but also that their difference lies on a particular dimension, the dimension represented in the word "Man".

This technique has become standard practise in Natural Language Processing (NLP), to the point that performing this same operation can be seen as the first layer of any deep-learning language model. When faced with unlabelled data, a technique known as word embedding is often used, whereby labels are "created" by assuming that words that appear in some relation to one another (close to one another, within the same sentence, within the same document) are indeed related. Popular word embedding techniques are *Word2Vec*, *GloVe*, and *FastText*. These popular techniques set up the problem such that they can learn, directly, the relationship of one word to another[24]. This report used Facebook Research’s *StarSpace* library (see Box 3.1 for more details).

[23] A continuous representation of words allows representing documents as the normalised sum of their words, in the same metric space. This is advantageous over techniques such as a word-count or term frequency–inverse document frequency (tf-idf) vector, as it allows to learn similarities and complementarities in the use of words, and encode that in the embedding itself.

[24] While powerful, it is not always clear how to then turn these word embeddings into document embeddings. One can average the vectors, or take a weighted average of the vectors, which can work relatively well in practise, but there is no clear best approach.

Box 3.1. Applying the StarSpace embedding

To turn word embeddings into document embeddings[25], the *StarSpace* library and technique from Facebook's AI research team, published in 2017 was used. *StarSpace* is uniquely built such that it optimises the embedding of full sentences and documents directly, as normalised sums of their component words. The end result is, like the other techniques, a dictionary of word embeddings[26], which in the case of StarSpace are optimised explicitly to be summed together into "groups", such as sentences or documents, rather than intended to have meaning in-and-of themselves as word representations. The result is a document-embedding procedure that proves to be more accurate in comparing semantic textual similarity between sentences and documents.

In this context, the algorithm works via the following steps:

1. Picking a sentence in a repository;
2. Picking the other sentences in the repository as the "positive labels";
3. Picking a handful of random sentences from other repositories as the "negative labels";

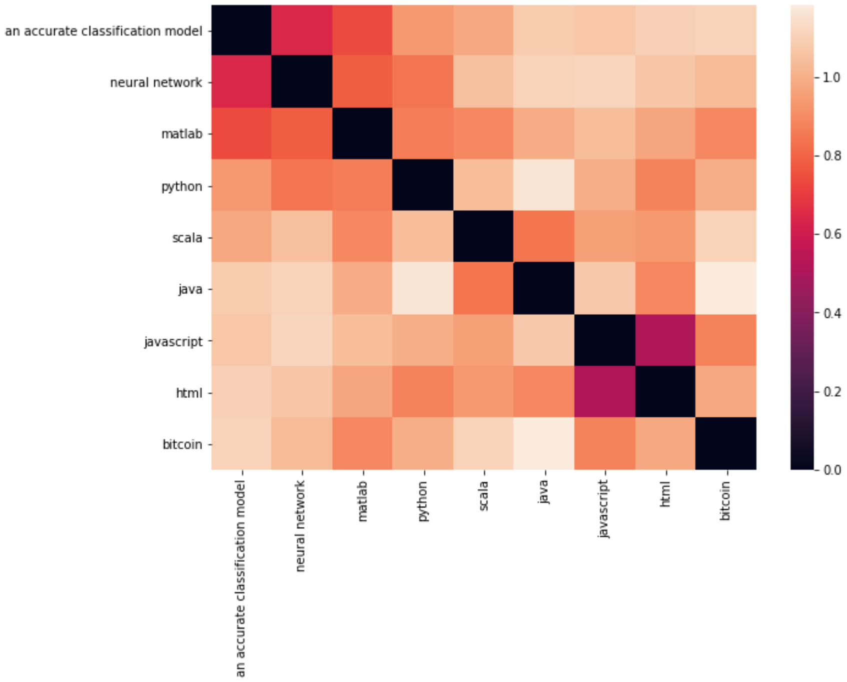
We then learn a dictionary of word embeddings, ***F***, such that the similarity measure is maximised between the chosen sentence and the positive labels, while minimised between the negative labels, when each sentence is represented as:

$$s = \frac{\sum_{i \in a} F_i}{|| \sum_{i \in a} F_i ||}$$

where ***a*** contains the dictionary index of each word in the sentence.

To show the power of this technique, we embed a few test words and a few test sentences and look at the heatmap of their distance in the embedded space, as shown in Figure 3.1.

Figure 3.1. Heatmap showing distance between key terms in the embedded space



Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

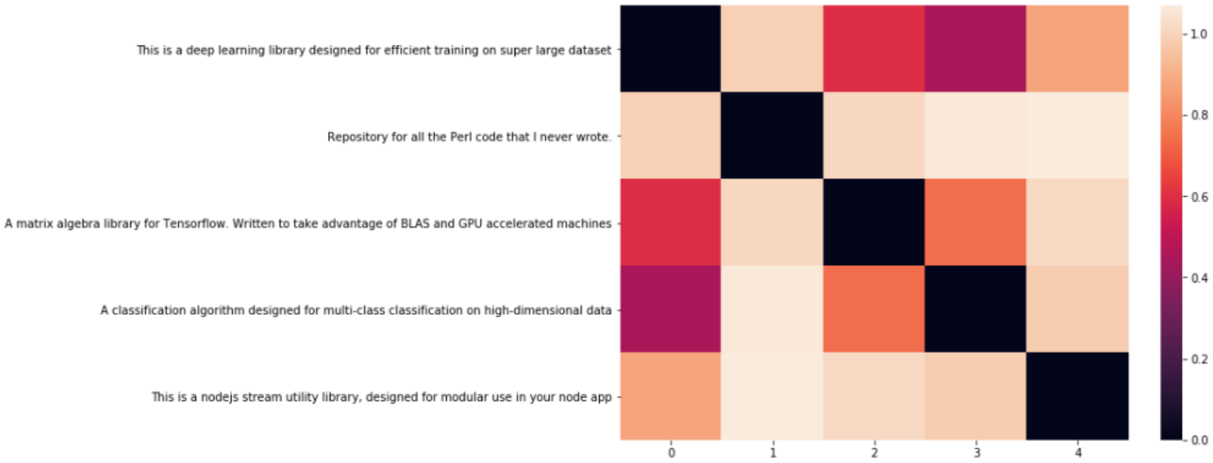
[25] This was done to compute similarities between documents rather than between words.

[26] Dictionaries of word embeddings are fixed-width continuously-valued vectors that represent each word.

Figure 3.1 clearly shows that e.g. "neural network" and an "accurate classification model" are considered very close in the space. Similarly, "javascript" and "html" (programming languages that run together on web pages) appear very close to one another. However, “javascript” and “html” are both very far from “neural networks”. This is exactly what one would expect, as, in the realm of programming, they have relatively little to do with one another (except when one might visualise neural networks on a webpage).

The example in Figure 3.2 shows how complex sentences that share no keywords can still be accurately modelled in their distance to one another. The embedding has learned the relationship of the concepts. This does not occur, instead, when relying on a simple word-count or tf-idf vectors are used to embed documents: in such cases, all of these documents would appear essentially unrelated.

Figure 3.2. Heatmap showing distance between selected sentences in the embedded space



Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

Using the embedding, however, reveals relationships between sentences that share no keywords: software focused on matrix algebra and GPU's (row 3) is shown as very close to software which runs a classification algorithm on high-dimensional data (row 4), which is in turn very close to a deep-learning library (row 1). None of these are particularly close to a "nodejs stream utility" (row 5), a library that is used in building servers to handle web requests, for instance.

3.4. Classification algorithm

The Support Vector Machine (SVM) is a classic machine-learning algorithm that draws a boundary in the geometric space[27] of the problem, classifying all points on one side of the boundary as positive and those on the other side as negative.

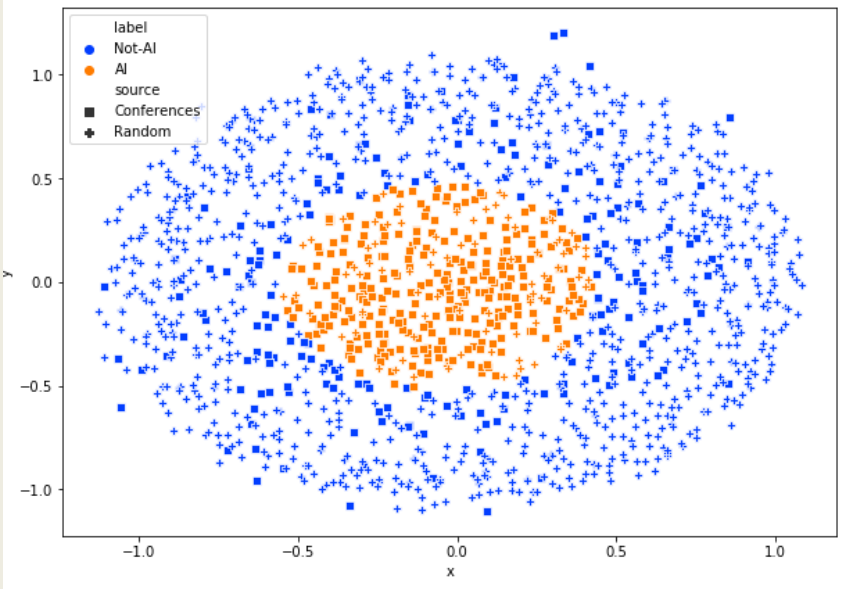
One-class SVM seeks to draw this boundary after having only seen positive points. It can be thought of as drawing a sphere around the positive points it is trained on, and shrinking that sphere as much as possible: trading off between minimising the volume of the sphere and maximising the amount of points it keeps inside of it (correctly classified).

[27] In the present case, the space created by a Gaussian kernel.

Box 3.2. Tuning the AI classifier

Drawing a small random sample from the embedded space around the medoid[28] AI repository, the work of the classifier can be visualised in the following figure. The model has one tuning parameter: ***v***, with which to set an upper bound, in fractional terms, on how many positive labels to leave outside of the decision boundary (incorrectly labelled). In this case, up to 50% of the original readmes were left outside the boundary (the squares with the "Not AI" label in the figure below).

Two dimensional projection of the classification algorithm



Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

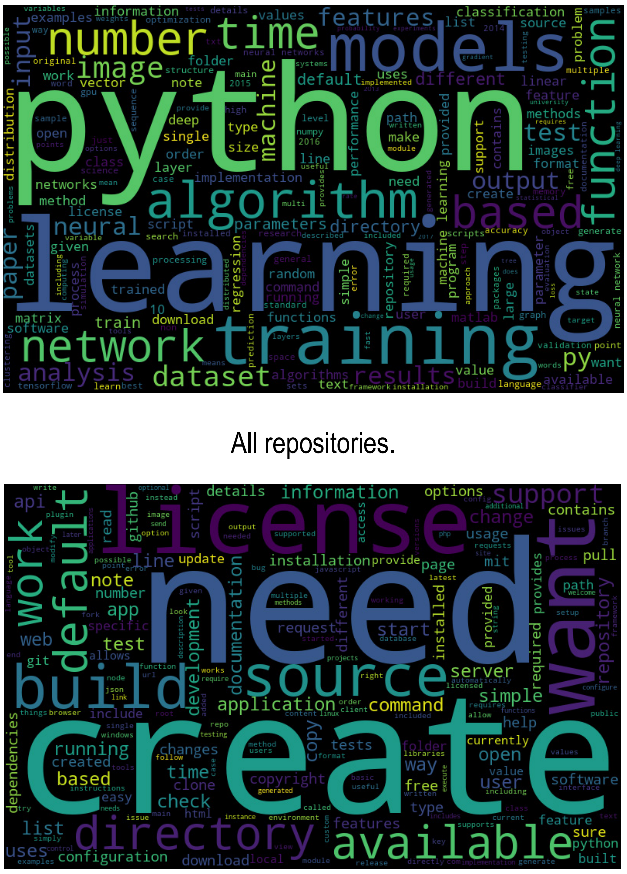
3.5. Results

From a sample of 13 thousand AI conference publications, 13.7% were referenced in about 3 thousand GitHub repository *Readme* files. Using these *Readme* files, out of 2.7 million GitHub repositories, around 11 and half thousands (i.e. 0.4%) were identified as “unambiguously” AI-related – meaning this is likely to be a lower-bound estimate given the conservative tuning choices made for the parameters.

The following word cloud (Figure 3.3) illustrates the difference between the AI repositories and the overall population of GitHub repositories. The main words on the top (which are AI-related) contain words much more related to AI and machine learning than on the bottom word cloud, for instance: “learning”, “algorithm”, and “training”.

[28] The medoid is the data point which has the lowest average distance to other data points in a given group.

Figure 3.3. World clouds with most frequent terms in repositories


AI repositories.

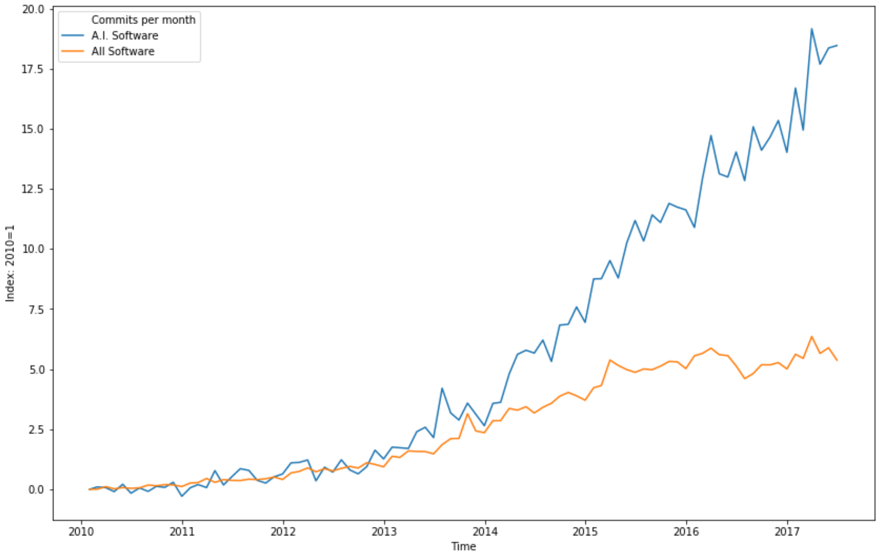
All repositories.

Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

In addition, one can track the rapid growth in AI-related repositories: in 2010, there were 50 active AI GitHub repositories that had gathered 1 350 “commits” (i.e. changes to code) from contributors, making up 0.26% of total commits on GitHub that month. In June 2017, AI software activity had increased to 26 275 commits on 1 533 projects, making up 0.74% of total commits on Github.

As can be seen in Figure 3.4, most of this growth has taken place since 2014. In the subsequent 3 years, A.I. open-source software grew about three times as much as the rest of open-source software.

Figure 3.4. Growth in commits to AI and all repositories, relative to 2010

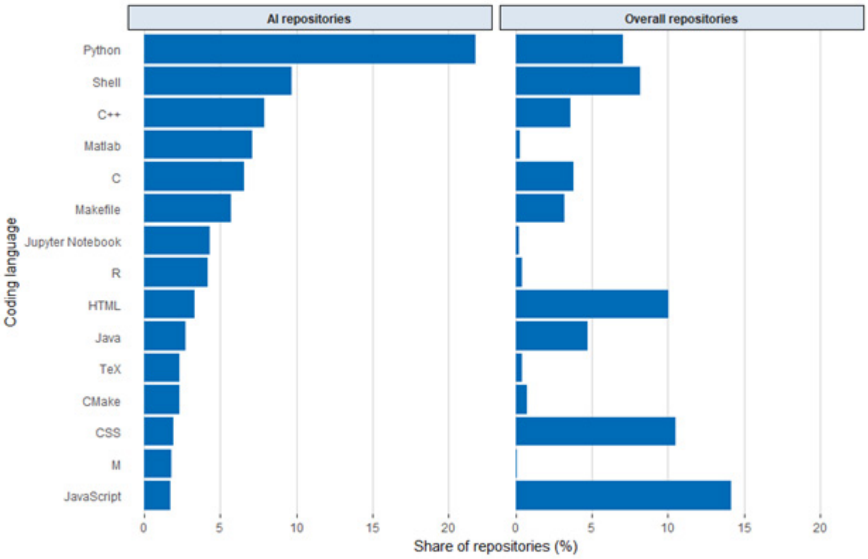


Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

Figure 3.5 displays the top 15 coding languages in AI repositories by share (cumulatively worth over 80% of all AI repositories), and their share within GitHub repositories as a whole. Whereas the top languages for software on GitHub overall are JavaScript, CSS and HTML (used for web development), the top languages for AI repositories are Python (over 20% of AI repositories), followed by Shell and C++, and includes some (such as Matlab, Jupyter Notebook and R) which hardly appear among GitHub repositories as a whole. This suggests that the software tools used by AI practitioners differ significantly from those used by the overall community on GitHub.

Figure 3.5. Top 15 coding languages on GitHub for AI repositories vs all repositories

Top languages by number of repositories as a percentage share of total number of repositories in each group



Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

What topics are covered in Artificial Intelligence in software? Topic modelling[29] is a technique that can generate broad themes covered in AI documents. Both words and documents (here, *Readme* files) are allocated in probability to each topic (with the sum of probabilities equal to 1). The topics generated must then be interpreted; in this case, the top 10 words by topic are displayed, and an overall title is assigned to each topic (Table 3.1). With 10 topics, one can see some of the broad topics covered by AI repositories on GitHub: machine learning (including deep learning) make up an important part, as well as statistics, mathematics and computational methods. The topics also provide insights into a few of the specific fields and techniques that AI is used for: text mining, image recognition, and biology. It important to note that this is not a comprehensive list of topics:

Table 3.1. Top 10 words by topic

Statistics	Simulation	Mathematics	Text mining	Deep learning	Computational methods	Image recognition	Biology	Machine learning courses	Other machine learning techniques
gaussian	simul	search	word	tensorflow	execut	segment	cell	machin learn	cluster
observ	figur	equat	languag	batch	parallel	recognit	gene	graph	label
modul	signal	solver	embed	gpu	simul	face	name	languag	tree
prior	cell	graph	sentenc	gradient	modul	vision	protein	cours	forest
bayesian	respons	numer	corpus	loss	cuda	video	sequenc	topic	svm
likelihood	rang	element	txt	architectur	thread	caff	express	mine	cross valid
popul	concentr	integr	label	deep learn	git	demo	score	toolbox	data set
simul	dynam	simul	task	epoch	memori	opencv	cluster	infer	accuraci
covari	argument	dimension	name	kera	name	box	measur	open sourc	score
sequenc	control	dynam	token	net	gpu	adversari	genom	engin	kernel

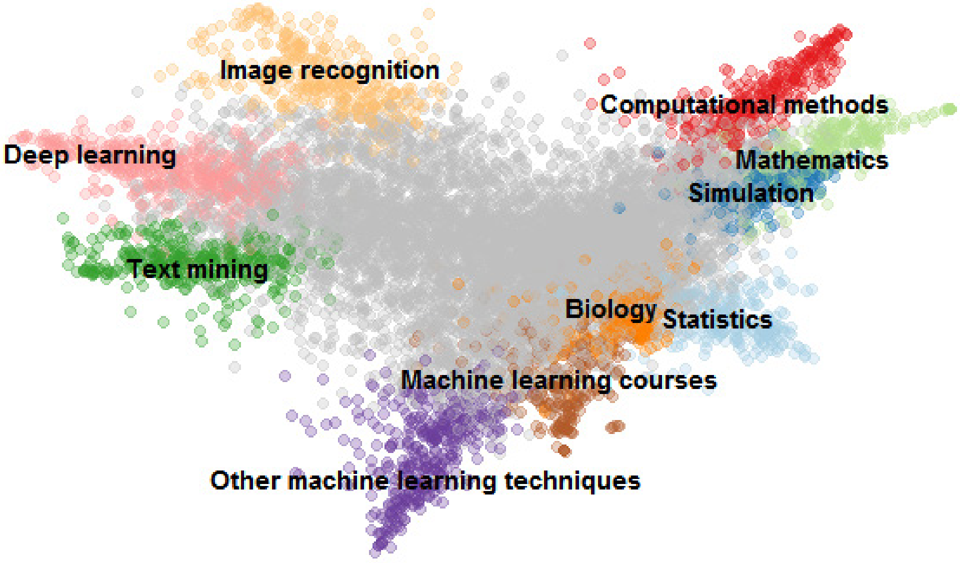
Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

Figure 3.6 provides a visual illustration of the mixed nature of AI repositories in terms of topics. The documents for which the highest topic probability is above 50% were coloured (by highest topic), with the rest in grey. One can see that while there are a number of *Readme* files for which there is relatively high certainty regarding their topic (where one of the 10 topics is above 50% probable), many documents cover a mix of the topics listed. It is important to note that this is not a comprehensive list of topics; 10 were chosen for ease of interpretability and to provide a broad picture, but more would be needed for greater accuracy and comprehensiveness.

[29] Using the Latent Dirichlet Allocation algorithm

Figure 3.6. Two dimensions projection of AI *Readme* files

Coloured by topics for probabilities above 50%



Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

3.6. Sensitivity analysis

As all measurement and analytical endeavours, the results presented in this sections, and their implications, are sensitive to a number of decisions. These include: the choice of embedding procedure (both the technology and the number of dimensions for the word vectors); the choice of the “core” AI documents; and the parameter ***v*** used in the one-class SVM classifier, which affects how many “core” AI documents are left out in the final classification.

A number of sensitivity tests, shown in Annex A, have thus been performed to shed light on the impact that the choice of those parameters may have on the identification and mapping on what is to be considered AI-related and what is not. While thorough, this exploration has been far from exhaustive, and future work will aim at testing higher-dimensional vectors, longer training, and more extensive hyper-parameter tweaking to check whether improvements can be brought to the current measurement endeavour.

4. Technological developments in AI – a patent-based measure

Patents represent a fairly standardised output measure of inventive activities and a detailed source of data (Griliches 1990). Patent data are used in this section to identify and map AI-related inventions and new technological developments embedding some AI components that occur in any technological domain. The present work builds on the experimental definition of AI-related patents proposed by the OECD in 2017. In the STI Scoreboard 2017 (OECD, 2017), AI patents were defined as patents filed in those International Patent Classification (IPC) fields belonging to Information and Communication Technologies (ICT) and related to human interface and cognition and meaning understanding (Inaba and Squicciarini, 2017).

The OECD (2017) experimental definition resulted in a broad measure of AI-related patenting that is refined and narrowed down in the present work. To this end, different approaches were pursued, including using keywords to search patents' abstracts, as well as relying on patents that cite references to AI-related scientific papers.[30]

4.1. Identifying AI-related patents using data contained in patent data

Artificial intelligence is often considered a General Purpose Technology (GPT), able to pervade many technological areas. This is reflected in the different definitions of AI-related inventions proposed in recent studies, some of which narrowly focus on the development of the algorithms at the basis of AI, while others encompass applications occurring in a many domains (e.g. robotics, autonomous vehicles, etc.).

The difficulty of mapping developments in AI is further demonstrated by the fact that most studies restrict their attention to inventions occurring or protected in specific geographical area, mainly the United States or Europe. Doing so they nevertheless end up proposing a curtailed view of AI-technological developments occurring worldwide.

In addition, as AI relies on software coding, it is important to track technological innovations contained in software. However, the patentability of software-based technologies varies across countries and depending on the Intellectual Property Office (IPO) where patent protection is sought. For instance, the US Patent and Trademark Office (USPTO) allows software to be patented as such, whereas software becomes patentable at the European Patent Office (EPO) only in the context of "Computer implemented inventions" (CII).[31].

[30] An alternative approach was also pusued, using data related to AI-related startups from the Crunchbase dataset, and linking these to patent data. However, no obvious pattern emerged from the analysis of the patent portfolio of those firms with respect to AI technologies identification.

[31] Computer implemented inventions are defined as inventions "involving the use of a computer, computer network or other programmable apparatus, where one or more features are realised wholly or partly by means of a computer program". For more details, see the EPO guidelines for examination at https://www.epo.org/law-practice/legal-texts/html/guidelines/e/j.htm

4.1.1. Existing patent taxonomies related to AI

Fujii and Managi (2017) rely on USPTO patent records and identify as AI-related those patents mainly allocated to the IPC code "G06N", i.e. "Computer Systems based on Specific Computational Models". This corresponds to the US Patent Classification (USPC) code 706 ("Data processing, Artificial Intelligence").

In Cockburn, Henderson and Stern (2017), the scope of AI-related patents encompasses in addition to the USPC code 706 the patents allocated to the USPC code 901, i.e. "Robots". They further consider as being AI-related those patents identified performing keyword searches on patent titles. Cockburn et al.'s (2017) rely on USPTO patents only and AI-related patents are then allocated to three distinct subfields, namely Robotics, Learning Systems and Symbol Systems.

The EPO has developed an alternative approach in their report on *Patents and the Fourth Industrial Revolution* (2017). The EPO (2017) considers as AI-related those patents applied with respect to technologies enabling machine understanding. EPO AI-related patents are identified using a number of Cooperative Patent Classification (CPC) codes, including code G06N, and cover different fields of application, including health (e.g. methods for diagnostic purposes) or vehicles (e.g. control for combustion or traffic).

Finally, as mentioned, in the *OECD, Science, Technology and Industry Scoreboard 2017* (OECD, 2017), patents are considered as being AI-related in so far as they belong to the Human interface and Cognition and meaning understanding categories listed in the OECD 2017 taxonomy of ICT technologies (and detailed in Inaba and Squicciarini, 2017), in addition to those in IPC class G06N. To account for AI-related inventions occurring worldwide, the OECD taxonomy relied on data about "IP5 patent families". The latter denote inventions protected in at least two jurisdictions, at least one of which needs being one of the Five IP Office, which are: the EPO, the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the USPTO and the National Intellectual Property Administration of People's Republic of China (CNIPA)[32].

4.1.2. Comparing AI-related patent taxonomies

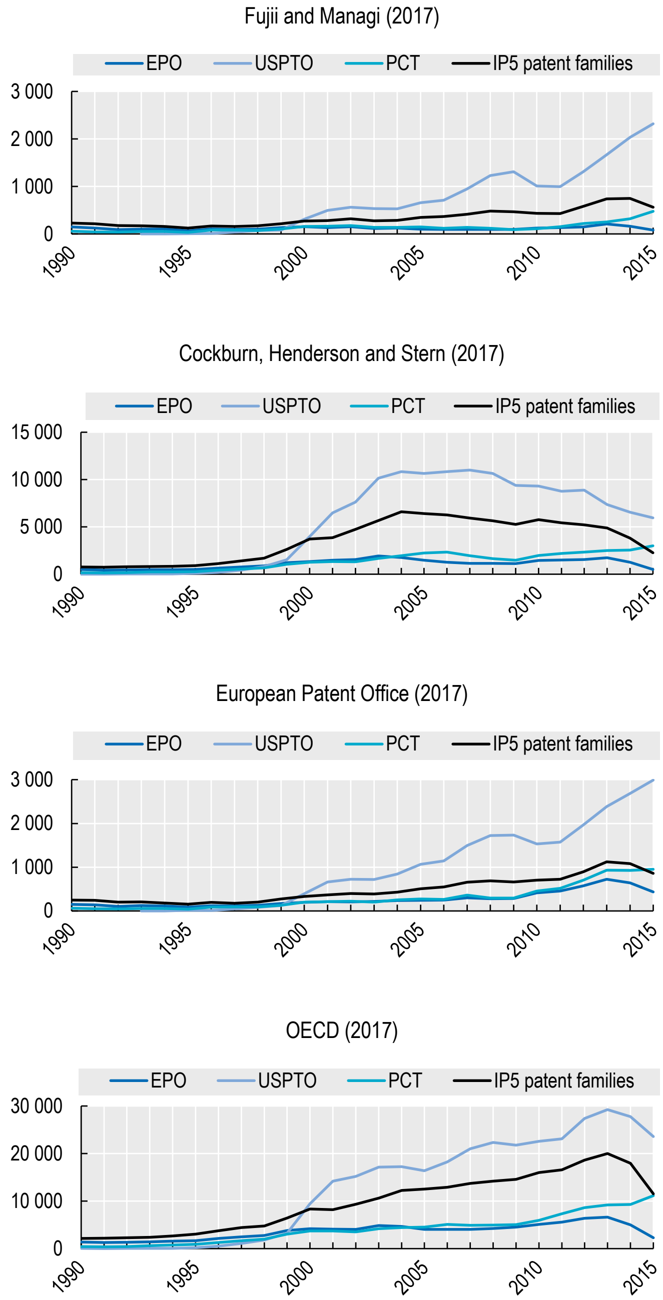
Figure 4.1 shows how different the volumes of AI-related patents appear when different data sources are used, for each definition provided above, and across the four taxonomies considered. The figures, which depict the number of applications of AI-related patents during the period 1990-2015, clearly highlight that considering EPO, USPTO, IP5 patent families or patent applications filed under the Patent Cooperation Treaty (PCT) leads to very different results.[33] Unless otherwise specified, patent-based statistics presented in this paper refer to the filing date of the patent application or the earliest filing date of the patent family.

[32] See Dernis et al. (2015) for more details about IP5 patent families.

[33] To this end, CPC or USPC codes were translated to IPC codes, to enable the identification of patents filed in IP offices outside the European or US patent systems (e.g. IP5 patent family members). However, no direct correspondence with the IPC can be found for USPC class 901 "Robots" used by Cockburn et al (2017). The robotics portion of the definition therefore only relies on the keyword search. The EPO definition relies on list of new CPC codes with no corresponding IPC codes. Consequently, CPC codes A61B5/7264-A61B5/7267, F01N2900/04-F01N2900/0422 and F05D2270/00-F05D2270/71 are not covered for patent filed at offices not using the CPC classification system.

Figure 4.1. Trends in AI-related patents according to different definitions, 1990-2015

Number of patents filed at EPO, USPTO or via the PCT and number of IP5 patent families



Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

While the definitions developed by Fujii and Managi (2017) and EPO (2017) give similar levels (about 750 and 1 100 AI-related patent families, respectively in 2013), the number of AI-related patents identified by Cockburn et al. (2017) is much larger, with 5 000 to 6 000 AI-related patent families filed every year since the mid-2000s. The OECD preliminary selection of IPC codes, which was broader than all the others in scope, is by far the largest.

Table 4.1 details the extent to which definitions overlap. By construction, Fujii and Managi (2017) AI-related patents are fully included in all other definitions; and 70-72% of EPO (2017) AI-patents are also covered in the other definitions.

Table 4.1. Definitions overlap, IP5 patent families

%	EPO	Fujii & Managi	Cockburn et al.	OECD
EPO		100	10	4
Fujii & Managi	70		9	3
Cockburn et al.	72	100		12
OECD	72	100	32	

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Text mining techniques applied on the abstracts[34] of IP5 patent documents and on USPTO patent claims[35] allowed uncovering commonalities and differences of inventions across definitions, and the way these evolve over time. To this end, the text contained in patent abstracts or claims was curated and parsed using "Bag of Word" techniques, in order to compile word frequencies, either for single words (unigram) or combinations of words (bigram / trigram)[36], in AI-related patent documents. Results showed that while Fujii and Managi's and EPO's selection of AI-related patents seemingly focus on data analysis, and feature some neural networks components, the two other AI-related definitions relate more to applications of AI (see Annex Figure 1). These include: treatment of information, devices, analysis of images, display devices and AI-related applications related to mobile phones.

Terminology employed in the abstract of patents under Fujii and Managi's definition is the closest to that of the EPO (2017), which in turn is closely related to that of patents identified using Cockburn et al. (2018). Correlations across samples are strongest when looking at simple word frequencies. However, the nature of the inventions, which was assessed by looking at similarities in the word combinations written in the abstracts, vary significantly across definitions, with the exception of the EPO (2017) and Fujii and Managi (2017), which remain closely linked, as shown in Annex Table 1 and Annex Figure 2).

The trend, volume and text analyses performed confirm that the four taxonomies all convey a different idea of what artificial intelligence is and does, and that no standard or agreed definition of AI-related inventions has thus far emerged. Taxonomies such as those of Fujii and Managi (2017) and the EPO (2017) appear somewhat conservative, as they focus mainly on computational models, whereas OECD (2017) experimental definition is rather geared towards AI applications, including image processing or digital devices, as is the one of Cockburn et al. (2018), especially with respect to robotics. This calls for alternative and more refined approaches aimed at better defining the scope of AI-related patents.

[34] If written in English.

[35] USPTO patents are at present the only one for which claims are readily usable. Future analysis will aim at scanning the claims of patents filed in other jurisdictions as well.

[36] Curation implied removing non-informative terms and harmonising words spellings using stemming and replenishment techniques. As discussed in Box 4.1. Text mining of patent data, the textual analysis is mainly performed on the abstracts.

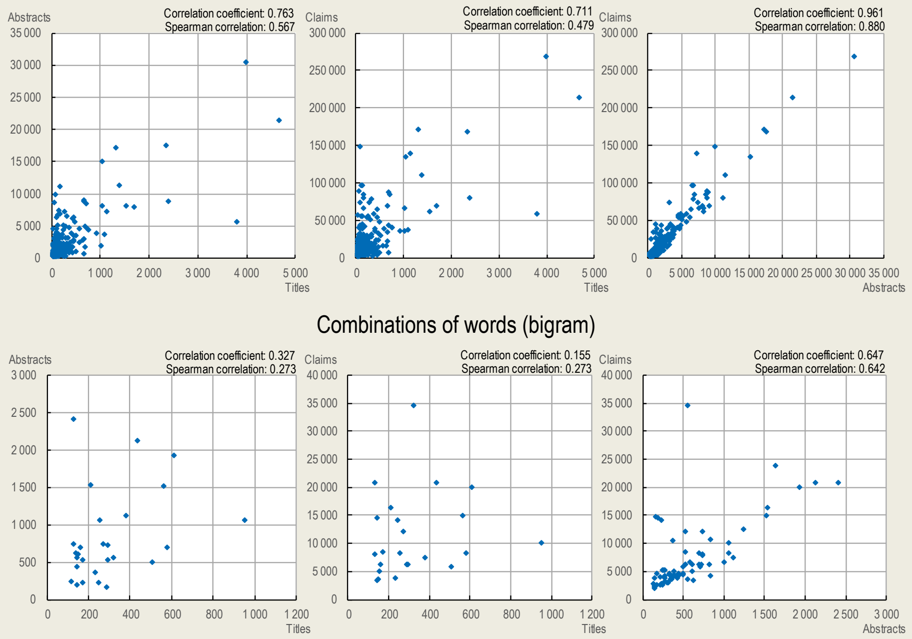
Box 4.1. Text mining of patent data

The level of details of patent records on which text mining techniques are applied differs across IPOs. It is frequently argued that the text contained in the patent description, especially in the claims, is the closest to the invention, as it is claims that set the legal boundaries of what can be defended in court. However, the EPO Worldwide Patent Statistical Database (PATSTAT, Spring 2018), which represent the main source of information for patents applied worldwide, does not include patents claims. Conversely, USPTO patent claims are available for download from the USPTO online portal. In the case of other IP offices, text mining can be performed only on titles or abstracts.

To check the extent to which results may change depending on the use of titles, abstracts or claims when performing keyword-based analyses aimed to identify AI-related patents, it is important to check the ex refinement exercise, top words and top combinations of words (bigram) were extracted from titles, abstracts and claims provided for a sample set of USPTO patents identified using the OECD (2017) experimental definition of AI. The analysis was performed on data for the years 2000, 2005, 2010 and 2013. The strong correlation emerging between the word frequencies extracted from titles, abstracts and claims suggest that abstracts represent a good proxy of the content of patents, in case claims text is not available.

In what follows, the analysis is therefore performed using the abstract of patents filed at IP5 offices.

Correlation of top frequent words across patent text source, titles, abstracts and claims



Simple words (unigram)

Combinations of words (bigram)

		Unique words						Word combination (Bigram)					
		Correlation coefficients			Spearman correlation			Correlation coefficients			Spearman correlation		
		Titles	Abstracts	Claims	Titles	Abstracts	Claims	Titles	Abstracts	Claims	Titles	Abstracts	Claims
Titles	2000		0.68	0.63		0.61	0.46		0.11	-0.05		0.12	0.13
	2005		0.72	0.67		0.62	0.45		0.23	0.14		0.27	0.29
	2010		0.74	0.69		0.58	0.44		0.33	0.23		0.38	0.36
	2013		0.76	0.71		0.57	0.48		0.33	0.15		0.27	0.34
Abstracts	2000	0.68		**0.96**	0.61		**0.84**	0.11		**0.64**	0.12		**0.40**
	2005	0.72		**0.96**	0.62		**0.88**	0.23		**0.66**	0.27		0.34
	2010	0.74		**0.96**	0.58		**0.89**	0.33		**0.63**	0.38		**0.51**
	2013	0.76		**0.96**	0.57		**0.88**	0.33		**0.65**	0.27		**0.64**
Claims	2000	0.63	**0.96**		0.46	**0.84**		-0.05	**0.64**		0.13	**0.40**	
	2005	0.67	**0.96**		0.45	**0.88**		0.14	**0.66**		0.29	0.34	
	2010	0.69	**0.96**		0.44	**0.89**		0.23	**0.63**		0.36	**0.51**	
	2013	0.71	**0.96**		0.48	**0.88**		0.15	**0.65**		0.34	**0.64**	

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

4.1.3. Identifying patents through keyword searches

To refine the IPC classification-based approach initially pursued, keyword searches were performed on IP5 patent documents (i.e. on titles and abstracts for all patents; also on full claims' text for USPTO patents only). The keywords used for the purpose were those identified through the bibliometric analysis performed in section 2, the "AI-193" list. In order to avoid overestimating the number of AI-related patent, however, the conservative approach pursued entailed considering AI-related only those patents featuring at least two AI-193 keywords. This reduced significantly the number of patents retrieved.

The sample of patents identified through the keywords search seems to only partially overlap with those identified on the basis of the taxonomies presented in the previous section. Out of the almost 21 thousand patents extracted using keyword searches, 11% were also flagged following Fujii and Managi's (2017) definition, 13% the EPO (2017) definition, 29% the Cockburn et al.'s (2018) taxonomy, and 55% the OECD definition (see Table 4.2).

Table 4.2. Overlap of AI patents identified by keywords with other definitions, 2000-16

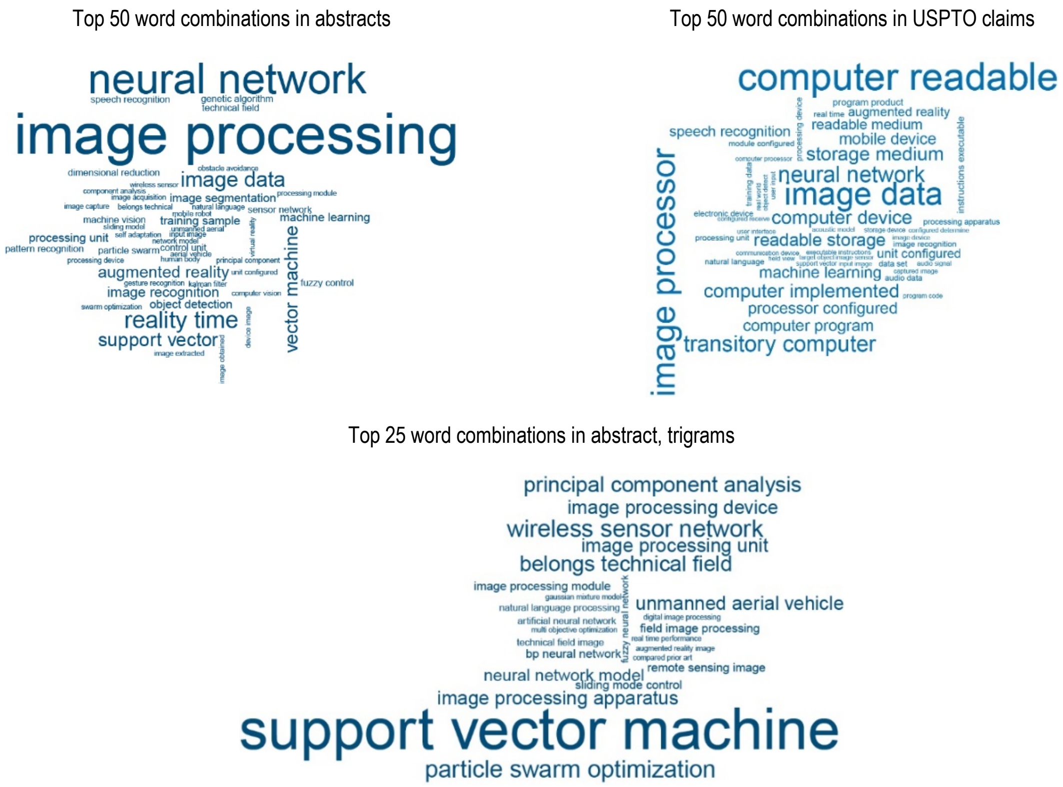
Share in total patents owned by AI companies, percentages

	2000-09	2010-16
Fujii & Managi	9.1	11.3
Cockburn et al	27.4	28.5
EPO	10.8	12.7
OECD	51.0	54.9

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Text mining performed on the abstracts and/or claims allowed shedding light on the content of the patents identified by means of the AI-193 keyword exercise. The resulting top frequent words combinations (bigram or trigram) in AI-related patents are displayed in the form of word clouds in Figure 4.2. Inventions related to *image processing*, *treatment of image data*, *control*, *neural networks* types of algorithms appear to account for an important part of AI-related technological developments.

Figure 4.2. Top word combinations in AI-patent identified by keywords, 2013



Top 50 word combinations in abstracts

Top 50 word combinations in USPTO claims

Top 25 word combinations in abstract, trigrams

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

In order to check the extent to which the IPC class based approach and the keyword-based one overlap, Table 4.3 shows the IPC codes to which the patents identified through the AI-193 exercise belong. This let emerge the importance of IPC group G06K (*Recognition of data; presentation of data; record carriers; handling record carriers*), G06T (*Image data processing or generation, in general*), and G06F (*Electric digital data processing*).

The code G06K9/00 '*Methods or arrangements for reading or recognising printed or written characters or for recognising patterns*' accounts for about 14% of the keyword identified AI patents filed since the year 2000. In the early 2010s, codes G06K9/62 '*Methods or arrangements for recognition using electronic means*', G06T7/00 '*Image data processing or generation, in general*', and G06F17/30 '*Database structures for information retrieval'* account for 13%, 9% and 7% of AI-related patents in the sample respectively. These IPC codes are also frequently associated with G06K9/00, as can be seen in Figure 4.3, which shows the most frequent combinations of IPC codes found in AI-patent identified through the keyword exercise, for the year 2013[37]. In the 2000s, G06T1/00 '*General purpose image processing*' was the second most frequent code allocated to the patents in AI keyword-based sample.

[37] As 2013 is the most recent year for which we can be confident that data are not truncated, given the fact that we are considering IP5 families and the time it takes for patent applications and publications to emerge, that year is used for most of the analysis.

Noteworthy, the IPC class G06N is not predominant among the most frequent IPC codes in AI-related patents identified using keywords, while the corresponding code emerges when looking at the CPC classification, mainly used by the EPO and USPTO.[38]

Table 4.3. Top 30 IPC codes in AI-patent identified by keywords, 2000-16

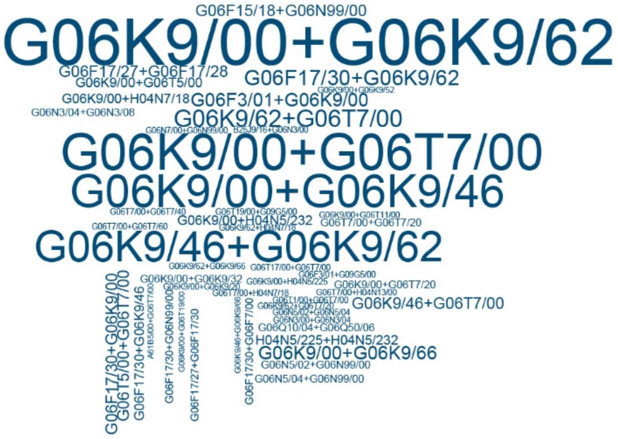
Share of IP5 patents by IPC code in AI-related patents, percentages

IPC code	Description	2000-09	2010-16
G06K9/00	Methods or arrangements for reading or recognising printed or written characters or for recognising patterns, e.g. fingerprints	13.8	14.3
G06K9/62	Methods or arrangements for recognition using electronic means	6.4	13.0
G06T7/00	Image data processing or generation, in general	7.5	9.1
G06F17/30	Database structures for information retrieval	7.4	7.0
G06K9/46	Extraction of features or characteristics of the image	3.5	5.8
G06N3/08	Computer systems based on biological models: Learning methods	1.4	3.8
G06F3/01	Input/output arrangements for interaction between user and computer	0.9	3.2
G06F19/00	Digital computing or data processing equipment or methods, specially adapted for specific applications	5.2	3.0
G05D1/02	Control of position or course in two dimension	1.7	2.6
H04N7/18	Closed-circuit television systems	3.2	2.5
G06K9/66	Methods or arrangements for recognition using simultaneous comparisons or correlations of the image signals with learning process	0.9	2.5
G05B13/04	Electric adaptative control systems involving the use of models or simulators	1.0	2.4
G06T7/20	Analysis of motion (motion estimation for coding, decoding, compressing or decompressing digital video signals)	2.2	2.3
G06N99/00	Other computer systems based on specific computational models	0.5	2.3
G06T19/00	Manipulating 3D models or images for computer graphics	0.3	2.3
G06T5/00	Image enhancement or restoration, e.g. from bit-mapped to bit-mapped creating a similar image	2.2	2.3
B25J9/16	Programme controls (centrally controlling a plurality of machines)	1.4	2.1
G06N3/02	Computer systems based on biological models using neural network models	1.5	2.1
G06F17/27	Automatic analysis, e.g. parsing, orthograph correction, handling natural language data	1.6	2.0
H04N5/232	Remote control for television cameras	1.3	1.9

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

[38] At the USPTO, patents are allocated to technology areas using the US Patent Classification system, that is in turn translated into the IPC classification system. The allocation of IPC codes, for the same invention, may differ when the patent is filed at another office. A sensitivity analysis was conducted on the frequency counts of IPCs, replacing the IPC codes of US patents with their non-US equivalent patents. However, this phenomenon has a little impact on the final IPC rankings (correlation coefficients were above 0.99).

Figure 4.3. Top combinations of IPC codes in AI-patent identified by keywords, 2013



Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

4.2. Patents citing scientific publications related to AI

Yet another approach which could be pursued to identify AI-related patents is to identify those patents that are closely linked to scientific papers in AI. Published patent documents in fact contain references to documents that are closely related to the invention to be protected, and that are considered as prior art. These references can be to previous patents or to non-patent literature (NPL). The Max Planck Digital Library has developed a method to link NPL with scientific reference data (see Knaus and Palzenberger, 2018), which minimises type I and type II errors.

For the purpose of the present exercise data from Max Planck Institute covering a set of patents citing papers presented at AI conferences are used. More than 28 000 patents filed within the IP5 offices were thus identified, whose overlap with AI-patents selected using the IPC code and keywords methods is shown in Table 4.4. In 2010-16, between 5% and 7% of patents citing AI-papers were also identified using Fujii and Managi (2017), EPO (2017) definitions or the AI-193 list of keywords. 60% of them were also found in the OECD (2017) experimental definition.

Table 4.4. Overlap of patents citing AI-related papers with other definitions, 2000-16

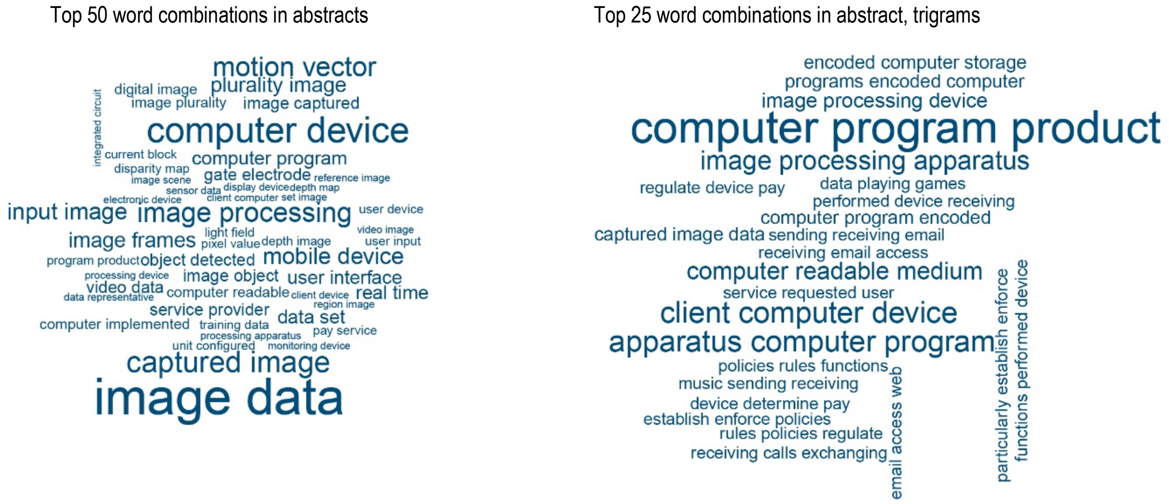
Share in total patents citing AI-related papers, percentages

	2000-09	2010-16
Fujii & Managi	3.1	5.0
Cockburn et al	14.6	12.6
EPO	4.1	6.5
OECD	55.0	57.9
Keywords	4.0	4.5

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Figure 4.4 presents the top frequent co-occurences of terms in IP5 patents citing AI-related papers. The analysis of the patent documents, based on the abstract or claims (for USPTO patents), shows such patented inventions to mainly relate to *processing of image data*, *programmes embedded in computer*, or *other IT devices*.

Figure 4.4. Top words and combinations in patents citing AI papers, 2013


Top 50 word combinations in abstracts

Top 25 word combinations in abstract, trigrams

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Inspecting the IPC codes of the patents in the sample reveals a technological focus that is similar to the one identified in the other AI-patent samples. In 2010-16, about 20% of patents citing AI documents were allocated to code G06K9/00 ‘*Methods or arrangements for reading or recognising printed or written characters or for recognising patterns*’ (Table 4.5). Other IPC codes relate to database structures (G06F17/30) and image data processing (G06T7/00, G06K9/46, G06T5/00).

Table 4.5. Top 20 IPC codes in patents citing AI-related papers, 2000-16

Share of IP5 patents by IPC code in patents, percentages

IPC code	Description	2000-09	2010-16
G06K9/00	Methods or arrangements for reading or recognising printed or written characters or for recognising patterns, e.g. fingerprints	17.2	20.0
G06F17/30	Database structures for information retrieval	13.5	12.8
G06T7/00	Image data processing or generation, in general	6.1	10.5
G06K9/62	Methods or arrangements for recognition using electronic means	4.4	7.7
G06K9/46	Extraction of features or characteristics of the image	4.3	7.0
G06F17/50	Computer-aided design (for the design of test circuits for static stores)	5.3	3.6
G06T5/00	Image enhancement or restoration, e.g. from bit-mapped to bit-mapped creating a similar image	2.8	3.3
G06T7/20	Analysis of motion (motion estimation for coding, decoding, compressing or decompressing digital video signals)	2.7	3.3
G06F3/01	Input/output arrangements for interaction between user and computer	2.2	3.3
H04L29/06	Communication control characterised by a protocol	2.6	2.9
H04N5/232	Remote control for television cameras	0.8	2.9
G09G5/00	Control arrangements or circuits for visual indicators	4.7	2.8
G06F15/16	Combinations of two or more digital computers for a simultaneous processing of several programs	4.0	2.8
H04N7/18	Closed-circuit television systems	2.0	2.7
H04N13/02	Image signal generators	0.6	2.7
G06F19/00	Digital computing or data processing equipment or methods, specially adapted for specific applications	3.0	2.7
G06K9/32	Aligning or centering of the image pick-up or image-field	1.8	2.6
G06K9/36	Processing the image information without deciding about the identity of the image	4.1	2.6
G06F17/00	Digital computing or data processing equipment or methods	5.7	2.6
G06T15/00	3D image rendering	3.2	2.6

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

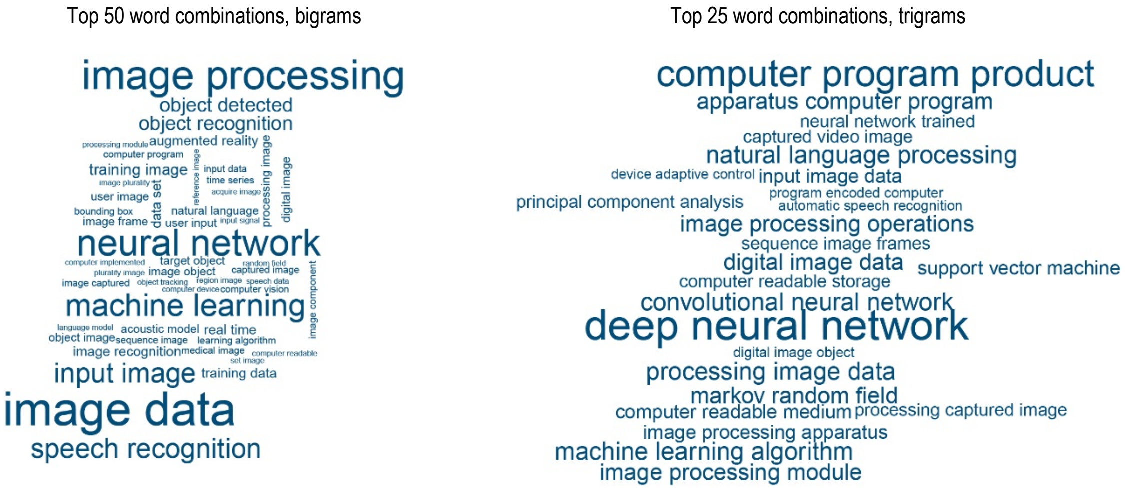
4.3. Converging towards a unique taxonomy of AI-related patents

The patent analysis performed so far clearly highlights how different results may be, in terms of both number of patents and of technological breadth, when different identification and mapping approaches are pursued.

The keyword-based approach and the sample of patents citing AI-related documents emerge as being similar in terms of technological content, over time (see Annex Table 2). To examine the commonalities emerging among the different samples, in the attempt to identify those patents that are ‘really’ related to AI, the set of AI-related patents was reduced to 1 028 patents filed in IP5 offices, i.e. those belonging to all samples (called “overlap sample”).

Figure 4.5 shows the most frequent combinations of words contained in the AI-related patents in the overlap sample for the period 2010-16. As can be seen, i*mage data* and *image processing* are the most frequently combined words in the patent abstracts in 2010-16, followed by *neural network* and *machine learning*. Conversely, when bundles of three words are considered, *deep neural networks* and *computer program product* appear as the most frequent, followed by a number of combinations that were already emerged in previous frequency analysis and that are very much coherent with what emerged from the software analysis and the keyword one.

Figure 4.5. Top words and combinations in AI-related patents, overlap sample, 2010-16



Top 50 word combinations, bigrams

Top 25 word combinations, trigrams

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

In terms of technological areas, in the period 2010-16, almost 60% of the AI patents in the overlap sample appeared to be allocated to IPC codes related to the automated recognition of patterns, G06K9/00 and G06K9/62, as reported in Table 4.6. The treatment and the analysis of images came next, with 16% and 15% of patents allocated to categories G06T7/00 or G06K9/46. Those IPC codes were also the most frequently combined in patents filed during that period (see Annex Figure 3).

Table 4.6. Top 20 IPC codes in patents, overlap sample, 2000-16

Share of IP5 patents by IPC code in patents, percentages

IPC code	Description	2000-09	2010-16
G06K9/00	Methods or arrangements for reading or recognising printed or written characters or for recognising patterns, e.g. fingerprints	38.0	35.3
G06K9/62	Methods or arrangements for recognition using electronic means	21.2	23.6
G06T7/00	Image data processing or generation, in general	10.8	16.0
G06K9/46	Extraction of features or characteristics of the image	10.4	14.9
G06F17/30	Database structures for information retrieval	7.8	8.9
G06F15/18	Learning machines	13.9	7.9
G06N99/00	Other computer systems based on specific computational models	0.8	6.2
G06T7/20	Analysis of motion (motion estimation for coding, decoding, compressing or decompressing digital video signals)	3.9	5.2
G06N5/02	Computer systems utilising knowledge representation	2.5	5.0
G06N3/08	Computer systems based on biological models: Learning methods	2.0	4.6
G06F17/27	Automatic analysis, e.g. parsing, orthograph correction, handling natural language data	4.7	4.4
G06F17/28	Processing or translating of natural language	5.1	4.4
B25J9/16	Programme controls (centrally controlling a plurality of machines)	1.2	4.2
G06T19/00	Manipulating 3D models or images for computer graphics	0.6	4.1
G10L15/00	Speech recognition	4.9	3.9
G06F3/01	Input/output arrangements for interaction between user and computer	2.0	3.9
G06K9/66	Methods or arrangements for recognition using simultaneous comparisons or correlations of the image signals with learning process	3.5	3.7
G06K9/32	Aligning or centering of the image pick-up or image-field	2.4	3.7
G06N3/04	Architecture of computer systems based on biological models using neural network models	0.2	3.5
G06K9/34	Segmentation of touching or overlapping patterns in the image field	4.7	3.5

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Given that, regardless of the approach pursued, the codes presented in Table 4.6 were identified as being AI-related, these codes can serve as a basis to pre-select IPC codes that would enable the compilation of a “core” AI taxonomy. The list of IPC codes was refined by looking in detail at the technology classes thus identified, to make sure not to include any IPC codes not strictly related to AI, based on its detailed definition. To improve the accuracy of the IPC-based taxonomy for AI patents, patents would need to be allocated to at least one of the top IPC codes in Table 4.7, and their text feature at least one of the AI 193 keywords, for patents to be considered as being AI-related.

The baseline IPC-based taxonomy proposed in Table 4.7, as well as the proposed list of 193 keywords resulting from the bibliometric investigation, was later reviewed and validated by patent examiners, who are the key experts in the field.

Table 4.7. Proposed baseline IPC-based taxonomy for AI-related patents

IPC code	Description
G06F15/18	Learning machines
G06F17/20	Handling natural language data
G06F17/27	*Automatic analysis, e.g. parsing, orthograph correction, handling natural language data*
G06F17/28	*Processing or translating of natural language*
G06F17/30	Database structures for information retrieval
G06K9/00	Methods or arrangements for reading or recognising printed or written characters or for recognising patterns
G06K9/46	Extraction of features or characteristics of the image
G06K9/48	*by coding the contour of the pattern*
G06K9/50	*by analysing segments intersecting the pattern*
G06K9/52	*by deriving mathematical or geometrical properties from the whole image*
G06K9/62	Methods or arrangements for recognition using electronic means
G06K9/64	*using simultaneous comparisons or correlations of the image signals with a plurality of references*
G06K9/66	*with references adjustable by an adaptive method, e.g. learning*
G06K 9/68	*using sequential comparisons of the image signals with a plurality of reference, e.g. addressable memory*
G06K 9/70	*the selection of the next reference depending on the result of the preceding comparison*
G06K 9/72	*using context analysis based on the provisionally recognised identity of a number of successive patterns*
G06K 9/74	Arrangements for recognition using optical reference masks
G06K 9/76	*using holographic masks*
G06K 9/78	Combination of image acquisition and recognition functions
G06K 9/80	Combination of image preprocessing and recognition functions
G06K 9/82	*using optical means in one or both functions*
G06N	Computer systems based on specific computational models
G06T1/40	General purpose image data processing using neural networks
G06T7/00	Image analysis
G06T7/20	Analysis of motion
G06T 7/207	*for motion estimation over a hierarchy of resolutions*
G10L15	Speech recognition

Source: WIPO, International Patent Classification (IPC), available at: http://www.wipo.int/classifications/ipc.

5. Proposed measures of AI-related scientific and technological developments

A preliminary version of this paper was brought to the attention of several OECD working groups in 2018, as well as to members of the OECD-led IP Statistics Task Force, asking for feedback and advice. During this consultation phase, the lists of keywords and the proposed baseline IPC taxonomy for patents were reviewed. In this way, an agreement with experts was reached with respect to the criteria to be used to identify AI-related scientific developments contained in scientific papers and conference proceedings, as well as AI-related technological developments outlined in published patent documents. First statistics illustrating the results of this consultation phase are presented at the end of the section.

5.1. Refining the list of keywords and classification codes

In January 2019, the OECD organised a workshop to further discuss the proposed methodology and to agree on a common approach to measure developments in AI, especially in terms of what can be intended to be AI-related patents.

One of the main takeaways of the workshop was that, while machine-learning techniques have been used by different offices to help identifying AI related developments, their design looks still non-trivial and results far from perfect. Experts agreed that, at present, their use was a desirable complement but that machine learning approaches could not represent the only solution. Experts agreed that more traditional approaches were still very much needed at present.

Developing a more traditional classification and/or keyword approach were considered among the best options by the research community, as they can provide a first good proxy of AI developments. Of course, limitations exist in this case as well. Identifying AI-related developments by means of relying on a set of classification codes or keywords may restrict our ability to detect emerging AI-related science and technologies, which are not well known or established, as the field is evolving very rapidly. Furthermore, even though patent classification schemes are regularly revised to account for the emergence of new technologies, it may take some time for a new classification code to appear in patent databases. This argues for the need to periodically revise and refine the list of keywords currently identified, as AI evolves.

When meeting at the OECD in January 2019, patent examiners and experts went through the full list of patent classification codes and keywords, to validate or challenge them. In particular, representatives of IP Australia, the Canadian Intellectual Property Office (CIPO), the EPO, the Israel Patent Office (ILPO), the Italian Patent and Trademark Office (UIBM), the National Institute for Industrial Property of Chile (INAPI), the United Kingdom Intellectual Property Office (UK IPO), and the USPTO contributed importantly to this exercise. The group went through each of the IPC codes proposed in section 4.3, signalling whether all patents classified in the identified class would need be considered as AI-related, whether the patents filed in a certain class would need to be searched using at least one of the identified AI-related keywords, or else. Additional codes of the Cooperative Patent Classification (CPC) scheme were also complement the search.

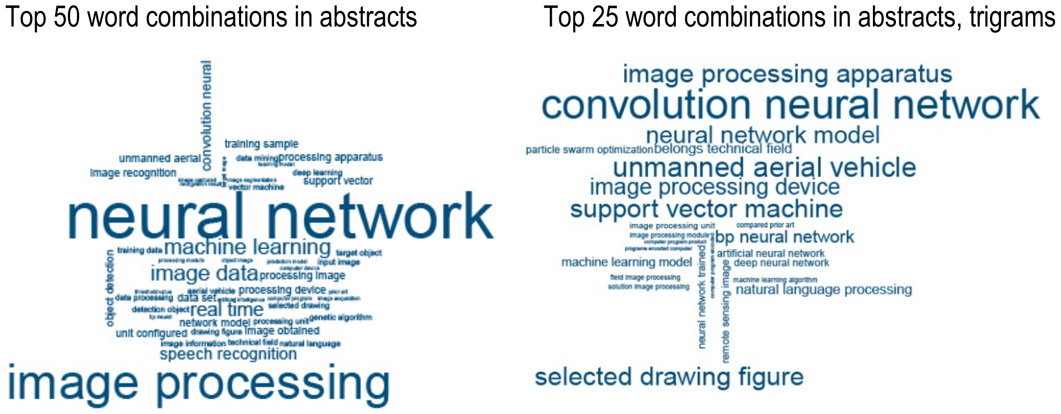
During the review process, an approach was delineated as the (currently) most accurate way to identify AI-related patents. In particular it was agreed that: in certain cases, all patents filed in purposely identified IPC codes are to be considered as being AI-related; for another group of IPC/CPC codes, keyword searches are to be performed (to avoid false positives); finally, and to identify AI developments happening in other areas, keywords-only searches using combinations of at least three keywords (listed in the Annex Table C.1), should be implemented.

Later, in the course of the 2019, the UK IPO has implemented a similar search strategy for patents in its AI report (UK IPO, 2019), building on the work of the IP Statistics Task Force and on WIPO's report (WIPO, 2019). The additional classification codes and keywords identified in the UK IPO's report have been included in the OECD search, as can be seen in Annex C.

Summarising, the methodology implemented here to identify and measure scientific and technological developments related to AI is as follows:

- Scientific articles in AI are those featuring in their abstract at least two of the AI-related keywords listed in Annex Table C.1. Only articles, books, business articles, chapters, conference papers, articles in press and reports are considered for the purpose.
- Patents in AI are those:
    - classified in one of the IPC codes listed in Annex Table C.2.1; or
    - classified in one of the IPC codes listed in Annex Table C.2.2. and featuring in their English abstract or claims at least one of the keywords Annex Table C.1..; or
    - classified in one of the CPC codes listed in Annex Table C.2.3. and featuring in their English abstract or claims at least one of the keywords listed in Annex Table C.1..; or
    - featuring at least three of the keywords in Annex Table C.1. in their English abstract or claims, in the patent document.

Figure 5.1. Top word combinations in AI-related patents, 2014-16



Top 50 word combinations in abstracts

Top 25 word combinations in abstracts, trigrams

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, July 2019.

Figure 5.1 displays the top terms that occur most frequently together in the abstracts of IP5 patents, when using two or three keywords to identify AI-related patents. The World cloud refers to the patents filed in the period 2014-16. *Neural networks* and *convolutional neural networks* terms appear very often, in about 15% of AI-related patent documents, followed by *image processing* (11%).

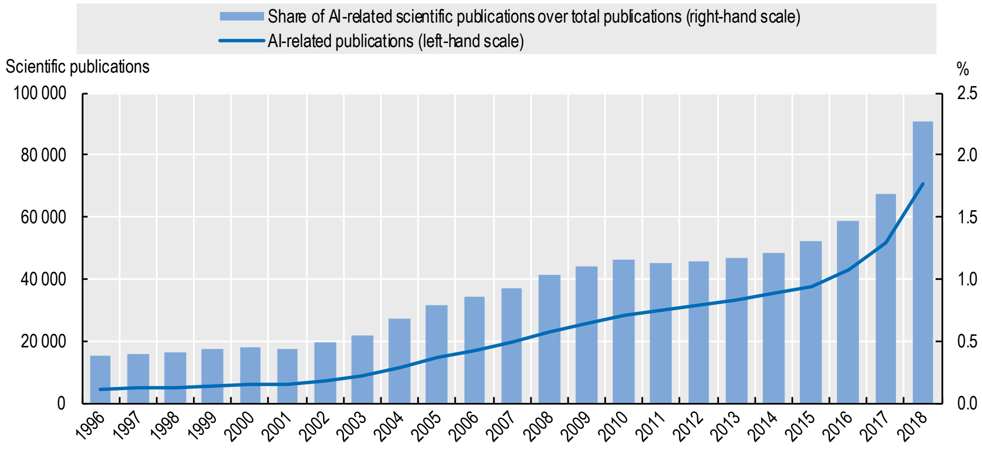
5.2. AI-related developments in science and technology: first statistics

The revised search strategy detailed above was implemented on the most recent available editions of the Scopus® database (version 5.2019) and PATSTAT (EPO, Spring 2019).

Figure 5.2 shows the evolution of scientific publications in AI from 1996 until 2018. As can be seen, the number of publications in AI accelerated in the early 2000s, at a rate of 21% a year on average between 2000 and 2005, and continued growing at a steady pace of 10% a year on average until 2015. Since then, the number of AI papers has been again significantly increasing, by 23% on average a year since 2015. About 71 000 publications could be identified as being AI-related in Scopus® for the year 2018, nearly twice as much the level observed in 2015. The share of AI-related publications in total publications also shows a sharp increase in the recent years, passing from less than 0.5% in 2000 to about 1.1% in the late 2000s, and more than 2.2% over all publications in 2018.

Figure 5.2. Scientific publications related to AI, 1996-2018

AI-related publications and share of AI in total publications

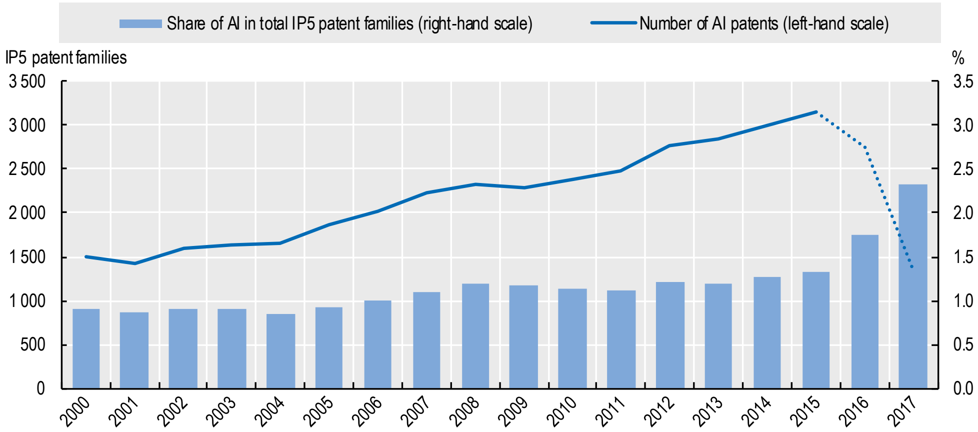


Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 5.2019.

AI-related inventions, as measured by IP5 patent families, evolved at a relatively slower pace of 5% a year on average between 2005 and 2015, as shown in Figure 5.3. Although the latest records of AI-related patents are not yet complete because of publication lags (complete statistics can only be displayed up to 2015, using the earliest application date), the available data reports a marked increase in the proportion of AI-related inventions in IP5 patent families in the latest years. This ratio averaged to about 1.1% in the late 2000s, and increased smoothly to over 1.3% in 2015. According to the preliminary figures for the latest years, over 2.3% of IP5 patent families refer to AI, more than twice the 2010 level.

Figure 5.3. Trends in AI-related patents, 2000-17

Number of IP5 patent families in AI and share of AI –related patents in total IP5 families

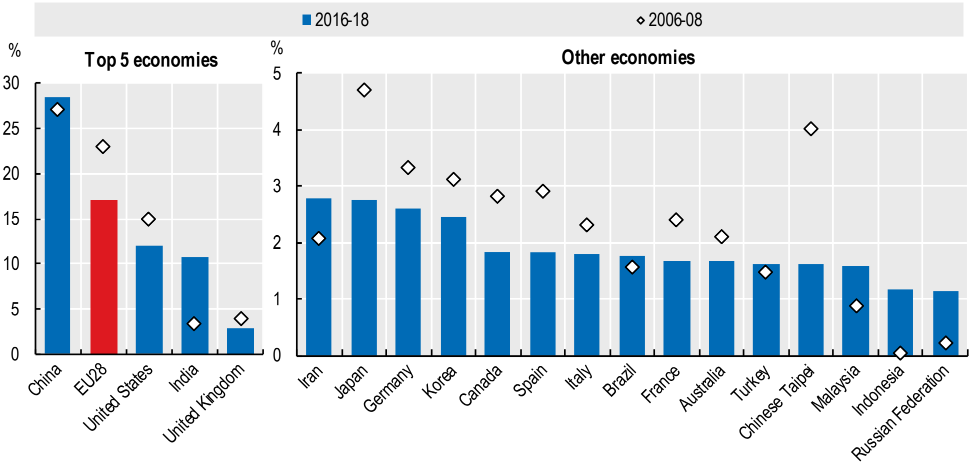


Note: Counts of IP5 patent families are reported according to the earliest filing date of patents that belong to the same family. Data from 2016 are truncated due to unpublished patent data.
Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, July 2019.

Figure 5.4 displays the geographical location of authors contributing to AI-related papers. When it comes to publications appearing in 2016-18, 28% of the world AI-related papers can be attributed to authors located in China. Over time, the share of AI publications originating from EU28, the United States and Japan decreased, as compared to the levels observed in 2006-08. It fell from the 23% of the first period to about 17% of AI papers in 2016-18 for the EU28; from 15% to 12% for the United States, and from 5% to less than 3% in Japan in the same reference periods. It is also interesting to remark that the number of papers from India-based authors jumped from 3.4% in 2006-08 to nearly 11% in 2016-18.

Figure 5.4. AI-related scientific publications by economies, 2006-08 and 2016-18

Top 20 economies with AI publications

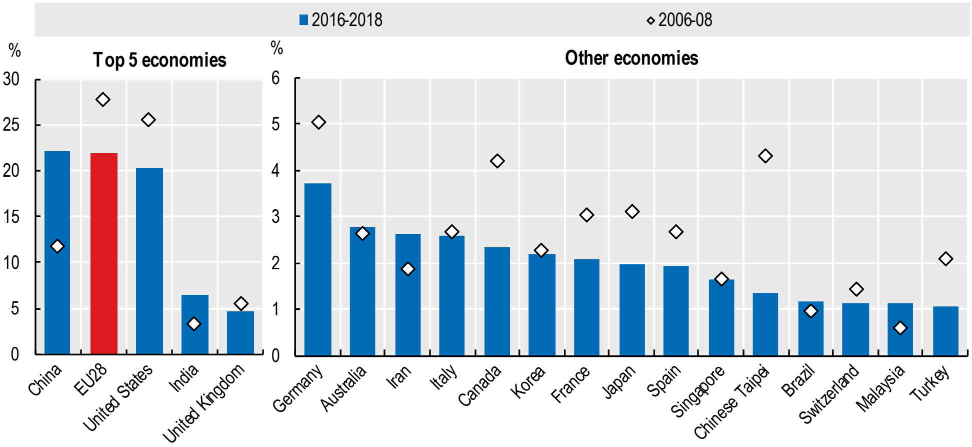


Note: The number of publications by author's location is based on fractional counts.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 5.2019.

In 2016-18, China also ranked first in terms of top cited AI publications, with a share of 22.1%, as shown in Figure 5.5 below, surpassing the levels of EU28 (21.9%) and of the United States (20%). The contribution of China to highly cited papers in AI doubled the level observed ten years earlier, while the relative share of EU28 and the United States dropped by about five percentage points during the same period. India now features in the fourth position of the ranking, and has also doubled its share during the last decade.

Figure 5.5. Top-cited AI-related scientific publications, 2006-08 and 2006-18

Economies with the largest number of AI-related documents among the 10% most cited publications



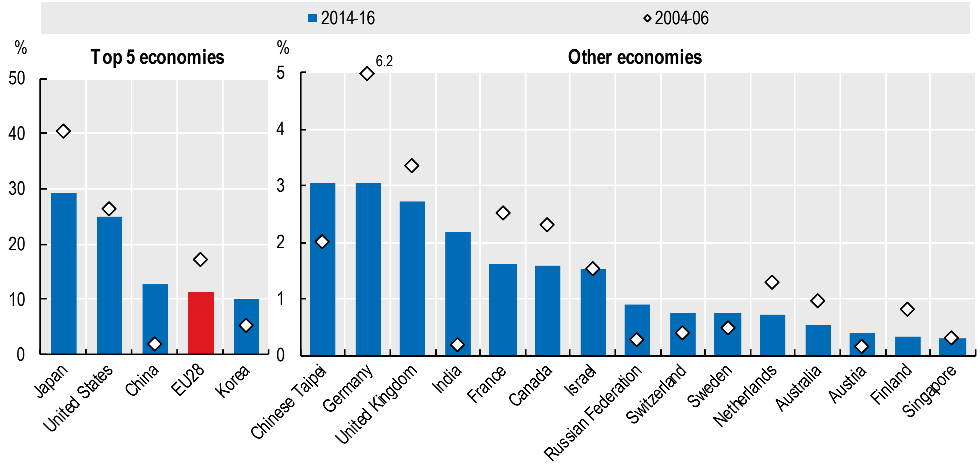
Note: Top-cited publications are the 10% most-cited papers normalised by publication journal scientific field(s) and type of document (articles, reviews and conference proceedings). The Scimago Journal Rank indicator is used to rank documents with identical numbers of citations within each class. This measure is a proxy indicator of research excellence. Estimates are based on fractional counts of documents by authors affiliated to institutions in each economy. Documents published in multidisciplinary/generic journals are allocated on a fractional basis to the ASJC codes of citing and cited papers.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 5.2019

Although China has the largest number of top-cited AI scientific publications in absolute numbers, the share of top-cited AI over total AI scientific publications for the 2016-18 period is 11.7% whereas it is 25.4% for the United States.

As seen in Figure 5.6, inventors located in Japan were responsible for about 29% of IP5 patent families in AI in 2014-16, a decrease compared to the level of 2004-06 (40%), while the contribution of US inventors remained at a level of 25-26%. An upward trend is conversely displayed by other Asian economies during the last decade, notably China, Korea, Chinese Taipei and India.

Figure 5.6. Top inventor's economies in AI-related patents, 2004-06 and 2014-16

Share of economies in AI-related patents, IP5 patent families, top 20 economies



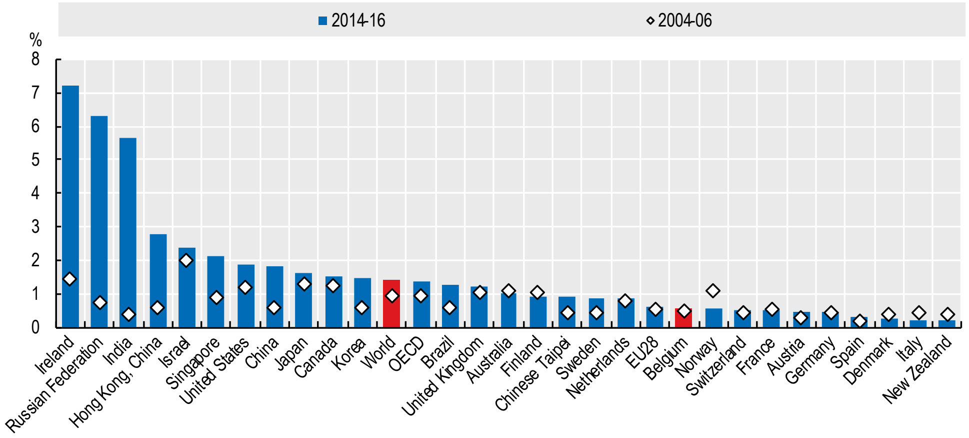
Note: Data refer to IP5 patent families in AI-related technologies, by earliest filing date and inventor's location, using fractional counts. Data for 2016 are incomplete.
Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, July 2019.

The contribution of China-based inventors to AI-related patents multiplied more than six fold, from less than 2% in the mid-2000s to nearly 13% in the last period, thus ranking third in the most recent period (Figure 5.6). Between 2004-06 and 2014-16 for Korea the share nearly doubled, increasing from 5.4% to nearly 10% of AI-related patents. Chinese Taipei increased its share by 1 percentage point, up to 3% in 2014-16, while Indian inventors saw their contribution passing from 0.2% in 2004-06 to 2.2% in 2014-16. The relative participation of most EU28 economies in AI-related IP5 patent families dropped from 17% to 11%, among which Germany, the United Kingdom, France, the Netherlands and Finland stand out for their contribution.

Figure 5.7 provides additional insights on the ownership of AI-related patents, showing the proportion of such patents in the patent portfolio of economies. The share of AI-related patents in the portfolio of applicants increased in almost all economies, from a world average of 0.9% in 2004-06 to 1.4% in 2014-16. Some economies displayed a significant increase of their AI-related contributions between the two periods: 7% of IP5 patent families owned by Ireland protected inventions in AI in 2014-16, compared to 1.4% ten years before, while the Russian Federation and India experienced an increase in their share of AI-patents from 0.7% and 0.4% respectively to about 6% in 2014-16.

Figure 5.7. AI-related inventions in total patents, by economies, 2004-06 and 2014-06

Share of AI-related patents in total IP5 patent families owned by economies



Note: Data refer to IP5 patent families in AI-related technologies, by earliest filing date and applicant's location, using fractional counts. Only economies owning more than 500 IP5 families in the time periods considered are included. Data for 2016 are incomplete.
Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, July 2019.

Conclusions

Artificial intelligence is high on the agenda of businesses and policy makers alike. It is therefore important to operationally define and map AI developments, as in the absence of measurement and empirical evidence, policies may be become ineffective if not distortive.

To this end, this work proposes a three-pronged approach aimed at identifying and measuring AI developments in science, as captured in scientific publications; technological developments, as proxied by patents; and software data, and in particular open source software.

The search strategy presented in the paper intends to provide the research community and analysts with a common operational definition of scientific and technological developments related to AI. The complete search strategy, applied to scientific paper and patents, is provided in A.C, and will need to be repeated (and possibly revised) periodically in order to account for future AI-related developments.

Finally, it is important to remember that AI is a complex technological paradigm, which is unfolding along a number of technological trajectories. A better understanding of what constitutes AI per se and what represents applications of AI to other fields or domains is therefore needed. This will be fundamental not only to shed light on what is developed, where and by whom, but also to understand how many and which parts of the overall phenomenon are needed for AI to become the welfare and productivity enhancing technology everybody hopes it will be. Also, this will be important to inform the design of a wide range of policies concerned with AI developments and applications.

References

Artificial Intelligence Index (2017), Artificial Intelligence Index: 2017 Annual Report, http://cdn.aiindex.org/2017-report.pdf

Bohannon, J (2016), "Who's the Michael Jordan of computer science? New tool ranks researchers' influence", *Science*, http://doi.org/10.1126/science.aaf9939

CISTP (2018). China AI Development Report 2018. http://www.sppm.tsinghua.edu.cn/eWebEditor/UploadFile/China_AI_development_report_2018.pdf

Cockburn I.M., R. Henderson and S. Stern (2018), "The Impact of Artificial Intelligence on Innovation", *NBER Working Paper* No. 24449. https://doi.org/10.3386/w24449

Craglia M. (Ed.), Annoni A., Benczur P., Bertoldi P., Delipetrev P., De Prato G., Feijoo C., Fernandez Macias E., Gomez E., Iglesias M., Junklewitz H, López Cobo M., Martens B., Nascimento S., Nativi S., Polvora A., Sanchez I., Tolan S., Tuomi I., Vesnic Alujevic L., *Artificial Intelligence - A European Perspective*, EUR 29425 EN, Publications Office, Luxembourg, 2018, ISBN 978-92-79-97217-1, http://doi.org/10.2760/11251

Dernis, H., Squicciarini M. and R. de Pinho (2016), "Detecting the emergence of technologies and the evolution and co-development trajectories in science (DETECTS): A 'burst' analysis-based approach", *Journal of Technology Transfer*, Vol. 41/5, pp. 930–960. http://doi.org/10.1007/s10961-015-9449-0

Dernis H., Dosso M., Hervás F., Millot V., Squicciarini M. and Vezzani A. (2015). World Corporate Top R&D Investors: Innovation and IP bundles. *A JRC and OECD common report*. Luxembourg: Publications Office of the European Union. http://doi.org/10.2791/741349

Elsevier (2018), *Artificial Intelligence: How knowledge is created, transferred, and used*. https://www.elsevier.com/connect/resource-center/artificial-intelligence

EPO (2017), *Patents and the Fourth Industrial Revolution*, http://documents.epo.org/projects/babylon/eponet.nsf/0/17FDB5538E87B4B9C12581EF0045762F/$File/fourth_industrial_revolution_2017__en.pdf

Fujii, H, and S Managi (2017), "Trends and Priority Shifts in Artificial Intelligence Technology Invention: A global patent analysis," *RIETI Discussion Paper Series 17-E-066*, https://www.rieti.go.jp/jp/publications/dp/17e066.pdf

Griliches, Z. (1990), "Patent statistics as economic indicators: A survey". *Journal of Economic Literature*, 28(4), pp1661– 1707, http://www.jstor.org/stable/2727442

Inaba, T. and M. Squicciarini (2017), "ICT: A new taxonomy based on the international patent classification", *OECD Science, Technology and Industry Working Papers*, No. 2017/01, OECD Publishing, Paris, https://doi.org/10.1787/ab16c396-en

IPO (2019), *Artificial Intelligence: A worldwide overview of AI patents and patenting by the UK AI sector*. https://www.gov.uk/government/publications/artificial-intelligence-a-worldwide-overview-of-ai-patents

Jordan M. (April 19th, 2018), 'Artificial Intelligence—The Revolution Hasn't Happened Yet', *Medium,* https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7

Knaus, J, and M. Palzenberger (2018), “PARMA. A full text search based method for matching non-patent literature citations with scientific reference databases. A pilot study”, Technical report by the Max Planck Digital Library, Big Data Analytics Group.

Krizhevsky, A., Sutskever, I., & G. E. Hinton (2012), “ImageNet classification with deep convolutional neural networks”, *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, pp1097-1105, https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf .

Martínez, Catalina. (2011). Patent families: When do different definitions really matter?. *Scientometrics*. 86 (1). 39-63. https://doi.org/10.1007/s11192-010-0251-3 .

McKinsey Global Institute (April 2018), *Notes from the AI frontier: Insights from hundreds of use cases,* https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/mgi_notes-from-ai-frontier_discussion-paper.ashx

Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation and OpenAI (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, adopted on 22 May 2019 after a proposal by the Expert Group on Artificial Intelligence at the OECD (AIGO). http://oe.cd/ai

OECD (2017), *OECD Science, Technology and Industry Scoreboard 2017: The digital transformation*, OECD Publishing, Paris, https://doi.org/10.1787/9789264268821-en .

OECD and SCImago Research Group (CSIC) (2016), *Compendium of Bibliometric Science Indicators. OECD*, Paris. Accessed from http://oe.cd/scientometrics

Samuel, Arthur (1959), “Some studies in machine learning using the game of Checkers”, *IBM Journal of Research and Development*, pp71-105, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.368.2254

Shoham, Y., R. Perrault, E. Brynjolfsson, J. Clark, J. Manyika, J.C. Niebles, T. Lyons, J. Etchemendy, B. Grosz and Z. Bauer (2018), "The AI Index 2018 Annual Report”, *AI Index Steering Committee, Human-Centered AI Initiative*, Stanford University, Stanford, CA, December 2018. http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

Squicciarini, M. and H. Dernis (2013), "A Cross-Country Characterisation of the Patenting Behaviour of Firms based on Matched Firm and Patent Data", *OECD Science, Technology and Industry Working Papers*, No. 2013/05, OECD Publishing, Paris, https://doi.org/10.1787/5k40gxd4vh41-en .

Van Eck, N.J., and L, Waltman (2018), “VOSviewer Manual”, available at: www.vosviewer.com/documentation/Manual_VOSviewer_1.6.8.pdf.

WIPO (2019). *WIPO Technology Trends 2019: Artificial Intelligence*. Geneva: World Inellectual Property Organization. https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf

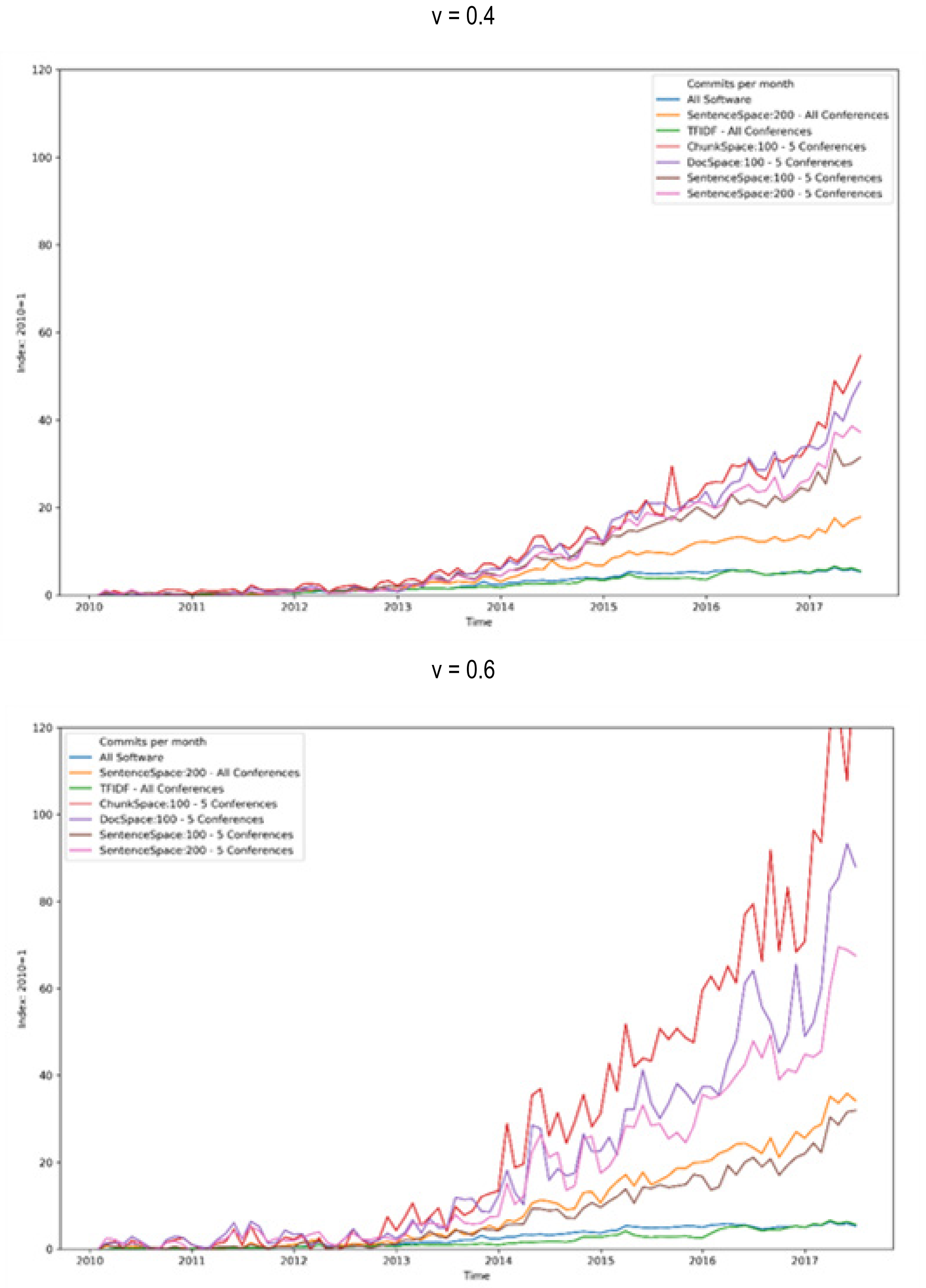
Annex

A. Sensitivity analysis for the SVM classifier

Annex Figure A.1illustrates the impact in the choice of those parameters. Both graphs show the growth in commits through time (similar to Figure 3.4). The various coloured lines show different embedding procedures, with the blue line (slowest growth) representing overall growth in commits to repositories on GitHub. The following observations can be made:

- The bottom graph displays stronger growth in commits to AI software than the top graph, which tells us there is a positive relationship between ***v*** and growth in commits to AI software. This is because the higher ***v*** is, the more the classifier is willing to leave out "core" AI documents from our AI classification, and the "purer" (more restrictive) our final classification of AI repositories. This implies, in turn, that commits to the more restricted group of AI repositories have grown considerably faster than to overall software. Choosing the optimal value for ***v*** is not obvious; this report has thus been conservative by choosing a high value for ***v***.
- The technology used for the embedding procedure matters: for instance, the tf-idf embedding (green line) displays almost no difference in growth compared to commits to overall software (blue line). This is because it is much less precise and classifies many non-AI repositories as AI, and thus tracks growth in overall software rather than specifically in AI software. One can also see that ChunkSpace and DocSpace display higher growth than SentenceSpace; this is to be expected as they train on word combinations larger than sentences and are thus likely to be more restrictive.
- The number of dimensions matter: the 200-dimensional SentenceSpace (light mauve line) suggests higher growth than the 100-dimensional SentenceSpace (brown line). This is because it clusters the original documents into a tighter space, classifying fewer external documents as AI-related for a given level of ***v***.
- Reducing the number of conferences (all conferences vs 5) from which to pick the "core" AI repositories also increases the growth in commits to AI repositories. This is probably because many publications in some of the conferences are not about AI.

Annex Figure A.1. Growth in commits to AI software for different methodologies



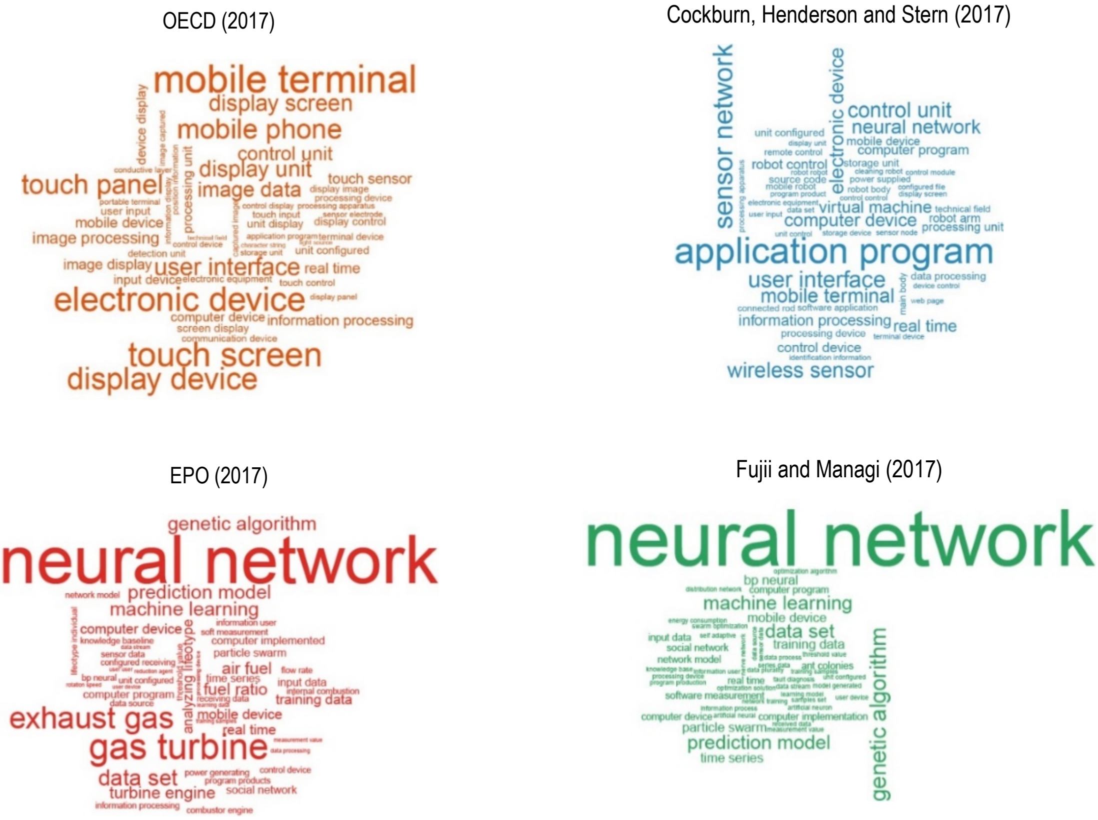
v = 0.4

v = 0.6

Source: OECD calculations based GitHub data from Google BigQuery and GitHub Search API, 2018.

B. Textual analysis of patents

Annex Figure 1. Top word combinations in AI-patent abstracts, 2013



OECD (2017)

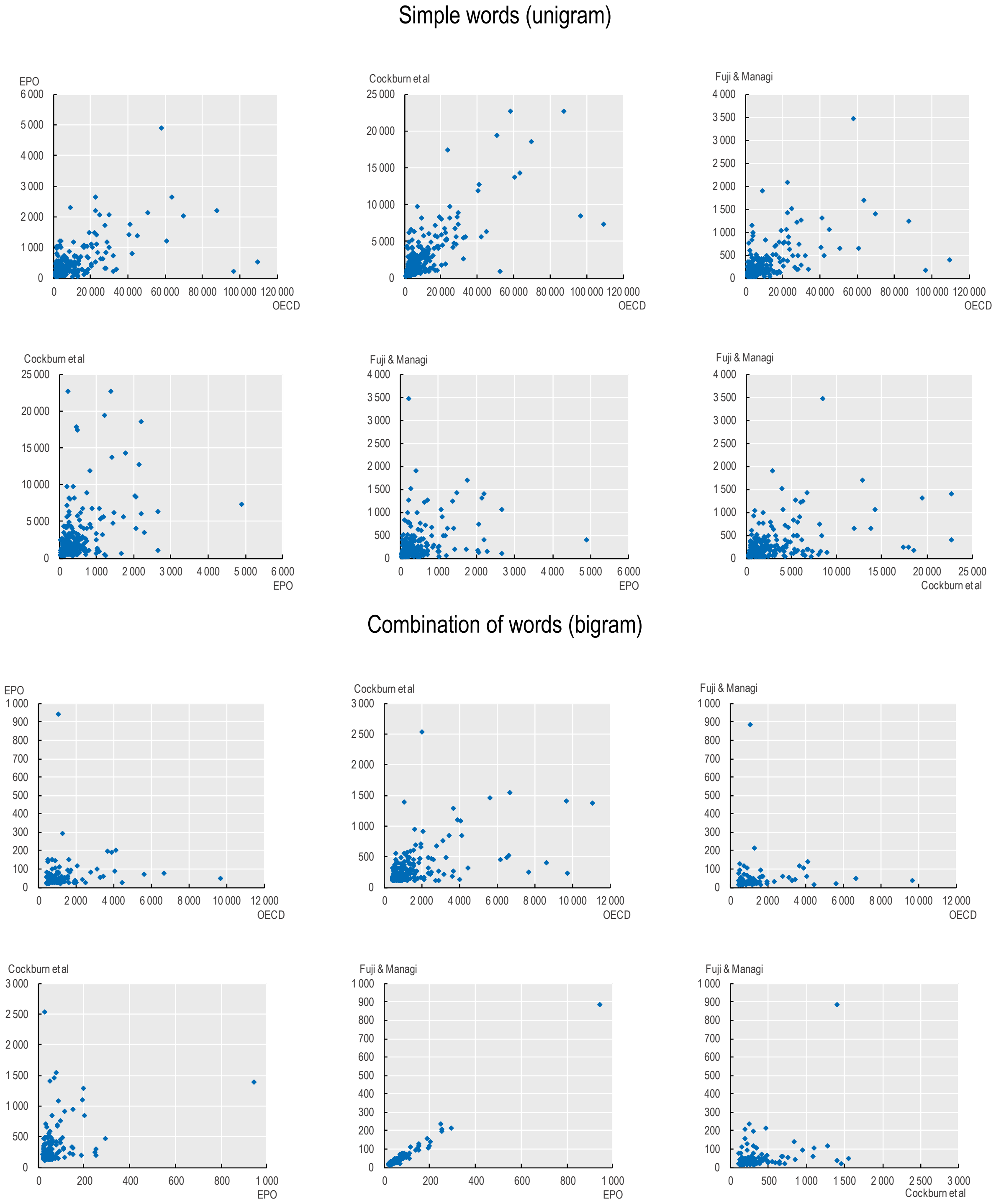
Cockburn, Henderson and Stern (2017)

EPO (2017)

Fujii and Managi (2017)

Note: Frequency of words based on the English abstracts of patents filed at IP5 offices.
Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Annex Figure 2. Distribution of word frequencies across existing patent taxonomies, 2013 based on English abstracts of IP5 patent documents



Simple words (unigram)

Combination of words (bigram)

Note: Data relies on the occurrence (or co-occurrence) of words contained in the abstracts of patents filed at IP5 offices in 2013.
Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Annex Table 1. Correlations coefficients across definitions, patent abstracts, 2013

Simple words (unigram)

	Correlation coefficients				Spearman correlation			
	OECD	EPO	Cockburn et al	Fuji & Managi	OECD	EPO	Cockburn et al	Fuji & Managi
OECD		0.590	0.760	0.543		0.563	0.756	0.575
EPO	0.590		0.712	0.968	0.563		0.604	0.932
Cockburn et al	0.760	0.712		0.647	0.756	0.604		0.616
Fuji & Managi	0.543	0.968	0.647		0.575	0.932	0.616	

Combination of words (bigram)

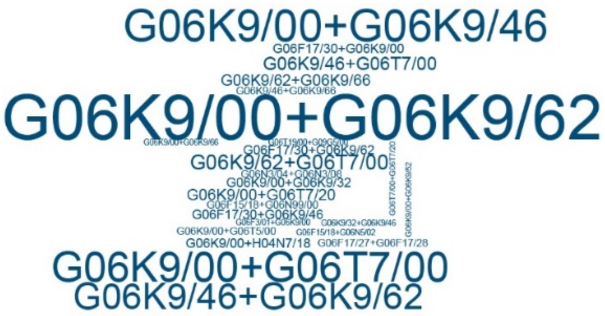
	Correlation coefficients				Spearman correlation			
	OECD	EPO	Cockburn et al	Fuji & Managi	OECD	EPO	Cockburn et al	Fuji & Managi
OECD		0.087	0.501	0.045		0.229	0.422	0.227
EPO	0.087		0.319	0.969	0.229		0.278	0.810
Cockburn et al	0.501	0.319		0.343	0.422	0.278		0.263
Fuji & Managi	0.045	0.969	0.343		0.227	0.810	0.263	

Note: Correlation measures were built on the occurrence (or co-occurrence) of words contained in the abstracts of patents filed at IP5 offices in 2013

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

Annex Figure 3. Top combinations of IPC codes in patent documents, 2010-16

Patents in the overlap sample



Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018

Annex Table 2. Correlation of IPC frequencies in AI-related patents' samples, 2005 and 2013

Correlations

	2005						2013					
	Fuji & Magani	Cockburn et al	EPO	OECD	Keywords	AI-NPL citation	Fuji & Magani	Cockburn et al	EPO	OECD	Keywords	AI-NPL citation
Fuji & Magani		0.36	0.96	0.19	0.33	0.29		0.21	0.93	0.14	0.27	0.28
Cockburn et al	0.36		0.43	0.18	0.24	0.33	0.21		0.27	0.14	0.16	0.22
EPO	0.96	0.43		0.21	0.40	0.37	0.93	0.27		0.19	0.39	0.44
OECD	0.19	0.18	0.21		0.54	0.64	0.14	0.14	0.19		0.50	0.58
Keywords	0.33	0.24	0.40	0.54		0.76	0.27	0.16	0.39	0.50		0.87
AI-NPL citation	0.29	0.33	0.37	0.64	0.76		0.28	0.22	0.44	0.58	0.87	

Spearman correlations

	2005						2013					
	Fuji & Magani	Cockburn et al	EPO	OECD	Keywords	AI-NPL citation	Fuji & Magani	Cockburn et al	EPO	OECD	Keywords	AI-NPL citation
Fuji & Magani		0.19	0.87	0.06	0.34	0.37		0.18	0.79	0.09	0.10	0.29
Cockburn et al	0.19		0.12	-0.32	0.08	0.17	0.18		0.31	-0.13	-0.02	0.37
EPO	0.87	0.12		-0.20	0.11	0.20	0.79	0.31		0.05	0.06	0.30
OECD	0.06	-0.32	-0.20		0.23	0.46	0.09	-0.13	0.05		0.41	0.37
Keywords	0.34	0.08	0.11	0.23		0.41	0.10	-0.02	0.06	0.41		0.52
AI-NPL citation	0.37	0.17	0.20	0.46	0.41		0.29	0.37	0.30	0.37	0.52	

Source: OECD, STI Micro-data Lab: Intellectual Property Database, http://oe.cd/ipstats, September 2018.

C. Proposed search strategy for AI-related patents and scientific papers

Annex Table C.1. List of AI-related keywords

action recognition human action recognition	activity recognition human activity recognition	adaboost
adaptive boosting	adversarial network generative adversarial network	*ambient intelligence*
ant colony ant colony optimisation	artificial intelligence human aware artificial intelligence	association rule
autoencoder	*autonomic computing*	*autonomous vehicle*
autonomous weapon	backpropagation	Bayesian learning
bayesian network	bee colony artificial bee colony algorithm	blind signal separation
bootstrap aggregation	*brain computer interface*	brownboost
chatbot	classification tree	cluster analysis
cognitive automation	cognitive computing	*cognitive insight system*
cognitive modelling	collaborative filtering	collision avoidance
community detection	computational intelligence	*computational pathology*
computer vision	*cyber physical system*	*data mining*
decision tree	deep belief network	deep learning
dictionary learning	dimensionality reduction	*dynamic time warping*
emotion recognition	ensemble learning	evolutionary algorithm differential evolution algorithm multi-objective evolutionary algorithm
evolutionary computation	face recognition	facial expression recognition
factorisation machine	feature engineering	feature extraction
feature learning	feature selection	*firefly algorithm*
fuzzy c fuzzy environment fuzzy logic fuzzy number fuzzy set intuitionistic fuzzy set fuzzy system *t s fuzzy system* *Takagi-Sugeno fuzzy systems*	gaussian mixture model	gaussian process
genetic algorithm	genetic programming	gesture recognition
gradient boosting gradient tree boosting	graphical model	*gravitational search algorithm*
hebbian learning	hierarchical clustering	high-dimensional data high-dimensional feature high-dimensional input high-dimensional model high-dimensional space high-dimensional system
image classification	*image processing*	image recognition
image retrieval	*image segmentation*	independent component analysis
inductive monitoring	instance-based learning	*intelligence augmentation*
intelligent agent intelligent software agent	intelligent classifier	intelligent geometric computing
intelligent infrastructure	Kernel learning	K-means
latent dirichlet allocation	latent semantic analysis	latent variable

layered control system	learning automata	link prediction
logitboost	long short term memory (LSTM)	lpboost
machine intelligence	machine learning extreme machine learning	machine translation
machine vision	madaboost	MapReduce
Markovian hidden Markov model	memetic algorithm	meta learning
motion planning	multi task learning	multi-agent system
multi-label classification	multi-layer perceptron	multinomial naïve Bayes
multi-objective optimisation	naïve Bayes classifier	natural gradient
natural language generation	natural language processing	natural language understanding
nearest neighbour algorithm	neural network artificial neural network convolutional neural network deep convolutional neural network deep neural network recurrent neural network	neural turing neural turing machine
neuromorphic computing	*non negative matrix factorisation*	object detection
object recognition	*obstacle avoidance*	pattern recognition
pedestrian detection	policy gradient methods	Q-learning
random field	random forest	rankboost
recommender system	regression tree	reinforcement learning
relational learning statistical relational learning	*robot* *biped robot* *humanoid robot* *human-robot interaction* *industrial robot* *legged robot* *quadruped robot* *service robot* *social robot* *wheeled mobile robot*	*rough set*
rule learning rule-based learning	self-organising map	self-organising structure
semantic web	semi-supervised learning	*sensor fusion* *sensor data fusion* *multi-sensor fusion*
sentiment analysis	similarity learning	simultaneous localisation mapping
single-linkage clustering	sparse representation	spectral clustering
speech recognition	speech to text	stacked generalisation
stochastic gradient	supervised learning	support vector regression
swarm intelligence	swarm optimisation particle swarm optimisation	temporal difference learning
text mining	text to speech	topic model
totalboost	trajectory planning	trajectory tracking
transfer learning	trust region policy optimisation	*unmanned aerial vehicle*
unsupervised learning	variational inference	vector machine support vector machine
virtual assistant	*visual servoing*	xgboost

Note: When accounting for combinations of keywords in a given abstract, only keywords belonging to different groups are considered. Keywords in italics were identified as somewhat general by patent examiners. For the AI-patent search, these words are only included in combination with IPC or CPC classes.

Annex Table C.2. List of IPC Codes

C.2.1. IPC codes only.

G06N3	G06N5	G06N20	G06F15/18
G06T1/40	G16C20/70	G16B40/20	G16B40/30

C.2.2. IPC codes to be combined with keyword search

G01R31/367	G06F17/(20-28, 30)	G06F19/24	G06K9/00
G06K9/(46-52, 60-82)	G06N7	G06N10	G06N99
G06Q	G06T7/00-20	G10L15	G10L21
G16B40/(00-10)	G16H50/20-70	H01M8/04992	H04N21/466

Annex Table C.3. List of CPC codes

CPC codes to be combined with keyword search

A61B5/(7264,7267)	B29C2945/76979	B29C66/965	B60G2600/(1876-1879)
E21B2041/0028	F02D41/1405	F03D7/046	F05B2270/(707-709)
F05D2270/(707-709)	F16H2061/(0081-0084)	G01N2201/1296	G01N29/4481
G01N33/0034	G01R31/367	G01S7/417	G05B13/(027-029)
G05B2219/33002	G05D1/0088	G06F11/(1476,2257,2263)	G06F15/18
G06F17/(20-28)	G06F19/(34,707)	G06F2207/4824	G06K7/1482
G06K9/00	G06K9/(46-52, 60-82)	G06N3	G06N5
G06N7	G06N10	G06N20	G06N99
G06Q	G06T2207/(20081,20084)	G06T3/4046	G06T7/(00-20)
G06T9/002	G08B29/186	G10H2250/(151,311)	G10K2210/(3024,3038)
G10L15	G10L21	G10L25/30	G11B20/10518
G16B40	G16C20/70	G16H50/(20,70)	H01J2237/30427
H01M8/04992	H02P21/0014	H02P23/0018	H03H2017/0208
H03H2222/04	H04L2012/5686	H04L2025/03464	H04L25/(0254,03165)
H04L41/16	H04L45/08	H04N21/(4662-4666)	H04Q2213/(054,13343,343)
H04R25/507	Y10S128/(924-925)	Y10S706	