Organisation for Economic Co-operation and Development

DSTI/STP/NESTI(2019)1

For Official Use **English - Or. English**

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION
COMMITTEE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY

Working Party of National Experts on Science and Technology Indicators

Identifying government funding of AI-related R&D projects

An initial exploration based on US NIH and NSF project funding data

This work has been carried out in the context of a feasibility study by NESTI to use project-level funding data as a basis for conducting fine-grained analysis of government funding of scientific R&D (Fundstat). This particular application is also intended to contribute to informing the monitoring of the recently agreed OECD Council Recommendation on Artificial Intelligence (AI), which states that governments "should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI".

A previous version of this document was distributed as a room document at the NESTI meeting held in October 2018. This version is presented for further feedback and methodological refinement. Delegates are invited to submit comments (preferably by 30 August) on the methodology of the study and its potential applicability to national funding databases across different countries, with a view to augmenting the scope and relevance of the work.

For further information, please contact: Economic Analysis and Statistics Division.
Izumi Yamashita (Izumi.Yamashita@oecd.org); Fernando Galindo-Rueda (Fernando.Galindo-Rueda@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.

Table of contents

Executive summary 4

Identifying government funding of AI-related R&D projects.. 5

1. Introduction and background 5

2. Methodology.. 7

2.1. Project funding data 7
2.2. Operational definition of AI... 10
2.3. AI document retrieval methodology 11
2.3.1. Selecting key terms 11
2.3.2. Tagging documents with the list of AI key terms 19

3. Results.. 21

3.1. Distribution of AI key terms... 21
3.2. Estimates of AI-related R&D projects and funding.. 23
3.2.1. NIH funding 23
3.2.2. NSF funding.. 25
3.3. AI research across different R&D domains and application areas 27
3.4. Concentration of AI funding... 29
3.5. Bias and robustness checks.. 30
3.5.1. Incidence of false positives (precision error) 30
3.5.2. Incidence of false negatives (recall error) 32
3.5.3. General assessment 33
3.5.4. Robustness analysis. Comparing to other lists 34

4. Conclusions and next steps 35

References 38

Annex. Lists of key terms... 39

Tables

Table 2.1. Fields included in the NIH project data analysed.. 8
Table 2.2. Fields included in the NSF project data analysed... 9
Table 2.3. MeSH tree structure for Artificial Intelligence 12
Table 2.4. AI term list from Cockburn et al. (2018)... 13
Table 3.1. Most frequent AI key terms in the NIH funding database 21
Table 3.2. Most frequent AI key terms in the NSF funding database 22
Table 3.3. Estimates of AI-related R&D in NIH funding 24
Table 3.4. Estimates of AI-related R&D in NSF funding 26
Table 3.5. Precision analysis of AI detection results in NIH and NSF data............................ 31
Table 3.6. Recall analysis of AI detection results in NIH and NSF data 33

Table A.1. Clustering and treatment of AI-related terms in the AI-journal corpus.................. 39
Table A.2. Selected list of AI key term 42

Figures

Figure 2.1. Outline of AI key term identification and tagging procedure 14
Figure 2.2. Cluster representation of core AI key terms from M- and C-lists........................... 15
Figure 2.3. Cluster representation of core and additional AI terms.. 16
Figure 2.4. Clustered vector representations of AI terms in the NIH funding corpus.............. 17
Figure 2.5. Clustered vector representations of AI terms in the NSF funding corpus.............. 18
Figure 2.6. Key term selection based on core list of AI terms 19
Figure 3.1. Occurrence of selected AI key terms in NIH and NSF data 23
Figure 3.2. Estimates of AI-related NIH-funding... 24
Figure 3.3. Funding amounts per application in NIH project funding data.............................. 25
Figure 3.4. Estimates of AI-related NSF-funding 26
Figure 3.5. Estimated NIH and NSF funding for AI-related R&D, 2001-2017 27
Figure 3.6. Estimated share of AI-related R&D funding within NIH institutes........................ 28
Figure 3.7. Estimated share of AI-related R&D funding within NSF disciplinary directorates 29
Figure 3.8. Funding concentration of AI-related and non AI-related projects 30
Figure 3.9. Illustration of sensitivity of results to using alternative AI term lists 35

Boxes

Box 3.1. Excerpts from sample of projects automatically retrieved as AI-related.................... 32

Executive summary

This document reports on the procedures and initial findings from an experimental text-based analysis of project-level R&D funding data, focused on measuring the extent and features of government R&D support for Artificial Intelligence (AI-related R&D). The study uses a quantitative case study approach, applying a set of text mining tools to specific project funding databases to identify AI-related R&D. The project-level funding data of the US National Institutes of Health (NIH) and US National Science Foundation (NSF), both openly available, provide useful and relevant ground for demonstration purposes. R&D project funding databases, while not comprehensive of total government R&D funding, can be used to trace and estimate a sizeable part of total government funding of AI related R&D, helping inform the recent OECD Council Recommendation on AI calling for increased government support for R&D in this area.

This study adopts a "key-terms" selection and matching approach for identification of AI-related R&D projects in the text corpus of funded projects by each organisation. The task is to predict a category using text data (project titles and abstracts) in the absence of a representative set of labelled data. The process involves: (1) Key term selection and categorisation on basis of AI-relatedness (core and non-core terms). (2) Matching key terms to documents, applying a selection rule to categorise documents that as AI-related, based on the presence of core and non-core terms.

The key term selection sets out to deliver a comprehensive list of AI-relevant key terms for matching. As there is no formal, already tagged representative training database for the population of funded projects, the process requires an initial (seed) list of terms that can be posited to be AI-relevant. The set of potential key terms is enriched by means of text analysis applied to a separate but relevant body of scientific publication abstracts in first instance and then to the project funding text corpora that are the object of this study. These corpora are analysed through a two-layer neural network model (Word2vec) designed to reconstruct linguistic contexts of terms. This method "embeds" all of the terms in each corpus into a multi-dimensional vector so that terms that occur in similar contexts have a similar digital "fingerprint". This process, which requires substantial human intervention in order to identify ambiguous terms, results in 104 terms (35 core and 69 non-core).

A simple selection matching process whereby a document is categorised as AI-related if it contains at least one core term or two non-core terms generates the list of AI projects. The results confirm the widely held view that AI-R&D funding has been rapidly increasing in both organisations. In 2017, AI-related funding appears to represent 10% of total NSF funding and 2.5% of NIH funding. Together, NIH and NSF dedicated USD 1.2bn in 2017 to AI-related R&D, namely 3.1% of their combined funding, up from 0.6% in 2001. The analysis also shows that for some areas with very large projects, it is important to consider monetary measures as well as count-based indicators such as those that may be derived from bibliometric analysis of scientific outputs. This work finds that outside the core funding for data infrastructure and computer science, where AI funding rates are very high, significant rates of AI funding can also be found within engineering (NSF) as well as within health applications requiring large amounts of data such as genomics and neurological research (NIH). A number of suggestions for future work are also provided in this paper, pointing at the possibility of carrying out similar, potentially federated analysis, across agencies and countries.

Identifying government funding of AI-related R&D projects

1. Introduction and background

1. This document reports on the procedures and initial findings from an experimental text-based analysis of project-level R&D funding data, focused on measuring the extent and features of government R&D support for Artificial Intelligence (AI-related R&D). The study uses a quantitative case study approach, applying a set of quantitative tools to identify AI-related R&D in specific project funding databases, for demonstration and illustration purposes. The project-level funding data of the US National Institutes of Health (NIH) and US National Science Foundation (NSF), both openly available databases for download and analysis, provide useful and relevant ground for demonstration purposes, as both organisations are among the largest world and US research funding organisations and leading sponsors of AI technology development.

2. The purpose of this study is twofold:

- To contribute to informing the policy discussion on public support for AI-related R&D, helping to understand the transformational role of AI as a general purpose technology that can also enable R&D and innovation in different scientific domains and application areas.
- To support the ongoing OECD project seeking to assess the feasibility of constructing a multi-country database on project funding for analytical purposes ("Fundstat"), including the identification of emerging scientific and technology domains.

3. Project level data can be an extremely rich source of information, particularly if it is comprehensive or at least representative of government funding. This depends on the extent to which project-based funding is the norm as well as the accessibility of such data. Data quality is a major consideration underpinning the decision on whether project funding databases are suitable for statistical analysis. Because of the administrative nature of the data, it should be borne in mind what purposes (by funders and applicants) underpin the data generation process. The depth and breadth of the information provided by applicants will depend on the incentives and constraints that they face. Tagging of projects by either applicants or administrators is also potentially subject to human error and inconsistency. Thus, beyond the concrete application to AI as a research subject, this work seeks to address the widespread demand for data resources, tools and methods that help identify features of R&D funding in thematic areas that are not easily captured by pre-defined and difficult-to-change taxonomies. This exercise is furthermore a demonstration of the possibilities of AI for text analysis as a potential tool for statistical measurement.

4. AI is indeed at the top of policy agendas for stakeholders and governmental institutions at both national and international levels, and reflected in the adoption by

the OECD Council of a Recommendation on AI[1,2]. As a result, this work is set within a wider range of OECD efforts in this area[3] and in particular to measure AI in scientific articles, patents, and other relevant data (OECD, 2019 forthcoming[1]). The OECD Council Recommendation explicitly states that governments "should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI […]".

5. In order to contribute to monitoring how OECD countries and partner economies invest in AI R&D, this study's primary operational objective is to demonstrate whether and how it is possible to identify AI-related projects, i.e. projects whose text description renders themselves suitable to be classified as seeking advances in AI or making an explicit and non-trivial use of AI systems to achieve their objectives. Based on the outcome of the (approximate) identification of the full set of AI-related projects in a corpus, the statistical goal is to estimate the volume and share of projects and funding amounts that fit into this category.

6. There is a wide and fast growing literature dealing with topic extraction on research fields drawing on several corpora, mostly publications, with some efforts looking at AI in particular, as documented in (Cockburn et al., 2018[2]) and previous OECD work aimed at identifying and measuring Artificial Intelligence (AI)-related developments in: science, as captured in scientific publications; technological developments, as proxied by patents; and software, and in particular open source software (OECD, forthcoming). The increasing public availability of project-level funding data, often set in the context of public transparency measures, is also enabling related efforts looking specifically at data about R&D funding. Funding organisations and a growing number of commercial providers of research support services have been not only compiling and offering access to data but also providing semantic search and analytical functionalities (Bode et al., 2018[3]).

7. Some studies have looked in general at the classification of NIH funding (Park et al., 2016[4]; Talley et al., 2011[5]), NSF (Kawamura et al., 2018[6]; Freyman, Byrnes and Alexander, 2016[7]), and EU's FP7 (Kawamura et al., 2018[6]), but not on AI-related research specifically. A number of policy documents have drawn attention to the level of efforts made at advancing research on AI. A White House report indicated that the United States invested USD 1.1 billion in "AI R&D" in 2015 (NSTC, 2016[8]). It has been estimated that the EU has spent around 13% of R&D budget in ICT since 2014 (EC, 2018[9]). The Engineering and Physical Sciences Research Council (EPSRC) of the UK has allocated GBP 300 million to fund research related to data science and AI (BEIS and DCMS, 2018[10]).

8. This document is structured as follows. **Section 2** describes the data used for analysis and the methodology applied to identify AI-related R&D funding in the absence of a proper training database where projects have been comprehensively ex-ante rated as AI-related. In a nutshell, this paper adopts a relatively simple keyword or "key-term" selection methodology which results in the body of documents being tagged as AI relevant or not. The procedure relies on natural language processing

[1] See http://www.oecd.org/going-digital/ai/initiatives-worldwide/

[2] See https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[3] See https://www.oecd.org/going-digital/ai/

models to select a comprehensive list of candidate keywords, taking as starting point a baseline expert list and expanding by reference to various corpora, seeking to minimise the risk of ignoring relevant AI projects. This is complemented with a human-informed procedure to penalise potentially ambiguous terms in the ultimate selection of AI documents. **Section 3** presents the key results for NIH and NSF funding, reporting fast growing levels and rates of AI-related funding, providing additional evidence on the distribution of AI related funding. Robustness checks are also presented. **Section 4** concludes with a series of possible next steps, including potential efforts to more fully automatise the process and to extend the work in a distributed fashion to other project funding databases across OECD countries.

2. Methodology

2.1. Project funding data

9. The two organisations under study offer each relevant insights on the methodology of project description analysis and the role of AI in research. NIH funding accounts for approximately 26% of US federal government funds for R&D (obligations[4]) in 2017 and NSF was 4.8 % (National Science Foundation, 2018[11]). It is important to put in a broader context the analysis of project funding data for any given organisation or group of them, in order to illustrate what type of R&D funding is excluded. In this case, the omission of other agencies results in the exclusion of significant levels of funding in the areas of defence, aerospace, energy and agriculture, as well as R&D funding provided by the individual states.

10. The data used in this study come from the two organisations, but the information is originally provided to them by different types of organisations and individuals as these submit. Only successful applications are available.[5] Both organisations operate primarily in the basic and applied research space of the R&D spectrum, fostering advances in fundamental knowledge and research into potential applications without going all the way to fund products or process that can be ready for commercialisation. Both organisations provide financial support in the form of grants, cooperative agreements and contracts, making extensive use of peer review as a resource allocation mechanism.

11. The **NIH** is the largest biomedical research funding organisation in the world. Its work is defined by human health being the main intended objective of all research its funds. NIH is made up of 27 institutes and centres (ICs), 24 of which can provide grant awards. These ICs award more than 80% of the NIH budget each year to support

[4] In the United States, obligations represent legally binding commitments by the government that will result in outlays or payments. Obligations follow the appropriation process in the government budget. Data on appropriations inform the most up to date R&D budgetary reported in the GBARD estimates reported by OECD.

[5] Rejected proposals are not included unless they turned out to be successful in raising funding at a later stage. The comparison of selected and rejected proposals can also be highly informative for a range of policy analysis purposes.

investigators across universities, medical schools, and other research organisations around the world. About 10% of the NIH's budget supports intramural research and scientific activity. The human health dimension of NIH's mission and research funding portfolio implies that the AI dimension within projects will be instrumental and a key driver of information contained in project description abstracts.

12. The **NSF** is the principal US federal agency in charge of supporting civil R&D across all fields of fundamental science and engineering, with the exception of medical sciences which are the competence of NIH funding.[6] Unlike NIH, NSF does not have its own intramural R&D activity. As the organisation responsible for funding research on Computer and Information Science and Engineering and Engineering, NSF supports several projects pushing the boundaries of AI, in addition to supporting projects across several disciplines that may make varying use of AI in their work.

13. This analysis deals with data on grant funding for projects in the NIH RePORTER database and in the NSF Award Search database between FY2001 to FY2017 for NIH and FY2001 to FY2018 for NSF. In FY2017, NIH was estimated to have been responsible for USD 30.7bn worth of obligations for R&D, with an outlay of USD 31.1bn; while NSF accounted for USD 5.6bn worth of funding (obligations), with a corresponding outlay of USD 5.2bn.[7]

Table 2.1. Fields included in the NIH project data analysed

NIH RePORTER item	Description
Application ID	Identification number for an application. Application is unit of observation in the database; a research project may contain multiple associated applications.
Project number	Identification number for a project.
Project title	Title of the project
Fiscal year	A 12 months period in which funding to the application is noticed.
Funding amount	The total amount of money allocated to the project by the government agency, in USD.
Abstract	Description of the contents of the project
Project terms	Terms that represent the characteristics of the project, automatically tagged after 2008.
NIH spending categories	Classification based on Research, Condition, and Disease Categorization (RCDC) taxonomy
Funding mechanism	The instrument through which the funding to the project is made
PI ID(s)	Identification number(s) for the principal investigator(s) of the project
PI Name(s)	Name(s) of the principal investigator(s) of the project
Organisation name(s)	The name(s) of organisation to which the PI(s) belong or the funding is distributed

Source: NIH RePORTER.

14. The data were downloaded from the NIH ExPORTER and the NSF Award Search website over multiple days in December 2018, with a total number of observations of 1 254 139 for NIH and 216 166 for NSF. The unit of observation is a

[6] Other major US R&D funding organisations with R&D funding levels higher than NSF are the Department of Defense (DOD, the largest by far in terms of funding, NASA and the Department of Energy (DOE).

[7] See https://ncsesdata.nsf.gov/fedfunds/2016/html/ffs2016_dst_004.html

successful / awarded "application" for NIH[8] and a successful / awarded project for the NSF database. In both cases, the data contain information provided in the project application or generated in the administrative process such as application/award ID and project number, project descriptions (e.g. project title, fiscal year, funding amount, abstract, and funding mechanism), and beneficiary information (PI ID, PI name, organisation name). Major attributes for the NIH RePORTER data are as presented in **Table 2.1** and for the NSF Award Search as in **Table 2.2**.

Table 2.2. Fields included in the NSF project data analysed

NSF Award Search Items	Descriptions
Award ID	Identification number for an award.
Award title	Title of the awarded project.
Award effective date	The date in that the award is effective.
Award expiration date	The date in that the award is expired.
Award amount	The total amount of money allocated to the project by the government agency, in USD.
Abstract narration	Description of the contents of the project
Award instrument / Program element	The instrument through which the funding to the project is made
Investigator Name(s)	Name(s) of the investigator(s) of the project
Institution(s)	The name(s) of organisation to which the investigator(s) belong or the funding is distributed

Source: NSF Award Search.

15. The embedded tagging of available records is an important source of information. Yet the thematic classification item in the NIH data (NIH spending categories), which is based on the Research, Condition, and Disease Categorization Process (RCDC) is not on its own an appropriate basis for identifying AI-related research because it is health objective-oriented and not informative about actual research methods. AI-related terms do not appear to be used comprehensively and tagging has not been consistent over the time period, as projects were manually tagged from 2001 to 2007 and automatically from 2008 to 2016. Furthermore, some documents do not include project description terms. In the NSF data, thematic classification items and project terms are not available.

16. Because of the lack of systematic tagging information within the NIH and the NSF databases that is relevant to the objectives of this study, the analysis focuses on text information included within project title and abstract[9]. By not depending on tagging features specific to any given database, this approach renders itself relevant for potential application to other project funding databases.

[8] In the NIH case, a given project may have different granted applications, reflecting applications at different points in time, for example to secure renewals and extensions.

[9] Although many of NIH data contain "project terms", the analysis does not rely on them in order to use a common methodology for the two funding agencies.

2.2. Operational definition of AI

17. Definitions guide statistical measurement work in different ways and their operationalisation will vary according to the type of data used. Because this exercise uses existing administrative data for a secondary identification/classification purpose that differs from their original agency-specific grant management purpose, the role of a definition for AI in this context is purely indicative and aimed at helping inform the consistency of the selection approach (the procedures used and their outcomes) with such definition.[10,11]

18. The analysis of taxonomic systems used by the funding organisations under analysis revealed the existence of an NIH definition for Artificial Intelligence as the "*theory and development of computer systems which perform tasks that normally require human intelligence. Such tasks may include speech recognition, learning; visual perception; mathematical computing; reasoning, problem solving, decision-making, and translation of language* (NIH MeSH)". This definition, extracted from the NIH MeSH (Medical Subject Headings) is a hierarchically organised set of keywords set managed by one of the NIH institutes (U.S. National Library of Medicine). It uses a list of potential application tasks and makes explicit reference to the concept of intelligence without defining it. The reference to "normally require HI" is indicative of the potential subjectivity and context-dependence of the concept. Over time and with growing levels of automation, a number of tasks will ultimately cease to be considered as AI depending on who is making the judgement.

19. The Advisory Expert Group at OECD on Artificial Intelligence (AIGO) has recently defined an AI system in a somewhat less context-dependent fashion as *a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. It does so by using machine and/or human-based inputs to: i) perceive real and/or virtual environments; ii) abstract such perceptions into models through analysis in an automated manner (e.g. with machine learning, or manually); and iii) use model inference to formulate options for information or action. AI systems are designed to operate with varying levels of autonomy*.

20. Both definitions underpin the work in this report, whose intention is not to propose new definitions of AI but to identify funded projects that relate to AI in the sense that they either make use of AI systems or contribute to their development from a theoretical or practical perspective. This is what is implied by the rather broad notion of "relatedness". Further refinements in the methodology might ultimately allow to differentiate more reliably between different types and grades of relatedness, in particular separating between research that makes use of available AI systems and tools (instrumental relatedness), research that seeks to develop brand new AI concepts, theories and tools (output relatedness), as well as possible instances when both are combined.

[10] A designed survey-based approach, for example, would use a definition as departing point for developing and testing question items aimed at implementing the chosen definition. A definition may also be offered to survey respondents to inform and support their own information retrieval and response processes.

[11] Ultimately, measurement and administrative purposes might align, although this may be realised too late or occur over a limited period of time.

2.3. AI document retrieval methodology

21. After consideration of several options, the analysis in this report adopts a "key-terms" (sometimes referred to as keyword for simplicity) selection and tagging approach for identification of AI-related R&D projects in the text corpus of funded projects for each organisation. The fundamental task is one of predicting a category using text data in the absence of labelled data. Project data (or a representative subsample of it) are not tagged as AI-related. The potential use of a fully unsupervised topic modelling – the identification of hidden semantic structures - in the full corpus of projects was discarded after an initial attempt. Examination of the results of conventional topic modelling schemes showed that topics proved difficult to associate with the notion of AI-relatedness as the AI signal was relatively weak, especially in the case of NIH data. In that database, topics are often dominated by the semantic weight of the objectives of the research such as health over the scarcer information about potential AI-related methods used in the research. Topic identification from internal or external linkages to other data (e.g. citations) was not possible either. With further refinements, it may be ultimately possible to revert to this type of modelling. For the time being, a somewhat simpler key term selection and tagging procedure has been followed.

22. Key term matching presents a number of challenges, since there is at present no consensus on a standard set of key terms available to comprehensively and unambiguously represent AI-related R&D and such standard is bound to vary over time. Failing to capture all possible relevant key terms that identify AI poses a significant underestimation risk. This problem can also arise when the title and abstract of project applications do not contain detailed information on the research methodology to be used in a given project. This is an underlying data problem which may be particularly acute when the abstracts focus on outlining the expected outcomes and AI plays an enabling role within a project (the notion of instrumental relatedness alluded to above). Ideally, the underlying project text corpus available for analysis should contain a "methods" section to facilitate a more effective data mining process as well as a better understanding of the role played by AI in various R&D fields.

23. Conversely, effective key term matching from a predefined menu of AI terms does not assure that the research is ultimately AI-related. Potential AI key terms may describe research paradigms or domains that do not necessary relate to AI. A common example is the term "neural network" that was "borrowed" by computer scientists from the neurosciences. Indeed, several terms in the AI domain are based on analogies to concepts used from other life science domains. Furthermore, the use of terms relating to statistical analysis tools known for several decades and recently popularised and adapted for AI applications (e.g. Markov decision processes) may result in non AI-related projects being mistakenly identified as AI related.

24. The process involves two main steps: (1) Key term selection and categorisation on basis of AI-relatedness; (2) matching key terms to documents and applying an AI-relatedness selection rule to documents and selecting documents that are AI-related. These two steps are complemented by additional bias and robustness checks.

2.3.1. Selecting key terms

25. The key-terms selection process is in its own right a multi-step process that sets out to deliver a comprehensive list of AI-relevant key terms for matching to the

documents in the NIH and NSF databases and undertaking a selection of documents. As there is no formal training database, the process requires an initial (seed) list of terms that can be posited to be AI-relevant according to the definition provided above.

Core AI terms

26. As previously noted, the MeSH taxonomy contains a heading for "**Artificial intelligence**". **Table 2.3** provides a description of the position of AI within its hierarchical structure. AI features in two separate MeSH domains: "Mathematical concepts" and "Information science". Subject subheadings include Biological Ontologies, Computer Heuristics, Expert Systems, Fuzzy Logic, Gene Ontology, Knowledge Bases, Machine Learning (including Supervised Machine Learning and Unsupervised Machine Learning), Natural Language Processing, Neural Networks (Computer), Robotics and Support Vector Machine. This structure does not provide alone a comprehensive source of all potentially relevant AI terms but provides a basic structure for the categorisation of research activity and outputs in the health domain. Unfortunately, the MeSH research tagging system is not yet applied either manually or automatically to funding applications and therefore such information is not available within the RePORTER data system. MeSH is principally applied to scholarly publications listed in MEDLINE, PUBmed and related sources, while the funding data do not contain readily available information on publication outputs associated to the funded projects.

Table 2.3. MeSH tree structure for Artificial Intelligence

Mathematical Concepts [G17]						
	Algorithms [G17.035]					
		Artificial Intelligence [G17.035.250]				
			Machine Learning [G17.035.250.500]			
				Supervised Machine Learning [G17.035.250.500.500]		
					Support Vector Machine [G17.035.250.500.500.500]	
				Unsupervised Machine Learning [G17.035.250.500.750]		
Information Science [L01]						
	Computing Methodologies [L01.224]					
		Algorithms [L01.224.050]				
			Artificial Intelligence [L01.224.050.375]			
				Computer Heuristics [L01.224.050.375.095]		
				Expert Systems [L01.224.050.375.190]		
				Fuzzy Logic [L01.224.050.375.250]		
				Knowledge Bases [L01.224.050.375.480]		
					Biological Ontologies [L01.224.050.375.480.500]	
					Gene Ontology [L01.224.050.375.480.500.500]	
				Machine Learning [L01.224.050.375.530]		
					Supervised Machine Learning [L01.224.050.375.530.500]	
						Support Vector Machine [L01.224.050.375.530.500.500]
					Unsupervised Machine Learning [L01.224.050.375.530.750]	
				Natural Language Processing [L01.224.050.375.580]		
				Neural Networks (Computer) [L01.224.050.375.605]		
				Robotics [L01.224.050.375.630]		

Source: U.S. National Library of Medicine, NIH. Extracted on 28 September 2018 from https://meshb.nlm.nih.gov/record/ui?ui=D001185.

Table 2.4. AI term list from Cockburn et al. (2018)

Symbols	Learning	Robotics
natural language processing	machine learning	computer vision
image grammars	neural networks	robot
pattern recognition	reinforcement learning	robots
image matching	logic theorist	robot systems
symbolic reasoning	bayesian belief networks	robotics
symbolic error analysis	unsupervised learning	robotic
pattern analysis	deep learning	collaborative systems
symbolic processing	knowledge representation and reasoning	humanoid robotics
physical symbol system	crowdsourcing and human computation	sensor network
natural languages	neuromorphic computing	sensor networks
image alignment	decision making	sensor data fusion
optimal search	machine intelligence	systems and control theory
	neural network	layered control systems

Source: Cockburn et al. (2018), The Impact of Artificial Intelligence on Innovation, http://www.nber.org/papers/w24449.

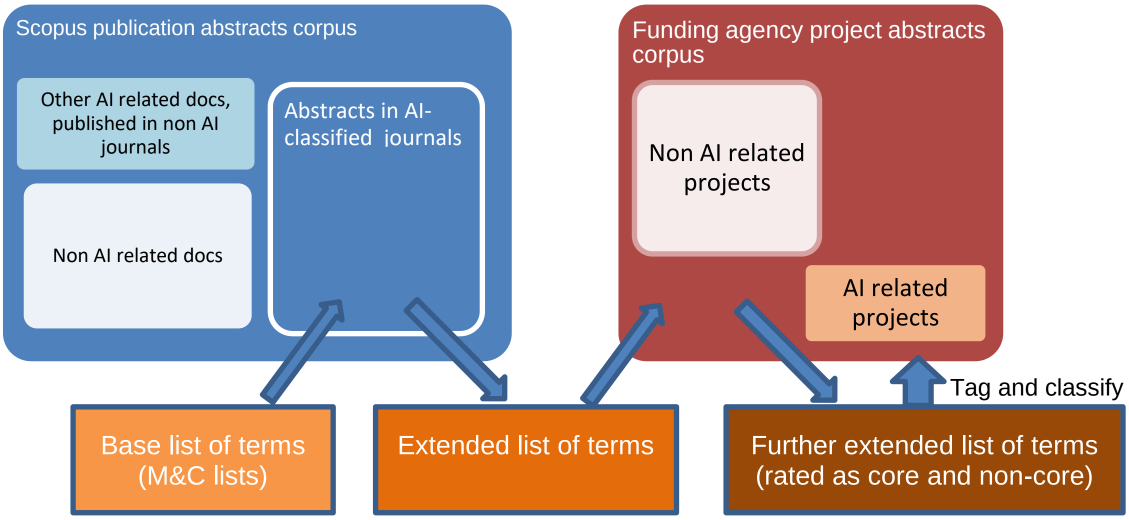
27. For this reason, the MeSH list of terms (the "M-list" from here onwards) is enhanced with an AI key term list produced by (Cockburn et al., 2018[2]), who analyse academic papers and patent documents to measure the impact of AI on innovation and derive a list of key terms related to AI as basis for their analysis. This list ("C-list") contains 38 terms classified into three categories (Symbols, Learning, and Robotics) and is reproduced in **Table 2.4**. The C-list is more comprehensive than the M-list, although the former contains terms that are not uniquely associated to AI and may therefore imply a lower degree of precision.

Extending the set of potential key terms

28. Combined into one, the M- and C- lists provide a basis for an initial set of key terms.[12] However, before proceeding further, additional work is required to minimise the risks of recall and precision errors. In particular, a process is required to retrieve additional key terms that provide relevant signals of AI-related research activity and minimise recall bias. This procedure is outlined in **Figure 2.1**.

[12] For further analysis to be described below, the terms in the combined M-C list been converted to lower case text and have been lemmatised, i.e. variations of the same terms have been converted to a single item (e.g. sees, saw, seeing, seen to see or books to book). The terms "supervised machine learning" and "unsupervised machine learning" have been converted to "supervised learning" and "unsupervised learning" as they overlap with "machine learning". The terms "decision making" and "natural languages" have been removed from the core combined M-C list as potentially ambiguous. Furthermore, in order to reduce the level of noise in the word embedding process, a number of terms with very low incidence in all the corpora used for text analysis, a few candidate terms are removed, namely "biological ontologies", "computer heuristics", "crowdsourcing and human computation", "image grammars", "layered control systems", "logic theorist", "physical symbol system", "symbolic error analysis", "symbolic processing". Thus, a shortlist of 32 candidates of AI key terms are retained.

Figure 2.1. Outline of AI key term identification and tagging procedure



Note: This figure provides a schematic representation of the procedure, which departs from the base list of terms extracted from the M and C-lists, the retrieval of terms used in similar contexts within the corpus of scientific publications in AI journals, and the ultimate applications of the definitive list of AI terms (which have been graded / scored according to their potential ambiguity) for tagging the documents in the project abstracts in the funding database corpus. A key limitation of this process is that it is not possible to learn about the distinctive features of AI-related science published in non AI journals unless the patterns are also present in the documents contained in the AI journals.

29. A reasonable data corpus for retrieving an extended set of baseline terms is the body of scientific publications (articles, conference proceedings and reviews) featured in journals and dissemination vehicles known to focus on AI. The Scopus Custom database used at OECD provides titles and abstracts for 713 016 documents published between 2001 and 2017 assigned to the All Science Journal Classification (ASJC) codes corresponding to Artificial Intelligence (ASJC1702) and Computer vision and pattern recognition (ASJC1707).[13]

30. The entire data corpus was cleaned[14] and tokenised (i.e. separated into 1 to 4-grams such as "robot", "deep learning", "natural language processing" and "knowledge representation and reasoning") . The tokens have been vectorised to be compared in a vector space by mathematical measures of similarity, e.g. cosine-based. This process generates a distributed representation of words or terms (also called "word embeddings" or continuous space representation of words). This has become a popular way of capturing distributional similarity (lexical, semantic or even syntactic) between different words based on co-occurrence patterns. The basic idea is to represent each word in a vocabulary with a real-valued vector of some fixed dimension.

31. This paper analyses the scientific publications "corpus" through a model of two-layer neural networks (Word2vec) that are trained to reconstruct linguistic

[13] The Scopus database does however not provide a full basis for training because there is no in built identification of AI documents outside the corpus of documents published in AI-classified journals.

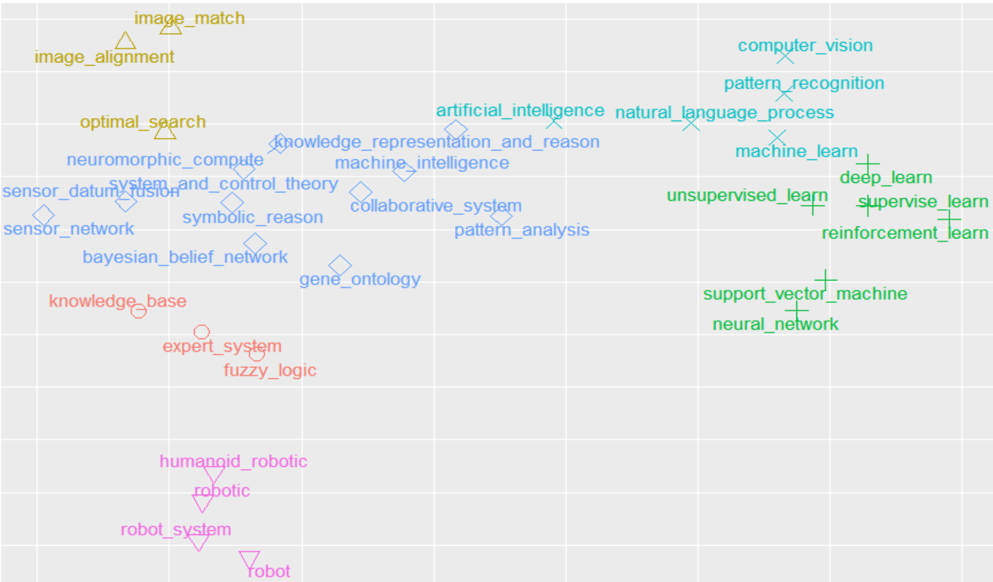
[14] The text was converted into lower case, lemmatised and punctuation was removed, changing all the numbers into "0". Hyphenated terms were also processed.

contexts of words (Mikolov et al., 2013[12]). These "embed" all of the terms in the corpus into a multi-dimensional vector (100 dimensions in this case) so that words that occur in similar contexts have similar embeddings (as captured by mathematical measures of similarity, e.g. cosine-based).

32. It is possible to examine the structure of the vector representations of the core 32 key term candidates through clustering analysis, as shown in **Figure 2.2**. This figure is a two dimensional representation of the proximity of such terms in the 100-dimension vector space, with the terms clustered in 6 groups through a nearest k-means algorithm. This visualisation only provides an indication of internal coherence as implied by the proximities across vector representations for these terms. The cluster represented by circles "○" refer to AI methodologies mostly used up until the early 2000s. The inverted triangles relate to robotics. One cluster (plus "+" signs) refers to types of automatic learning procedures, which interestingly is separate from the cluster that contains the more general machine learning and AI terms as such, as well as pattern recognition, natural language programming and computer vision. It is worth noting that a cluster denoted by upright triangles refer to image analysis occupies a different space to pattern recognition. The diamond cluster is in turn more internally heterogeneous, comprising in one same group statistical concepts used in AI, terms relating to sensors and gene ontology.

Figure 2.2. Cluster representation of core AI key terms from M- and C-lists

Two-dimensional cluster representation based on "term embeddings" in the Scopus AI-journal corpus



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

33. Terms inside the AI-journal Scopus corpus with high[15] cosine similarity to the core terms found in the M-C list are automatically selected as potential candidates to become additional AI- key terms. The procedure yields a total of 171 terms. 45

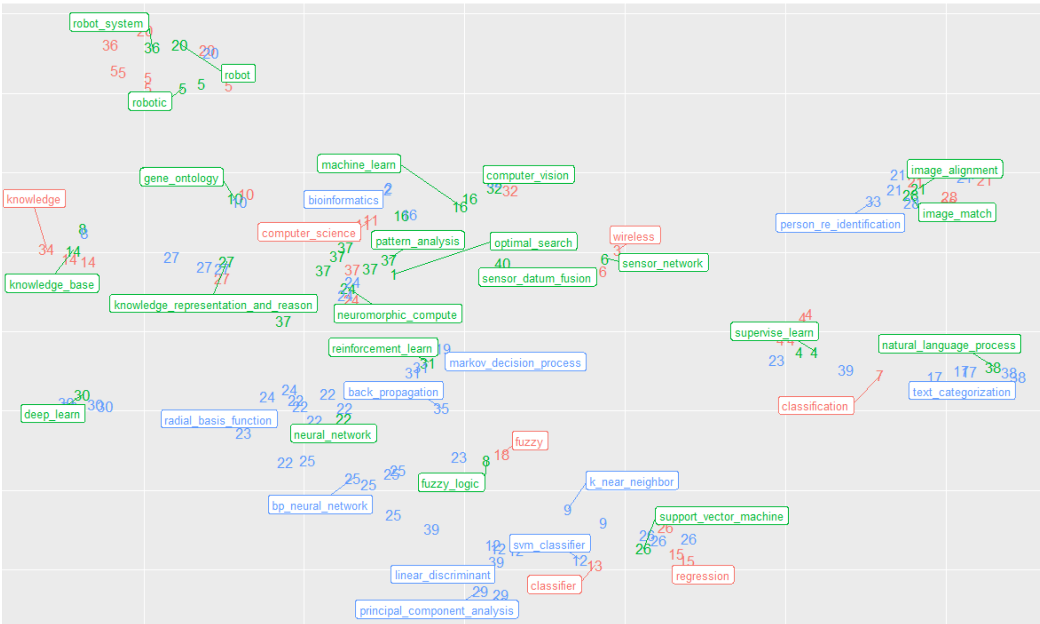
[15] The analysis sets a minimum threshold of cosine similarity as +0.65. As the cosine similarity drops, the likelihood of retrieving terms that have irrelevant meanings rises considerably.

duplicated terms are removed or merged. For example, "machine learning techniques" (duplication with "machine learning") is removed. A total of 27 abbreviations are also removed to avoid duplication. The remaining 99 terms are subject to cluster analysis alongside the original AI key terms.

34. Visualising this larger set of terms on a two-dimensional space is more challenging but still possible by making a more sparing use of labels, as shown in **Figure 2.3**.[16] This extended visualisation differs from the previous one because it represents connections across a broader set of terms, all derived from an entirely AI-related corpus (the AI journals). The visualisation helps identify the connection between core terms and co-occurring terms that may be also likely AI terms as well as those that may be ambiguous in relation to AI if extended to other corpora (e.g. wireless, classification, cognitive science, computer science, wheelchair, mosaicking, and registration). These are manually removed from the list to increase precision and reduce the risk of false positives. This ultimately results in the original list of 32 terms being expanded with 66 additional terms.

Figure 2.3. Cluster representation of core and additional AI terms

Two-dimensional cluster representation based on "term embeddings" in the Scopus AI-journal corpus



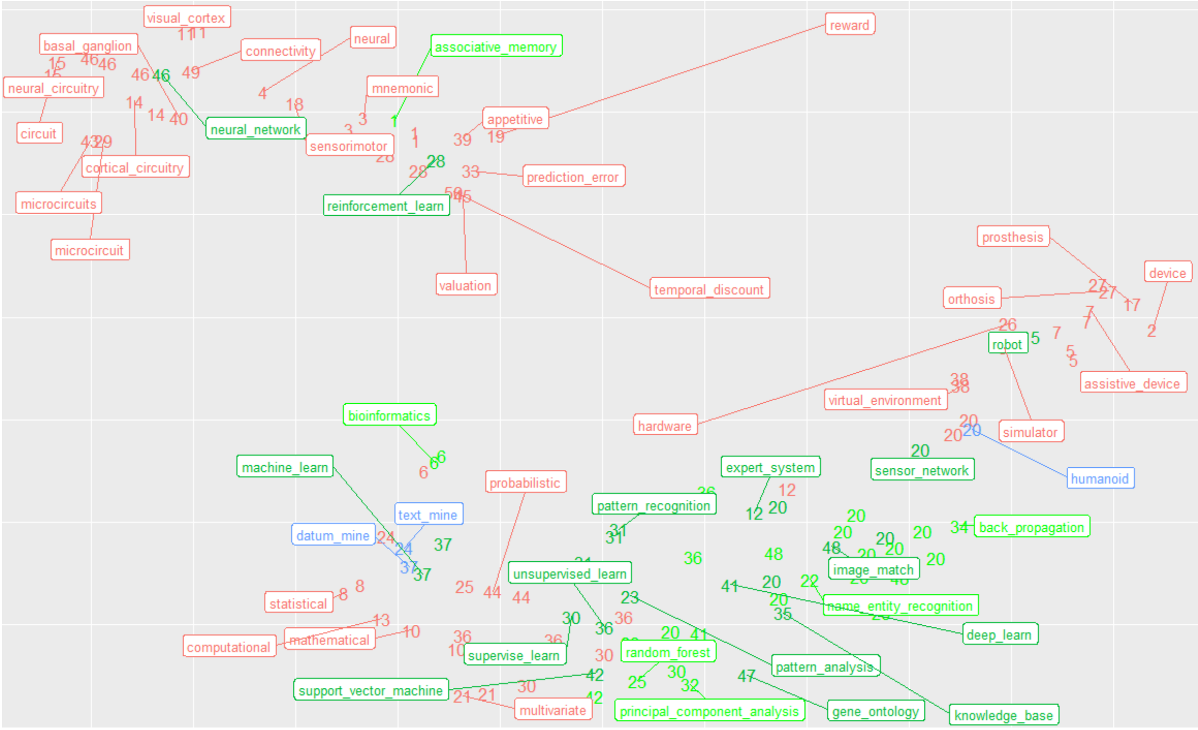
Note: One single label per cluster is presented to facilitate readability. The colour coding is as follows: Green for AI core terms, blue –terms selected as additional AI, and red, for terms removed not considered as AI related in the subsequent scoring procedure.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

35. Drawing on this extended list, the same methodology is applied to the NIH (**Figure 2.4**) and the NSF (**Figure 2.5**) databases in order to both assess the feasibility of identifying additional context specific additional AI key terms and also for the purpose of assessing the interpretability of the key terms from an AI-relatedness perspective. This stage of the analysis results in only 6 additional key terms. Overall,

[16] More detail on the individual terms is available in the annex section (**Table A.1**).

this results in 104 AI key terms, presented in **Table A.2**, which also indicates the corpus from which each term was identified in first instance.

Figure 2.4. Clustered vector representations of AI terms in the NIH funding corpus



Note: One single label per cluster is presented to facilitate readability, giving priority to the AI core terms or in its absence the most frequent label in the cluster. The colour coding is as follows: Green for AI core terms, blue –terms selected as additional AI, and red, for terms removed not considered as AI related in the subsequent scoring procedure.
Source: OECD calculations based on NIH Reporter data, accessed December 2018.

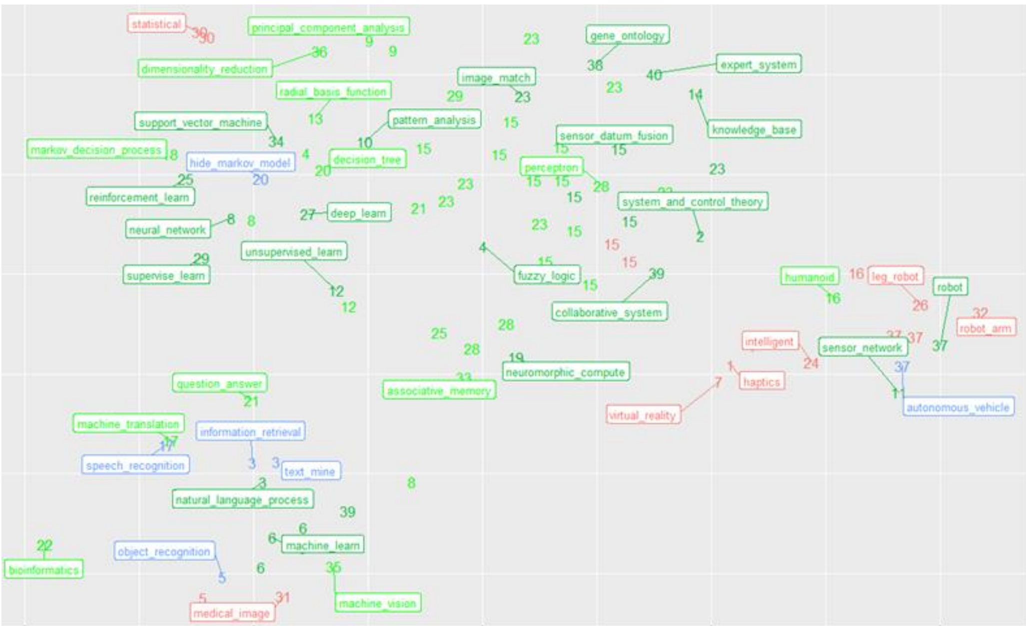
36. In the case of NIH data, the visual representation shows two main groups of potentially AI related terms. The top left hand corner of **Figure 2.4** reveals the high ambiguity for AI tagging purposes of the "neural network" term, which is clustered around neuroscience-related terms. The same applies to two AI terms (reinforcement learning and associative memory) that appear to be mostly found in the context of projects that appear to look into human cognitive processes. Although this might reveal the use of AI in (or connected to) the neuroscience domain[17], high co-occurring terms in this space such as "neural", "prediction error" or "circuit" are therefore not included in the list of key terms for document retrieval, while documents with retained AI terms will only selected if there are other AI signals. Within the region that appears to be less ambiguously related to AI, one broad cluster is defined by devices and hardware that tend to co-occur with references to the term "robot". A cluster of statistical terms is found to gravitate around the term "machine learning". These terms may also be weak signals of AI activity, but since there is a high risk that those terms are being used in

[17] In particular, research providing possible directions for replicating artificially how the human brain operates.

standard epidemiological analysis, they are excluded from the list of AI terms used for document tagging and retrieval.[18]

37. In the case of NSF data, the clustering exercise visualised in 2 dimensions in **Figure 2.5** presents a significantly larger set of AI terms given its coverage of core computer science and a different but similar set of clusters that exhibit a varying degree of ambiguity. The space in the bottom-left region of the figure appears to identify application areas for AI systems dealing with text, speech and image recognition, in turn linking to medical and life sciences applications. The terms in the centre-right region relates to automation and robots, while the top-left region incorporates a section clearly related to various types of machine learning techniques which links to more generic statistical terms. The top-right hand corner incorporates AI strands more detached from machine learning systems (expert systems, ontologies, and knowledge bases). The central space is more sparsely population, with terms related to neuromorphic computing[19] and associative memory.

Figure 2.5. Clustered vector representations of AI terms in the NSF funding corpus

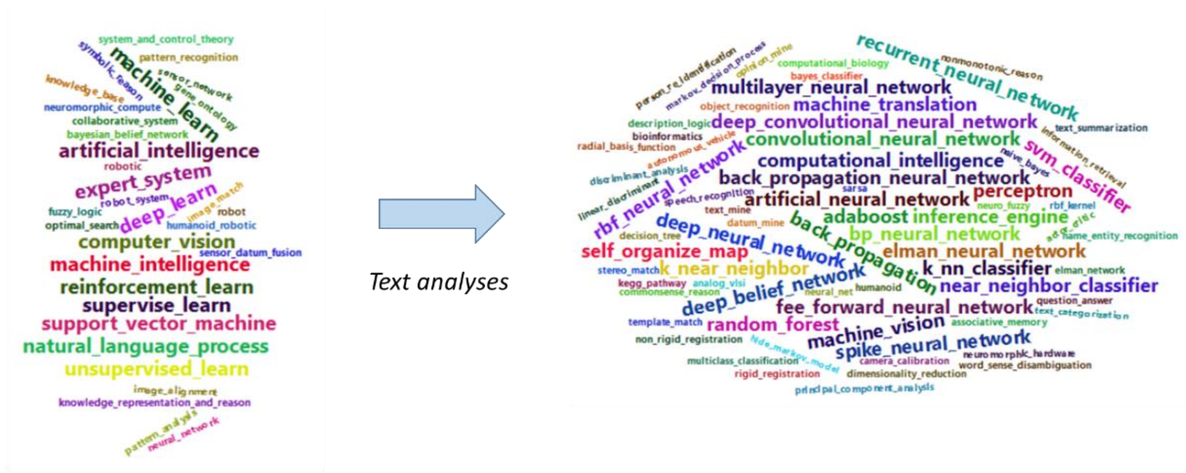


Note: One single label per cluster is presented to facilitate readability, giving priority to the AI core terms or in its absence the most frequent label in the cluster. The colour coding is as follows: Green for AI core terms, blue –terms selected as additional AI, and red, for terms removed not considered as AI related in the subsequent scoring procedure.
Source: OECD calculations based on NSF Award Search data, accessed December 2018.

[18] It is left for future work to visualise the entire NIH portfolio/vocabulary and its connectedness to AI terms.

[19] An emerging interdisciplinary field focused on designing hardware/physical models of neural and sensory systems

Figure 2.6. Key term selection based on core list of AI terms



Initial AI key terms from two keyword sets (MeSH and Cockburn)

Semi-automatically retrieved additional AI key terms from Scopus, NIH and NSF data

Note: This dual word cloud represents the extension from a core set of key AI terms based on seed "expert" lists to a text-mining extended list of AI terms for document retrieval. The full list of selected potential AI key terms is available in the Table A.2. The key terms in the figure are lemmatised (e.g. machine learning -> machine learn). Core AI terms are presented using a larger font size.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018; on NIH RePORTER; and on NSF Award Search, accessed December 2018.

2.3.2. Tagging documents with the list of AI key terms

38. Equipped with a list of key terms, it is now possible to identify which documents incorporate such terms. So far, the process has been largely oriented towards reducing the risk of poor recall and missing out on AI-related projects. As a result, the selection step requires further caution in order to avoid selecting documents that make use of AI key terms when they do not refer to actual AI systems use or development.

39. In order to minimise this risk of low precision, a very simple threshold and scoring approach has been applied that assigns different weights to various key terms and sets a minimum requirement. Up until now, overly ambiguous terms have been manually removed from the candidate list, while potentially ambiguous terms are tagged as "non-core" to reflect that there may be some ambiguity when examining a general corpus. Some terms included in the expert M and C-lists such as "neural network" are therefore categorised as non-core, because the term is not only used for describing an AI algorithm but also for network of neurons in a biological sense. In a similar context, "reinforcement learning" is classified as non-core as it can be used in other R&D contexts.

40. The project selection criterion imposes that a project description should at least contain either one core term or two or more distinct potentially ambiguous (non-core) terms.[20] In other words:

[20] To ensure distinctiveness, a number of terms are merged for the key term matching as they might appear combined into a single document without necessarily referring to AI. This manual

- A document is selected as (likely) AI-related if
  - o At least one core key term is found within its title or abstract; or
  - o Two or more distinct non-core terms are found
- Other documents are classified as (likely) non AI-related

41. This is a rather simple yet somewhat naïve procedure aimed at resolving potential ambiguity, as the implicit scoring and thresholds are defined somewhat arbitrarily. The idea is that by requiring that at least two potentially ambiguous terms are included to be considered AI, the likelihood of accepting a non AI-relevant document will be significantly reduced. A higher threshold (or lower scoring for such terms) would definitely improve precision but would on the other hand increase substantially the risk of poor recall. Another marked disadvantage of this procedure is the extent of manual intervention required to assign terms into the three possible categories (AI core, AI non-core and overly ambiguous).

42. Ultimately, these challenges stem from the impossibility of training a selection algorithm due to the lack of database that is comprehensive of the entire corpus on which the document detection is intended. In order to better understanding the resulting error, a manual examination of potential biases was carried out by extracting and analysing four samples of documents:

- Within the set of documents identified as AI-related, 100 documents were selected from each of the NIH and NSF corpora. This examination allows checking which documents might have been wrongly classified as AI-relevant (false positive rate).
- Within the set of documents identified as not AI-related, 100 documents were selected from each of the NIH and NSF corpora. This examination allows checking which documents might have been wrongly classified as non AI-relevant (false negative rate).

43. The human visual examination of those 400 documents seeks to establish whether the documents are unambiguously AI-or non AI- related, as well as whether, on the basis of the information available, it is impossible to tell either way without further information about the project. Within this latter class, it is often possible to separate between projects that lay out tasks for which AI tools are often required and others that do not provide explicit connection points with AI. This analysis helps provide an indication of the challenges associated to text mining project abstracts for information retrieval and classification according to AI-relatedness.

merge adjustment is implemented for “humanoid robotics”, “robot systems”, “robotic” which are merged into the term “robot” and “neural net” is merged to “neural network”.

3. Results

3.1. Distribution of AI key terms

44. **Table 3.1** shows the list of presumed AI key terms that frequently appear in documents corresponding to granted applications to NIH. It shows that the "bioinformatics" is the highest incidence term, followed by "knowledge base". Neither of them are considered as strong AI key terms, as they may be used in non AI-related contexts.[21] Among the 20 most frequent AI key terms, 12 are derived from the original expert lists for AI key terms while the rest are high co-occurring terms, which might be referred to as quasi-synonyms.

Table 3.1. Most frequent AI key terms in the NIH funding database

AI terms	Source	Key term status	Number of documents occurring	Incidence rate (per 10 000 documents)
bioinformatics	Scopus	non-core	24 089	192.08
knowledge base	M	non-core	5 435	43.34
computational biology	Scopus	non-core	3 885	30.98
neural network	M & C	non-core	3 820	30.46
robot	M & C	non-core	3 509	27.98
machine learn	M & C	CORE	2 827	22.54
pattern recognition	C	non-core	2 591	20.66
datum mine	NIH	non-core	2 019	16.10
speech recognition	NSF	non-core	759	6.05
object recognition	NSF	non-core	744	5.93
natural language process	M & C	CORE	621	4.95
gene ontology	M	non-core	541	4.31
computer vision	C	CORE	469	3.74
reinforcement learn	C	non-core	466	3.72
principal component analysis	Scopus	non-core	437	3.48
pattern analysis	C	non-core	434	3.46
hide markov model	NSF	non-core	405	3.23
associative memory	Scopus	non-core	320	2.55
support vector machine	M	CORE	309	2.46
artificial intelligence	M & C	CORE	302	2.41

Note: The "Source" column denotes de provenance of the AI term: "M" for MeSH, "C" for Cockburn et al. (2018), and "Scopus" refers to the quasi-synonyms retrieved from Scopus. "Key term status" refers to the scoring of the terms. "CORE" terms fully identify AI documents.
Source: OECD calculations based on NIH RePORTER data, accessed December 2018.

[21] Bioinformatics is considered in this context as weakly related to AI given the technical needs in this area. "Knowledge bases" is one of the MESH AI terms and refers to "collections of facts, assumptions, beliefs, and heuristics that are used in combination with databases to achieve desired results, such as a diagnosis, an interpretation, or a solution to a problem". One example of a project in an AI context proposes "to develop two separate Decision Support Systems using totally different approaches, a heuristic model (knowledge based expert system) and a predictive statistical system. These Systems will be developed from a database of 3500-4000 MAG3 studies and will be designed to acquire the study, generate images and curves […], check for errors, extract the relevant quantitative data and then use these data to interpret the study." An example of a non AI relevant project refers to "training goals […] will enhance the applicant's knowledge base in child and adolescent mental health services".

45. **Table 3.2** shows the equivalent list of AI key terms for the NSF database. The terms "robot" and "machine learn(ing)" are the two highest occurring, followed by "bioinformatics" and "knowledge base". Among the top 20 most frequently occurring AI terms, 11 are derived from the initial set of AI key terms and the rest are quasi-synonyms.

Table 3.2. Most frequent AI key terms in the NSF funding database

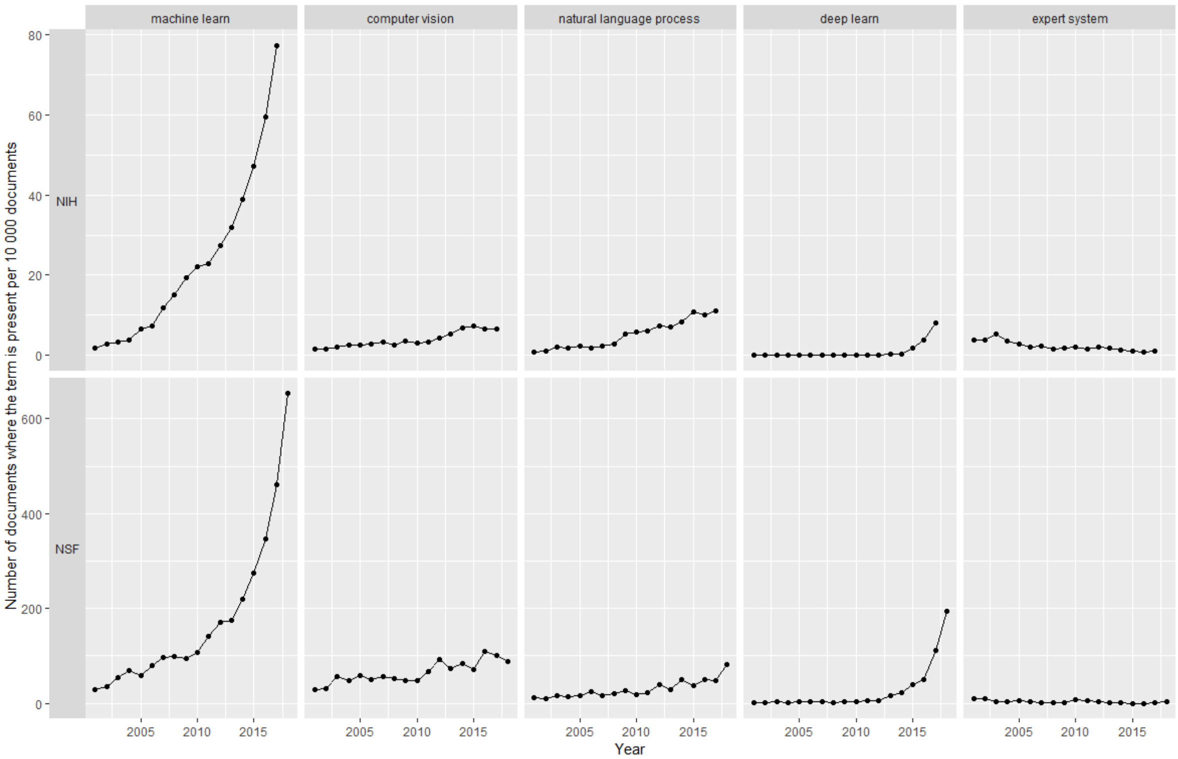
AI terms	Source	Key term status	Number of documents occurring	Incidence rate (per 10 000 documents)
robot	M & C	non-core	5 143	237.92
machine learn	M & C	CORE	3 877	179.35
bioinformatics	Scopus	non-core	2 531	117.09
knowledge base	M	non-core	2 005	92.75
sensor network	C	non-core	1 903	88.03
datum mine	NIH	non-core	1 791	82.85
computer vision	C	CORE	1 407	65.09
artificial intelligence	M & C	CORE	1 075	49.73
computational biology	Scopus	non-core	914	42.28
neural network	M & C	non-core	716	33.12
natural language process	M & C	CORE	653	30.21
deep learn	C	CORE	572	26.46
information retrieval	NSF	non-core	500	23.13
pattern recognition	C	non-core	487	22.53
speech recognition	NSF	non-core	434	20.08
autonomous vehicle	NSF	non-core	395	18.27
reinforcement learn	C	non-core	259	11.98
object recognition	NSF	non-core	225	10.41
machine vision	Scopus	CORE	198	9.16

Note: In the "Source" section, "M" means MeSH, "C" means Cockburn et al. (2018), "Scopus" means quasi-synonyms retrieved in Scopus. "Key term status" means how the terms in the listed treated in the analysis of the database. "CORE" terms fully identify AI documents.
Source: OECD calculations based on NSF Award Search data, accessed December 2018.

46. **Figure 3.1** shows the incidence of documents featuring selected AI key terms per 10 000 documents in the NIH and NSF databases over time (machine learn*, computer vision, and natural language process, deep learn* and expert system). Despite differences in scale (NSF data show a higher occurrence per document), the trends for all the terms are rather similar. "machine learn" shows a high increase from around 2010. "computer vision" and "natural language process" show stable growth. The term "deep learn" displays a sharp increase at around 2015, especially in the NSF database. The term "expert system", which started from a higher level than other AI terms, has instead flat-lined or even declined since 2005 in the case of NIH data.

Figure 3.1. Occurrence of selected AI key terms in NIH and NSF data

Number of documents where the term is present per 10 000 documents



Source: OECD calculations based on NIH RePORTER and NSF Award Search data, accessed December 2018.

3.2. Estimates of AI-related R&D projects and funding

3.2.1. NIH funding

47. Estimates of AI-related R&D projects (granted applications) and funding are based on the selection procedure described in the previous section for the period 2001 to 2017. As shown in **Table 3.3**, the number of projects identified as AI-related has more than quintupled over this period, growing its share from 0.2 to over 1.4% of all projects funded by the NIH. The amounts of R&D funding display a similar trend, going from 0.3% in 2001 of NIH's total funding allocation to nearly 2.4% in 2017, with USD 823 million being allocated on this last year for which figures are available.

48. **Figure 3.2** shows a sustained growth in the share of AI-related R&D funding punctuated by a large one off increase in 2009, coinciding with the allocation of additional funds under the American Recovery and Reinvestment Act (ARRA). The comparison of project count and funding data point to the existence of relatively large projects classified as AI-related. This is confirmed by further analysis plotting the distribution of funding amounts per granted application. **Figure 3.3** shows that there are some exceptionally large applications from 2009 onwards. As a result, applications with more than USD 100 million in funding have been manually examined, as they have high impacts on estimated AI funding shares.

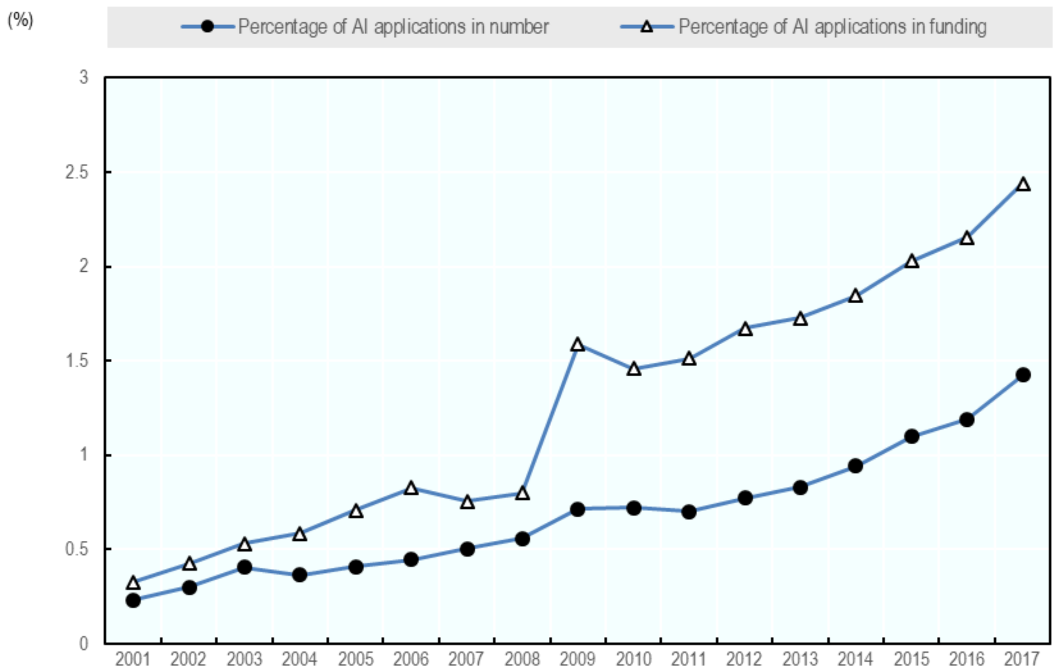
Table 3.3. Estimates of AI-related R&D in NIH funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	183	78 033	0.23	58	17 736	0.33
2002	240	79 912	0.30	84	19 618	0.43
2003	235	58 049	0.40	105	19 744	0.53
2004	273	74 367	0.37	129	21 917	0.59
2005	315	77 050	0.41	165	23 261	0.71
2006	341	76 195	0.45	193	23 309	0.83
2007	412	81 697	0.50	225	29 818	0.76
2008	445	79 576	0.56	239	29 830	0.80
2009	642	89 583	0.72	563	35 495	1.59
2010	607	84 195	0.72	535	36 689	1.46
2011	515	73 322	0.70	475	31 378	1.51
2012	531	68 610	0.77	512	30 624	1.67
2013	551	66 451	0.83	502	29 074	1.73
2014	618	65 540	0.94	552	29 920	1.84
2015	723	65 835	1.10	611	30 096	2.03
2016	798	67 030	1.19	690	32 002	2.15
2017	978	68 694	1.42	823	33 729	2.44

Note: These figures are results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on NIH RePORTER data, accessed December 2018.

Figure 3.2. Estimates of AI-related NIH-funding

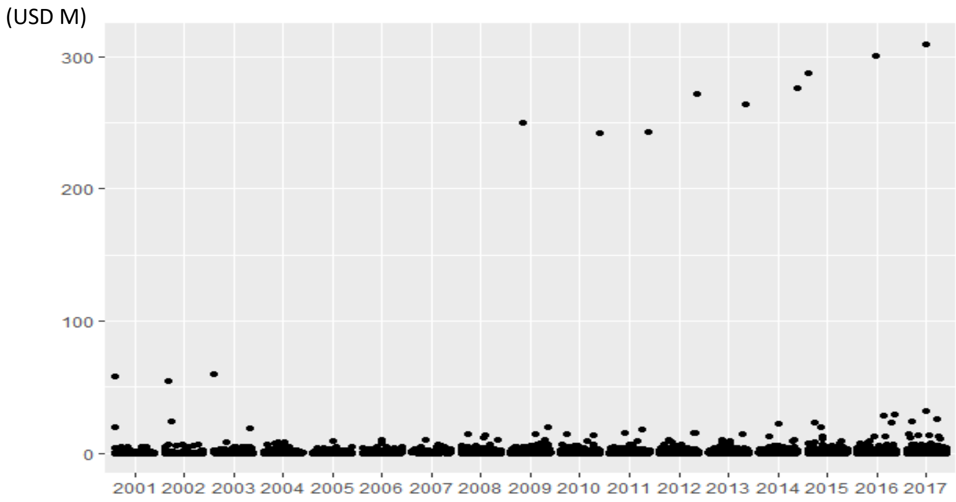
As a percentage of total NIH funding (number of projects and funding amounts)



Source: OECD calculations based on NIH RePORTER data, accessed December 2018.

49. The extreme values in excess of USD 100 million are accounted for by an infrastructure project classified as AI related. This project, called "National Biomedical Information Services", an intramural project within one of the NIH institutes, the National Libraries of Medicine. This project has annual records with text descriptions that are very similar each year. It was decided to retain this project as AI-related as it contributes to build and provide various information systems that use AI systems, such as natural language processing and tools for managing large bibliographic databases.

Figure 3.3. Funding amounts per application in NIH project funding data



Source: OECD calculations based on NIH RePORTER data, accessed December 2018.

50. The results do nonetheless confirm the sustained growth in AI-related funding over the period with a brief hiatus from 2006 to 2008. Growth is particularly fast both in terms of numbers of documents and funding in the 2010s, with no signs of deceleration in the final years for which data are available.

3.2.2. NSF funding

51. Following the same procedure as above, AI-related R&D funding is summed up for the NSF in **Table 3.4**. Over the period from 2001 to 2018, its share of AI-related R&D funded projects has increased from 1.8% in 2001 10% in 2018. The growth in terms of funded amounts is very similar, from 1.5% in 2001 to 11.5% in 2018.

52. Compared to NIH funding, NSF funding exhibits a higher degree of relatedness to AI from the start of the period, which is probably explained by the presence of computer science funding within the NSF portfolio. As seen in **Figure 3.4**, the time profile is also different, with a very fast growing rate of AI intensity towards the end of the period for NSF data. Both the absolute growth in AI-related funding in absolute terms and the overall decline in NSF funding from 2015 contribute to this.

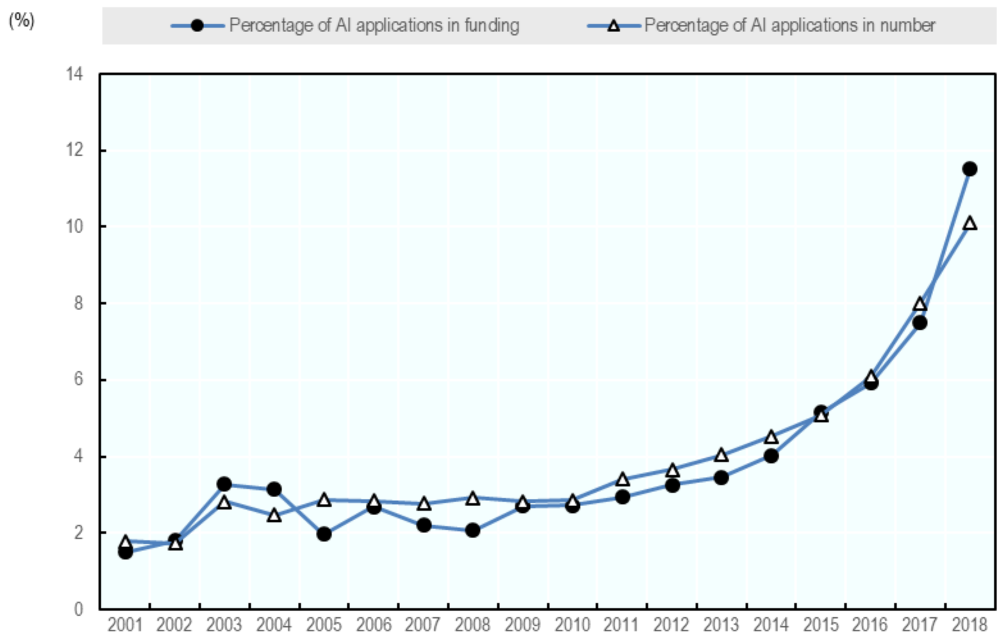
Table 3.4. Estimates of AI-related R&D in NSF funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	186	10 427	1.78	87	5 873	1.49
2002	189	10 939	1.73	95	5 300	1.79
2003	326	11 591	2.81	212	6 462	3.27
2004	269	10 927	2.46	156	4 964	3.14
2005	299	10 403	2.87	101	5 118	1.97
2006	308	10 886	2.83	147	5 481	2.68
2007	336	12 121	2.77	111	5 029	2.20
2008	343	11 742	2.92	150	7 238	2.07
2009	430	15 244	2.82	217	8 028	2.70
2010	394	13 818	2.85	201	7 413	2.72
2011	407	11 948	3.41	195	6 649	2.93
2012	451	12 311	3.66	201	6 181	3.26
2013	473	11 703	4.04	198	5 721	3.46
2014	543	11 984	4.53	242	6 007	4.02
2015	657	12 929	5.08	281	5 462	5.14
2016	790	12 994	6.08	359	6 069	5.91
2017	977	12 185	8.02	389	5 192	7.49
2018	1214	12 014	10.10	526	4 561	11.53

Note: These figures are results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on NSF Award Search data, accessed December 2018.

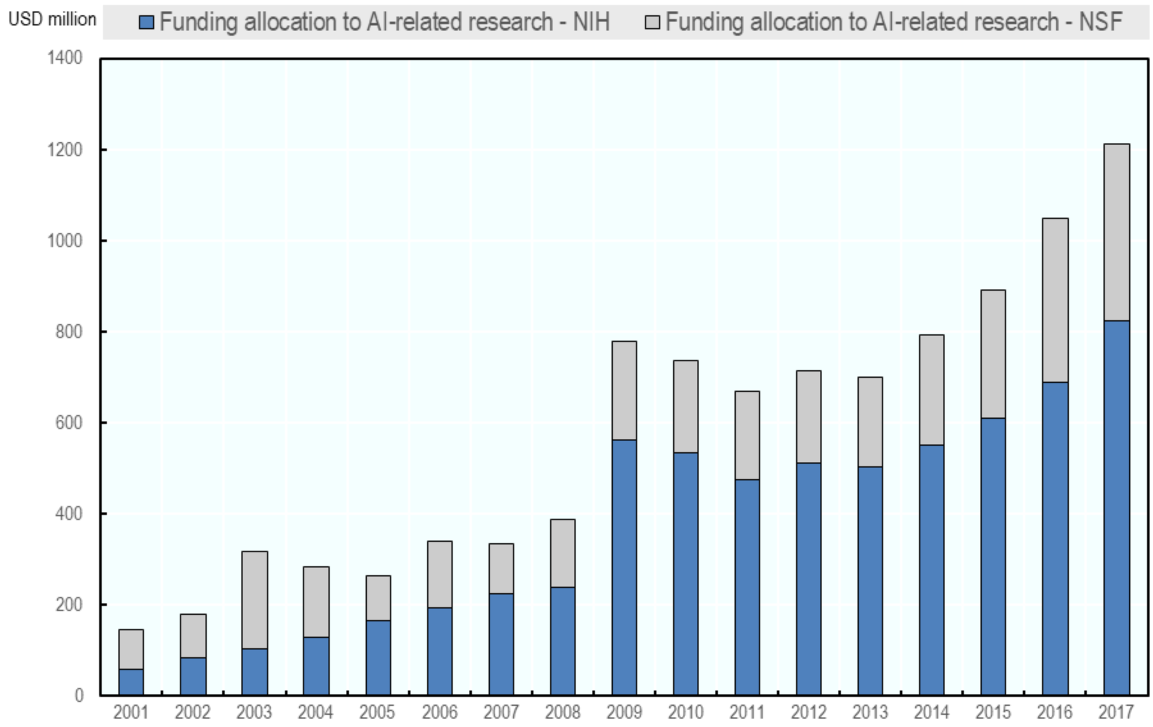
Figure 3.4. Estimates of AI-related NSF-funding

As percentage of total NSF funding (number of projects and funding amounts)



Source: OECD calculations based on NSF Award Search data, accessed December 2018.

Figure 3.5. Estimated NIH and NSF funding for AI-related R&D, 2001-2017



Note: The surge in funding in 2009 was not specific to AI R&D and is related to the additional funding facilitated by the economic stimulus package contained in the American Recovery and Reinvestment Act.
Source: OECD calculations based on NIH RePORTER and NSF Award Search data, accessed 1 December 2018.

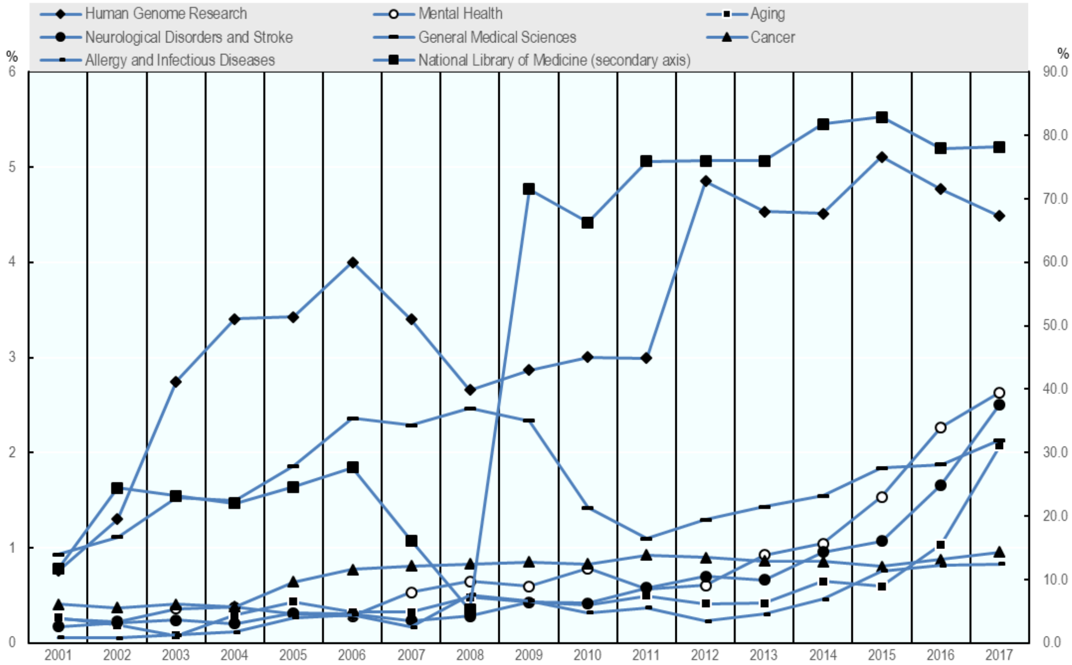
53. Despite a higher AI intensity of NSF funding, the difference is not large enough to offset the larger size of the overall NIH budget. As a result, NIH AI-related funding is systematically larger than NSF AI-R&D funding, about three times as large in recent years, compared to similar absolute amounts in 2001-2003 (**Figure 3.5**). Together, NIH and NSF dedicated USD 1.2bn in 2017 to AI-related R&D, namely 3.1% of their combined funding, up from 0.6% in 2001.

3.3. AI research across different R&D domains and application areas

54. Based on the tagging of funding portfolios by AI-relevance, it is possible to examine which areas account for the largest volumes of AI related funding and which ones exhibit a higher degree of AI-intensity. **Figure 3.6** shows the incidence of AI-related R&D within the selected NIH institutes that manage the funding and whose names describe their intended application areas. In 2017, the National Library of Medicine accounts for both the largest share of AI-related research within NIH (about one third of the total) and has the highest internal AI intensity at over 75%, followed by the National Institute of Biomedical Imaging and Bioengineering and the National Human Genome Research Institute close to 5%. In total funding terms, NLM is followed by the National Cancer Institute, which has an AI "intensity" of around 1%.

Figure 3.6. Estimated share of AI-related R&D funding within NIH institutes

"AI intensity" for selected institutes with the largest estimated amounts of AI funding

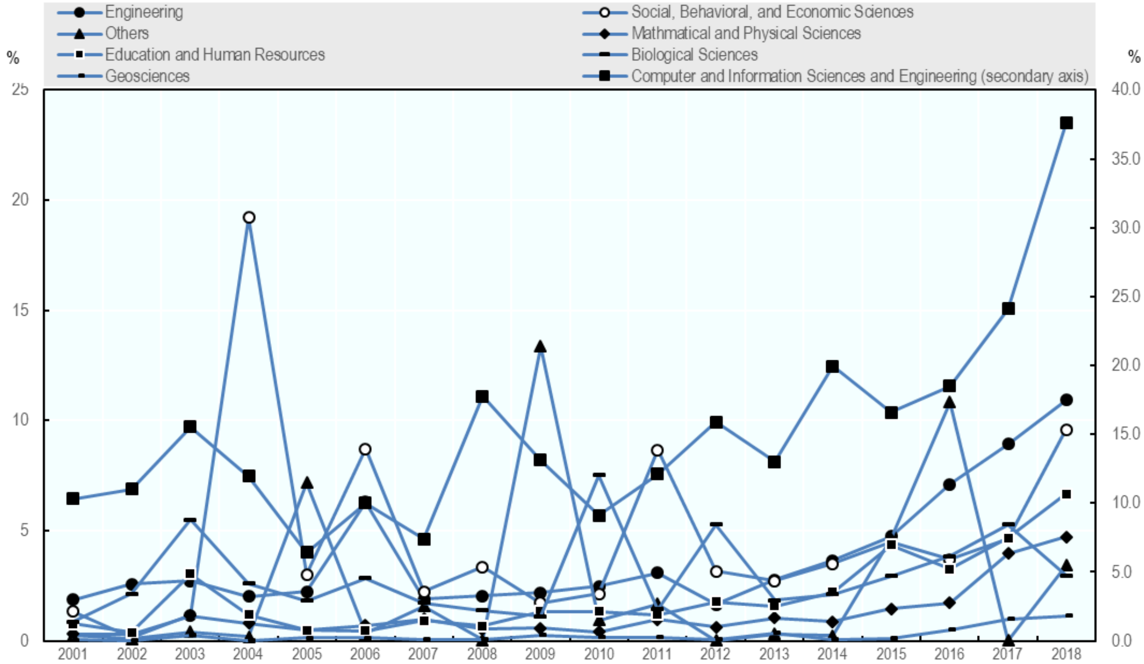


Note: This is an experimental indicator. For clarity of presentation, and with the exception of NLM, the names of the institutes are presented by referring solely to their missions.
Source: OECD calculations based on *NIH RePORTER* (database), accessed December 2018.

55. **Figure 3.7** shows the incidence of AI-related R&D within the NSF directorates with responsibility for managing the funding for different disciplinary domains. AI intensity in 2018 is more than 35% in the case of Computer and Information Sciences (see secondary axis), up from less than 10% in 2001. This is followed by Engineering (general) at 11%, up from nearly 2% in 2012. The surge in funding allocated by the NSF Directorate for Social, Behavioural and Economic Sciences in 2004 is accounted for two large projects dedicated to building learning centres. For one, there is a legitimate concern about whether it represents AI related research, as it appears that the use of the broader term "self supervised learning" refers to human learning activity. However does ultimately appear to have a significant AI content because it includes among its linked bibliographic outputs (not automatically examined in this analysis) a paper in a conference dedicated to machine learning.[22]

[22] The paper is entitled "Applying Programming by Demonstration in an Intelligent Authoring Tool for Cognitive Tutors" and was published as a technical report in the "AAAI Workshop on Human Comprehensible Machine Learning".
https://www.nsf.gov/awardsearch/showAward?AWD_ID=0354420&HistoricalAwards=false

Figure 3.7. Estimated share of AI-related R&D funding within NSF disciplinary directorates



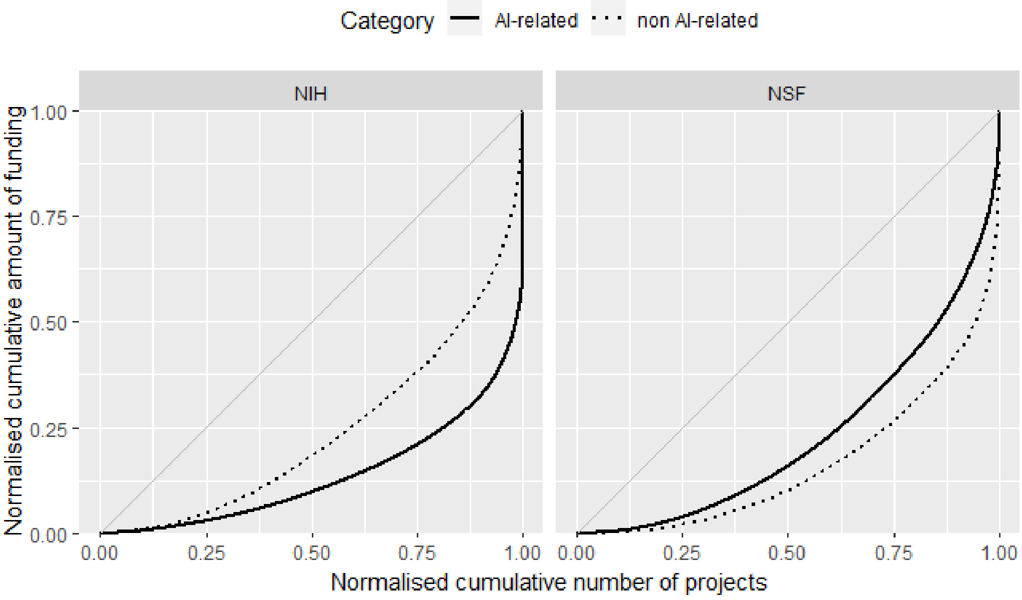
Note: This is an experimental indicator.
Source: OECD calculations based on NSF Award Search (database), accessed December 2018.

3.4. Concentration of AI funding

56. **Figure 3.8** shows Lorenz concentration curves of AI-related and non AI-related funding for both agencies. The funding distribution of AI-related R&D in the NIH portfolio is relatively concentrated, with 75% of AI R&D funding allocated to the 25% largest AI-related projects, compared to 37.5% for non AI-related R&D. This is partly explained by the AI-related large intramural data infrastructure projects previously identified for the NLM. A focus on purely extramural projects is likely to lead to a more similar picture, also in relation to NSF funding. In the case of NSF funding, non-AI related projects are relatively more concentrated than their AI counterparts but by a much smaller margin.

Figure 3.8. Funding concentration of AI-related and non AI-related projects

Lorenz curves of NIH and NSF funding



Note: The Lorenz curves represent the share (from 0 to 1) of funding (y-axis) that corresponds to the cumulative share of projects order from smallest to larger (x-axis) in each category.
Source: OECD calculations based on NIH RePORTER and NSF Award Search data, accessed December 2018.

3.5. Bias and robustness checks

3.5.1. Incidence of false positives (precision error)

57. In order to assess the robustness of the results, especially in light of the presence of somewhat ambiguous key terms that may bias these estimates (e.g. if two ambiguous AI terms or more are found in the same document), batches of 100 documents were selected at random from each of the 8 407 (NIH) and 8 592 (NSF) documents categorised as AI-related R&D. The titles and abstracts of these 200 documents were manually inspected and classified into four categories shown in **Table 3.5**. Documents making clear reference to the use or development of AI systems (category A) can be treated as unambiguous true positives. In addition to this class, there are also projects which from the description, can be deemed to represent AI-related R&D activity, based on the context (category B). These are likely true positives. The remainder can be conservatively considered as the measure of precision error, covering instances in which the description makes it apparent that the project has little to do with AI (category D) or not enough context is available to judge what the selected AI key terms actually refer to (C). Examples of projects allocated to different categories are provided in **Box 3.1**.

Table 3.5. Precision analysis of AI detection results in NIH and NSF data

Sample of randomly chosen documents from selected documents identified as AI-related

Status	NIH	NSF
A. Explicit AI-relevance.	60	56
B. Likely AI relevance	27	36
C. Possible false positive (insufficient information in text to tell)	***11***	***6***
D. Likely false positive	**2**	**2**
Total	100	100

Source: OECD analysis based on NIH RePORTER and NSF Award Search data, accessed December 2018.

58. In the case of the NIH funding data, the more conservative measure of precision bias is about 13% (categories C and D), or 2% for the strong measure of bias. In the case of NSF data, the broad measure precision error is lower at 8%, possibly reflecting the higher strength of the project methodology signal contained within project abstracts compared to NIH data and the avoidance of neurological or developmental analogue terms adapted in AI. Based on these samples, the narrow measure of precision error lies between close to zero and 7% for NIH and 4% for NSF with a 95% probability, while the broad measure lies in the 7-to-21% and 4-to-15% ranges, respectively.

59. The precision analysis allows us to identify two instances (one in each database) where lemmatisation mistakenly converts a term into one of the AI terms, namely when “deeper learning” was converted to the generic “deep learn” term.[23] Furthermore, an instance of the term “Supervised learning” was found to refer to actual instructional activity, not supervised machine learning. This suggests that the term may need to be partly penalised in order to reduce the risk of false positives.

[23] However, one of such instances may actually refer to an AI enabled project because as highlighted in one of the examples, the types of sensor enabled games may be supported by AI system. Further search confirmed that the company that was awarded the project reports on its website that “transforms training, sales, service, production and design by leveraging virtual and augmented reality (VR/AR), simulation, sensing, artificial intelligence and machine learning (AI/ML) across the totality of employee, customer and product life cycles. transforms training, sales, service, production and design by leveraging virtual and augmented reality (VR/AR), simulation, sensing, artificial intelligence and machine learning (AI/ML) across the totality of employee, customer and product life cycles.

Box 3.1. Excerpts from sample of projects automatically retrieved as AI-related

Projects with explicit AI relevance:

- **Neural Networks for Estimating and Compensating the Nonlinear Characteristics of Nonstationary Complex Systems.** "The approach is to use Echo State Networks and Simultaneous Recurrent Neural Networks with super fast learning algorithms (biological inspired algorithms such as particle swarm optimization), and other computational intelligence algorithms, to accurately measure the distortion by monitoring only voltage and current without the need for added transducers. Such fast and powerful **neural networks** could also be used for closed loop control of the offending nonlinear devices to mitigate the distortion." (NSF)
- **Automated NMR Assignment and Protein Structure.** "New algorithms and computer systems will be developed for determining protein structure from only four NMR spectra. The system will use algorithms similar to and adapted from physical geometric algorithms, **pattern recognition** and **machine vision**, signal processing, and **robotics**, in order to analyze spectra, assign spectral peaks to atom interactions, compute secondary structure, and estimate the global fold." (NIH)

Projects that appear to be incorrectly or ambiguously identified as AI-related:

- **Identification and characterization Of A Complex Involved In Bloom Syndrome.** "In **bioinformatics** searches of the human genome, we noticed that human genome contains proteins with OB-fold domains similar to those in RMI and RPA." (NIH)
- **Recombinant Multiepitope Mosaic Protein Design for Urine-based Diagnosis of Leptospirosis**. "Our approach involves the use of computational biology and **bioinformatics** to create, score, and select "mosaic" antigens from Leptospira spp. Antigenic properties of the mosaic antigens are evaluated by indirect ELISA using a panel of well-characterized human sera from clinical patients and apparently healthy individuals. We will then use recombinant DNA and protein engineering techniques to derive cognate chimeric proteins." (NSF)
- **Personalized sensor based digital media simulations for Biology and Health education.** "In this project, we present a set of sensor-enabled, multimodal, NGSS aligned, validated formative assessment games for biology and health education. Our emphasis will be on high engagement, **deeper learning** of the heart and cardiovascular function and diseases and valid formative feedback to guide next steps for teachers and to allow student to assess themselves". (NIH)
- **CAREER**: **Compiler and Runtime Support for Multi-Tasking on Commodity GPUs.** "GPU computing has become mainstream, as witnessed in various domains such as **machine learning**, graph analytics, and scientific simulation. This CAREER project aims at developing a set of compiler and runtime techniques to support multi-tasking on commodity GPUs in a transparent and efficient manner"(NSF)

3.5.2. Incidence of false negatives (recall error)

60. To assess the potential margin of error associated with using an incomplete list of key terms, a similar manual detection process is followed for documents not identified as AI-related. A total 100 documents were selected at random from each of the 1 254 139 (NIH) and 216 166 (NSF) documents categorised as non AI-related in order to establish what percentage of such documents were incorrectly excluded by the AI tagging procedure.

61. In the NIH case, none of the 100 projects examined were classified as type A or B (i.e. AI-related R&D). 10 documents were categorised as C and 90 as D. Very similar results are obtained for the NSF sample. These results imply a low rate of recall bias, in the order of 5 to 18% on the broadest possible measure, which is a likely overestimate (and 0 to 5% on a stricter measurement) with a 95% confidence probability.

Table 3.6. Recall analysis of AI detection results in NIH and NSF data

Categories	NIH	NSF
A. Explicit AI-relevance (false negative)	**0**	**0**
B. Likely AI relevance (likely false negative)	**0**	**1**
C. Possible false negative (insufficient information in text to tell)	**10**	**8**
D. Likely true negative	90	91
Total	100	100

Source: OECD calculations based on NIH RePORTER and NSF Award Search data, accessed December 2018.

62. The rejected documents assigned to the C category after examination (and therefore possible false negatives) are revealing of some of the challenges arising for AI relevance (or for that sake, the use or relevance of any other technology) when adoption rates are rapidly increasing. The example of a project untagged by the chosen key terms that seeks to "add strength in statistical methods for genetic data, clinical prediction, and paediatric oncology" raises questions as to, under what circumstances, it is plausible for statistical methods not to be AI-enabled. One should bear in mind that adding selected complementary inputs (genetic data) and applications (clinical prediction) into the selection process risks extending too much the selection window.

3.5.3. General assessment

63. While it may appear that there is worse precision error than recall error on a conditional basis[24], it is important to note that the recall error probability estimate applies to what is currently a much larger group, so that ultimately the results may actually underestimate the extent of true AI-related R&D funding by quite some margin. On the other hand, it is prudent to adopt a more conservative approach.

64. After examining 400 documents for estimating precision and recall errors, there are some potential lessons for refining the keyword matching approach. Firstly, the core set of AI-related terms appear to provide reasonably unambiguous predictors for AI relevance (e.g. "machine learning", "natural language processing" and "deep learning". Even the problems associated with the use of the term "deep learning" with a different intent in education science projects revealed in major instances that the projects had been carried out by AI experts and/or resulted in AI publications. This core set of documents could be extended provided that the text mining techniques used can cope with very large vocabularies.

65. A distinctive feature of key term identification in funding data, compared to working with counts of projects or documents, is the importance of prioritising the

[24] I.e. the probability of a wrong positive conditional on a project being tagged as AI vs the probability of wrong negative conditional on being tagged as non AI.

assessment of large projects, for the overall impact might be magnified through inaccurate measurement of individuals projects funded to the tune 10s of millions of USD or more. Larger projects also present difficult choices, in the sense of whether the entire amount allocated should be treated as AI or only fractionally. These are questions that lie beyond our current capability but could be ultimately tackled.

66. Furthermore, potential but non-exclusive terms used for AI should be further checked against the context in which they are use on a document-by-document basis, progressively substituting for the simple scoring approach. As noted, our approach requires some degree of human discretion in relation to key terms before applying the simple naïve rule. Results can also be rather sensitive to the decision on how many non-core items to require, probably more so than the actual selection of key terms.

67. The sensitivity test based on an assessment of random samples of the corpus reveals the fundamental inconclusiveness of the abstract portion of many project applications, particularly but not only for the projects identified as non-AI related. This points to the fundamental challenge of fitness for purpose of the abstracts. Abstracts are nonetheless convenient for use because they keep the data size manageable but also because they can be openly accessible, while detailed project descriptions are unlikely to qualify for public access.

3.5.4. Robustness analysis. Comparing to other lists

68. As a further robustness test, these results have been compared to those that would be derived using an alternative (and longer) final list of AI key terms. The measurement of AI-related activity has been recently considered in recent OECD work (OECD, 2019 forthcoming). The list of 193 terms derived in that context with the purpose of analysing the corpus of scientific publications (including conference proceedings) and also applicable in other contexts provides a useful check on the sensitivity of estimates to alternative methods to compile lists of identifying key terms.

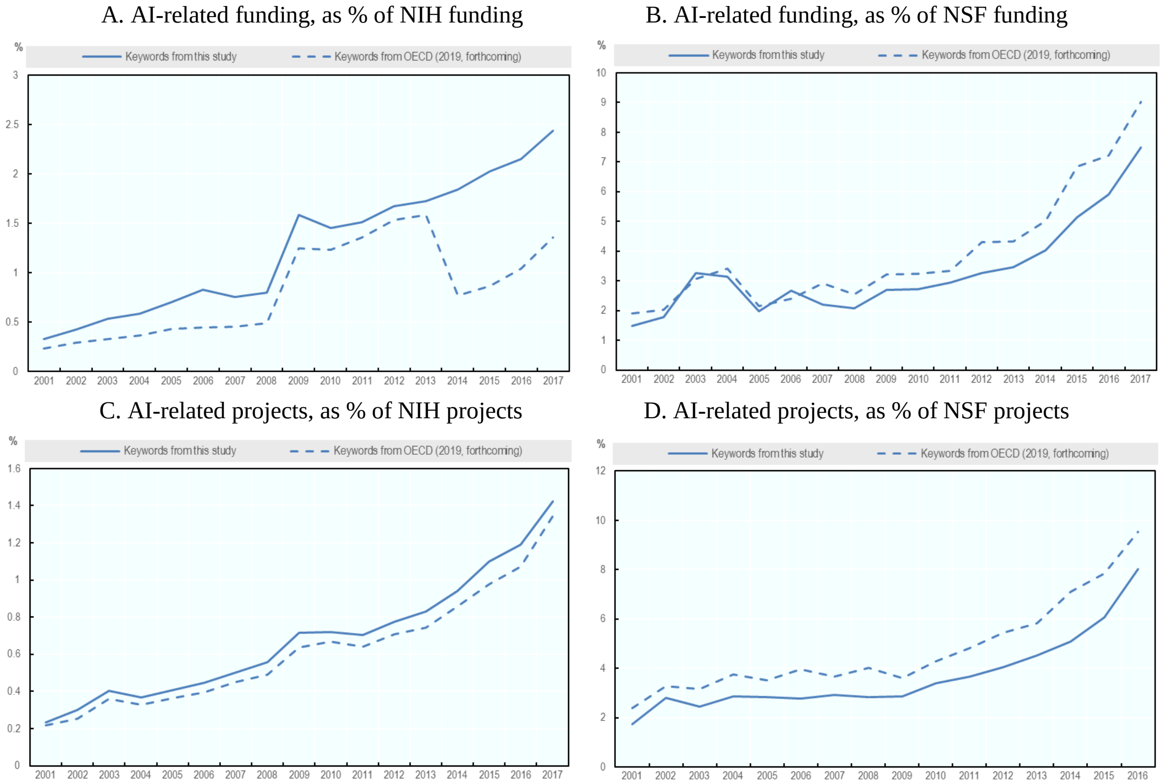
69. A comparison of both list of terms shows that while the overlap is very significant with 35 common terms, the OECD (2019, forthcoming) report contains a significantly longer list of terms. Our study does however contain a smaller but also significant number key terms that are not present in the comparison list. This suggests that it is possible to reduce further the risk of false negatives by combining both lists, although this might be at the expense of allowing for a higher rate of false positives. Bioinformatics is the most common term in this paper's list that is not found in the previous OECD report, while computational model is the most common term in that report that was not included in this paper's list.

70. Quantitatively, the application of the different lists does result in somewhat different estimates of AI related project counts and funding amounts, as shown in **Figure 3.9**. To make a like-for-like comparison, the same criterion of at least one core or two non-core terms is applied to both lists.

71. In the case of NSF funding, estimates in this paper are on average 15% lower than those obtained when using the 193 list. This gap is fairly stable since 2007. Between 2003 and 2006, estimates are virtually identical. The gap is rather similar for counts of projects. For NIH funding, estimates based on this paper's list are however significantly larger, from 40% larger in 2001 to about twice as larger in more recent years. This suggests that the health-oriented NIH corpus is quite distinct from that analysed and the specific semantic analysis of this corpus allows us to retrieve more

potentially relevant documents and funding. The analysis of the time profile for both NIH estimates also underlines the importance of the large scale intramural data infrastructure projects funded since 2009. The OECD (2019 forthcoming) list appears to capture them up to 2013 but these are subsequently lost as the identifying terms disappear from the abstracts which are found to be otherwise very similar. Indeed, project counts estimates are more similar than funding estimates. In spite of differences in terms of estimates for levels of R&D funding, both lists coincide in estimating annual growth rates in AI-related funding close to 15%.

Figure 3.9. Illustration of sensitivity of results to using alternative AI term lists



Note: The same decision criterion has been applied for the estimates applying to each list of key terms.
Source: OECD calculations based on NIH RePORTER and NSF Award Search data, accessed December 2018, and AI key terms in OECD (2019, forthcoming).

4. Conclusions and next steps

72. This document has presented the procedure and main results of a text-based analysis of administrative data at the project-level on governmental R&D funding, a potentially very valuable source of information about the size and directionality of public investments in science and technology. These data, although limited in their usability due to access restrictions and limited scope (i.e. they do not cover block funding for universities or institutions unless those institutions distribute funds on a project basis and make information about those projects available), are the best mechanism for timely insights on detailed aspects of R&D being funded by governments. With appropriate incentives, current research information systems

(CRIS) may in the future contain information about projects ultimately backed by block grants.

73. This quantitative case study, where a series of relatively simple text mining methods have been applied to publicly-reported funding data from the US National Institutes of Health and US National Science Foundation, two of the major R&D funding organisations in the world, highlights the potential for using project level data descriptions to carry out in depth analysis of the content and methods of R&D with such type of funding data.

74. The findings on the subject of interest, AI-related R&D, confirm the widely held view of an upward trend in R&D funding, reaching over 10% of NSF funding and 2.5% of NIH funding, slightly above the corresponding percentage of funded projects. Because the diversity of project sizes, it is important to go beyond counts of documents or projects. Working with funding data, it is possible to put a first and tentative monetary figure on AI R&D funding estimates (namely USD 1.2bn in 2017 for the combined NIH and NSF agencies). This document also shows that outside the core data infrastructure and computer science funding, where AI funding rates are very high, significant rates of AI funding can be found within engineering (NSF) as well as within health applications requiring large amounts of data such as genomics and neurological research. Funding within areas like cancer and infectious diseases are still characterised by particularly low and stable rates of AI-related projects.

75. This data-driven "classification and measurement" case study identifies challenges such as finding the right balance between mechanical procedures and personal judgements, the impact of large projects with high funding levels and multiple components. Data-based classification decisions on topics like AI, but also any other form of enabling technology, mechanically or manually implemented, require ultimately a view of when such applications or procedures are first significant enough to be systematically spelled out in applications so that they can be picked up within a vast body of information, and then become part of the "new normal" and references in project descriptions largely unnecessary. To be operationalised in a given corpus, the application of existing definitions of AI will be necessarily context dependent.

76. The study reveals that unstructured funding microdata contains valuable information that can shed insights on the structure of R&D efforts by governments. It also reveals that such efforts are highly dependent on comprehensive data infrastructure and exhaustive description of the projects. Text-mining of data containing superficial descriptions will fail to identify key features of projects, especially methodological, which can be critical for relevance identification purposes. In the case of R&D project abstracts, the information content about the intended application or research objective is likely to prevail unless access to more detailed methodological descriptions can be secured.

77. A number of possible next steps and extensions to the work may be envisaged focused around methodological improvements and extensions in the study's scope. Concerning the former, as highlighted in previous sections, it should be possible to develop a more fully automated non supervised classification procedure, achieving a better integration between the characterisation of key terms and the documents in the various corpora use, both auxiliary (e.g. publications) and target (funding data), making better use of information about the context in which potential key terms are used. This approach would therefore reduce the emphasis on lists of key terms as ultimate identification devices, basing the categorisation more on the similarity of

documents with respect to documents known to be AI related. Topic modelling could also be introduced to characterise the subset of projects identified as AI related, in order to investigate what domains and applications are being mostly impacted by the use of AI.

78. Regarding its scope, the project could proceed in different directions motivated by the text in the OECD AI recommendation on government funding of AI R&D. Firstly, keeping the focus on funding of scientific research, extending the analysis to other of funding agencies operating in this space is a necessary step towards developing nationally representative results that are ultimately suitable for international comparisons.[25] At the moment, preparatory work is assessing the basis for extending the analysis to the funding data in the UK Gateway to Research Database as well as data for Japan's KAKEN, as well as engaging in collaboration with other individual countries.

79. Secondly, it would be worthwhile to assess the possibilities for extending the work to different types of R&D support for which there may be textual descriptors. For example, in addition to grant data, public procurement data for R&D exhibits some similar features to R&D grant administrative data and may lend itself to this type of analysis. In many instances, this information will not be in public domain, but it may be possible for public bodies to carry out some form of coordinated analysis and share the results. This would be to some extent a "semantic" spin off for the ongoing OECD microBERD project, a project that ensures the preservation of business data confidentiality to inform the analysis of public support policies.

80. A third additional dimension of potential future NESTI work concerns the application of existing definitions of AI for use in "designed" statistical data sources, such as R&D or innovation surveys. For example, the US National Science Foundation already includes a question it is business R&D survey (BRDIS) on R&D activity *for* AI.[26] There is at present no international consensus on how such data should be collected. At the recent workshop on innovation surveys and the implementation of the 2018 *Oslo Manual*, participants consistently highlighted this topic as a high priority in the context of addressing recommendations on measuring "digital innovation".

81. To conclude, delegates are invited to submit comments on the methodology of the study and its potential applicability to national funding databases in their own countries, starting with scientific research funding agency data, with a view to augmenting the scope of the work and to publish the results as an OECD working paper.

[25] As previously hinted, there is a clear interest in evaluating the possibility of this work more globally to examine AI trends. This raises issues of data availability and consistency. In the case of public funding of research, different countries organise their support in different ways, particularly in terms of the extent to which public authorities grant support for specific projects or institutions and their employees have discretion on the use of funding allocations.

[26] The 2018 survey question asks "What percentage of the amount reported [in Question 5-2 – domestic R&D paid for and performed by your company] was for artificial intelligence (AI)?" and goes on to define AI as "A branch of computer science and engineering devoted to making machines intelligent. Intelligence is that quality that enables an entity to perceive, analyze, determine response, and act appropriately in its environment." https://www.nsf.gov/statistics/srvyindustry/about/brdis/surveys/srvybrdis-2018-brd-1.pdf

References

BEIS and DCMS (2018), *Industrial Strategy Artificial Intelligence Sector Deal*, https://www.gov.uk/government/publications/artificial-intelligence-sector-deal. [10]

Bode, C. et al. (2018), "A Guide to the Dimensions Data Approach A collaborative approach to creating a modern infrastructure for data describing research: where we are and where we want to take it Contents", http://dx.doi.org/10.6084/m9.figshare.5783094. [3]

Cockburn, I. et al. (2018), *The Impact of Artificial Intelligence on Innovation*, http://www.nber.org/papers/w24449. [2]

EC (2018), *Artificial Intelligence for Europe; Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions; COM(2018) 237 final*. [9]

Freyman, C., J. Byrnes and J. Alexander (2016), "Machine-learning-based classification of research grant award records", *Research Evaluation*, Vol. 25/4, p. rvw016, http://dx.doi.org/10.1093/reseval/rvw016. [7]

Kawamura, T. et al. (2018), "Funding map using paragraph embedding based on semantic diversity", *Scientometrics*, pp. 1-18, http://dx.doi.org/10.1007/s11192-018-2783-x. [6]

Mikolov, T. et al. (2013), "Distributed Representations of Words and Phrases and their Compositionality", *arXiv:1310.4546*. [12]

National Science Foundation (2018), *Survey of Federal Funds for Research and Development Fiscal Years 2016–17*, https://ncsesdata.nsf.gov/fedfunds/2016/ (accessed on 27 March 2019). [11]

NSTC (2016), "Preparing for the Future of Artificial Intelligence", https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (accessed on 21 June 2018). [8]

OECD (2019 forthcoming), "Identifying and Measuring Developments in Artificial Intelligence", *OECD Science, Technology and Industry Working Papers*, https://doi.org/10.1787/18151965. [1]

Park, J. et al. (2016), *Analyzing NIH funding patterns over time with statistical text analysis*. [4]

Talley, E. et al. (2011), "Database of NIH grants using machine-learned categories and graphical clustering", *Nature Methods*, Vol. 8/6, pp. 443-444, http://dx.doi.org/10.1038/nmeth.1619. [5]

Annex. Lists of key terms

Table A.1. Clustering and treatment of AI-related terms in the AI-journal corpus

Cluster number	Quasi-synonyms or AI key terms	Status
1	OPTIMAL_SEARCH	AI key term
2	bioinformatics	Selected as AI key term
2	computational_biology	Selected as AI key term
3	wireless	Removed
4	semi_supervise	Removed
4	semisupervised	Removed
4	supervise	Removed
4	transductive	Removed
4	unsupervised	Removed
4	SUPERVISE_LEARN	AI key term
4	UNSUPERVISED_LEARN	AI key term
5	autonomous	Removed
5	drone	Removed
5	mechatronic	Removed
5	rover	Removed
5	teleoperated	Removed
5	HUMANOID_ROBOTIC	AI key term
5	ROBOTIC	AI key term
6	ad_hoc_network	Removed
6	SENSOR_NETWORK	AI key term
7	classification	Removed
8	inference_engine	Selected as AI key term
8	EXPERT_SYSTEM	AI key term
8	FUZZY_LOGIC	AI key term
9	k_near_neighbor	Selected as AI key term
9	naive_bayes	Selected as AI key term
10	kegg_pathway	Selected as AI key term
10	protein_protein_interaction	Selected as AI key term
10	GENE_ONTOLOGY	AI key term
11	cognitive_science	Removed
11	computer_science	Removed
12	bayes_classifier	Selected as AI key term
12	k_nn_classifier	Selected as AI key term
12	near_neighbor_classifier	Selected as AI key term
12	svm_classifier	Selected as AI key term
13	classifier	Removed
14	ontological	Removed
14	ontology	Removed
14	KNOWLEDGE_BASE	AI key term
15	logistic_regression	Removed
15	regression	Removed
16	ARTIFICIAL_INTELLIGENCE	AI key term
16	computational_intelligence	Selected as AI key term
16	MACHINE_LEARN	AI key term
16	PATTERN_RECOGNITION	AI key term
17	name_entity_recognition	Selected as AI key term

Cluster number	Quasi-synonyms or AI key terms	Status
17	opinion_mine	Selected as AI key term
17	text_categorization	Selected as AI key term
17	text_summarization	Selected as AI key term
17	word_sense_disambiguation	Selected as AI key term
18	fuzzy	Removed
19	markov_decision_process	Selected as AI key term
20	humanoid	Selected as AI key term
20	humanoid_robot	Removed
20	wheelchair	Removed
20	ROBOT	AI key term
21	IMAGE_ALIGNMENT	AI key term
21	camera_calibration	Selected as AI key term
21	mosaicing	Removed
21	non_rigid_registration	Selected as AI key term
21	registration	Removed
21	rigid_registration	Selected as AI key term
21	stereo_match	Selected as AI key term
22	artificial_neural_network	Selected as AI key term
22	fee_forward_neural_network	Selected as AI key term
22	multilayer_neural_network	Selected as AI key term
22	neural_net	Selected as AI key term
22	perceptron	Selected as AI key term
22	recurrent_neural_network	Selected as AI key term
22	NEURAL_NETWORK	AI key term
23	neuro_fuzzy	Selected as AI key term
23	radial_basis_function	Selected as AI key term
23	self_organize_map	Selected as AI key term
24	analog_vlsi	Selected as AI key term
24	NEUROMORPHIC_COMPUTE	AI key term
24	associative_memory	Selected as AI key term
24	neuromorphic	Removed
24	neuromorphic_hardware	Selected as AI key term
24	spike_neural_network	Selected as AI key term
25	back_propagation_neural_network	Selected as AI key term
25	bp_neural_network	Selected as AI key term
25	elman_network	Selected as AI key term
25	elman_neural_network	Selected as AI key term
25	less_square_support_vector_machine	Selected as AI key term
25	rbf_neural_network	Selected as AI key term
26	adaboost	Selected as AI key term
26	decision_tree	Selected as AI key term
26	random_forest	Selected as AI key term
26	ensemble	Removed
26	SUPPORT_VECTOR_MACHINE	AI key term
27	KNOWLEDGE_REPRESENTATION_AND_REASON	AI key term
27	commonsense_reason	Selected as AI key term
27	description_logic	Selected as AI key term
27	nonmonotonic_reason	Selected as AI key term
27	reason	Removed
28	IMAGE_MATCH	AI key term
28	alignment	Removed
28	match	Removed
28	template_match	Selected as AI key term

Cluster number	Quasi-synonyms or AI key terms	Status
29	dimensionality_reduction	Selected as AI key term
29	discriminant_analysis	Selected as AI key term
29	principal_component_analysis	Selected as AI key term
30	DEEP_LEARN	AI key term
30	convolutional_neural_network	Selected as AI key term
30	deep_belief_network	Selected as AI key term
30	deep_convolutional_neural_network	Selected as AI key term
30	deep_neural_network	Selected as AI key term
31	REINFORCEMENT_LEARN	AI key term
31	actor_critic	Selected as AI key term
31	sarsa	Selected as AI key term
32	COMPUTER_VISION	AI key term
32	computer_graphic	Removed
32	machine_vision	Selected as AI key term
33	person_re_identification	Selected as AI key term
34	knowledge	Removed
35	back_propagation	Selected as AI key term
36	ROBOT_SYSTEM	AI key term
36	manipulator	Removed
37	BAYESIAN_BELIEF_NETWORK	AI key term
37	COLLABORATIVE_SYSTEM	AI key term
37	MACHINE_INTELLIGENCE	AI key term
37	PATTERN_ANALYSIS	AI key term
37	SYMBOLIC_REASON	AI key term
37	SYSTEM_AND_CONTROL_THEORY	AI key term
37	neuromolecular	Removed
38	NATURAL_LANGUAGE_PROCESS	AI key term
38	machine_translation	Selected as AI key term
38	question_answer	Selected as AI key term
39	linear_discriminant	Selected as AI key term
39	multiclass_classification	Selected as AI key term
39	rbf_kernel	Selected as AI key term
40	SENSOR_DATUM_FUSION	AI key term

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

Table A.2. Selected list of AI key term

Selected terms for document retrieval within NIH and NSF databases

Terms before lemmatisation	Terms after lemmatisation	Source	Term Status
actor critic	actor critic	Scopus	non-core
adaboost	adaboost	Scopus	CORE
analog vlsi	analog vlsi	Scopus	non-core
artificial intelligence	artificial intelligence	M & C	CORE
artificial neural networks	artificial neural network	Scopus	CORE
associative memory	associative memory	Scopus	non-core
autonomous vehicle	autonomous vehicle	NSF	non-core
back propagation	back propagation	Scopus	CORE
back propagation neural network	back propagation neural network	Scopus	CORE
bayes classifier	bayes classifier	Scopus	non-core
bayesian belief networks	bayesian belief network	C	non-core
bioinformatics	bioinformatics	Scopus	non-core
bp neural network	bp neural network	Scopus	Merged to back propagation neural network
camera calibration	camera calibration	Scopus	non-core
collaborative systems	collaborative system	C	non-core
commonsense reasoning	commonsense reason	Scopus	non-core
computational biology	computational biology	Scopus	non-core
computational intelligence	computational intelligence	Scopus	CORE
computer vision	computer vision	C	CORE
convolutional neural network	convolutional neural network	Scopus	CORE
data mining	datum mine	NIH	non-core
decision tree	decision tree	Scopus	non-core
deep belief network	deep belief network	Scopus	CORE
deep convolutional neural network	deep convolutional neural network	Scopus	CORE
deep learning	deep learn	C	CORE
deep neural network	deep neural network	Scopus	CORE
description logic	description logic	Scopus	non-core
dimensionality reduction	dimensionality reduction	Scopus	non-core
discriminant analysis	discriminant analysis	Scopus	non-core
elman network	elman network	Scopus	CORE
elman neural network	elman neural network	Scopus	CORE
expert systems	expert system	M	CORE
feed forward neural network	fee forward neural network	Scopus	CORE
fuzzy logic	fuzzy logic	M	non-core
gene ontology	gene ontology	M	non-core
hidden markov model	hide markov model	NSF	non-core
humanoid	humanoid	NIH	non-core
humanoid robotics	humanoid robotic	C	Merged to robot
image alignment	image alignment	C	non-core
image matching	image match	C	non-core
inference engine	inference engine	Scopus	CORE
information retrieval	information retrieval	NSF	non-core
k nearest neighbors	k near neighbor	Scopus	Merged to near neighbor classifier

Terms before lemmatisation	Terms after lemmatisation	Source	Term Status
k nn classifier	k nn classifier	Scopus	Merged to near neighbor classifier
kegg pathway	kegg pathway	Scopus	non-core
knowledge bases	knowledge base	M	non-core
knowledge representation and reasoning	knowledge representation and reason	C	non-core
linear discriminant	linear discriminant	Scopus	non-core
machine intelligence	machine intelligence	C	CORE
machine learning	machine learn	M & C	CORE
machine translation	machine translation	Scopus	CORE
machine vision	machine vision	Scopus	CORE
markov decision process	markov decision process	Scopus	non-core
multiclass classification	multiclass classification	Scopus	non-core
multilayer neural network	multilayer neural network	Scopus	CORE
naive bayes	naive bayes	Scopus	non-core
name entity recognition	name entity recognition	Scopus	non-core
natural language processing	natural language process	M & C	CORE
nearest neighbor classifier	near neighbor classifier	Scopus	non-core
neural net	neural net	Scopus	Merged to neural network
neural networks	neural network	M & C	non-core
neuro fuzzy	neuro fuzzy	Scopus	non-core
neuromorphic computing	neuromorphic compute	C	non-core
neuromorphic hardware	neuromorphic hardware	Scopus	non-core
non rigid registration	non rigid registration	Scopus	non-core
nonmonotonic reasoning	nonmonotonic reason	Scopus	non-core
object recognition	object recognition	NSF	non-core
opinion mining	opinion mine	Scopus	non-core
optimal search	optimal search	C	non-core
pattern analysis	pattern analysis	C	non-core
pattern recognition	pattern recognition	C	non-core
perceptron	perceptron	Scopus	CORE
person re identification	person re identification	Scopus	non-core
principal component analysis	principal component analysis	Scopus	non-core
question answering	question answer	Scopus	non-core
radial basis function	radial basis function	Scopus	non-core
random forest	random forest	Scopus	CORE
rbf kernel	rbf kernel	Scopus	non-core
rbf neural network	rbf neural network	Scopus	CORE
recurrent neural network	recurrent neural network	Scopus	CORE
reinforcement learning	reinforcement learn	C	non-core
rigid registration	rigid registration	Scopus	non-core
robot systems	robot system	C	Merged to robot
robotics	robotic	M & C	Merged to robot
robots	robot	M & C	non-core
sarsa	sarsa	Scopus	non-core
self organizing map	self organize map	Scopus	CORE
sensor data fusion	sensor datum fusion	C	non-core
sensor networks	sensor network	C	non-core

Terms before lemmatisation	Terms after lemmatisation	Source	Term Status
speech recognition	speech recognition	NSF	non-core
spike neural network	spike neural network	Scopus	CORE
stereo matching	stereo match	Scopus	non-core
supervised learning	supervise learn	M	CORE
support vector machines	support vector machine	M	CORE
svm classifier	svm classifier	Scopus	CORE
symbolic reasoning	symbolic reason	C	non-core
systems and control theory	system and control theory	C	non-core
template matching	template match	Scopus	non-core
text categorization	text categorization	Scopus	non-core
text mining	text mine	NIH & NSF	non-core
text summarization	text summarization	Scopus	non-core
unsupervised learning	unsupervised learn	M & C	CORE
word sense disambiguation	word sense disambiguation	Scopus	non-core

Note: In the “Source” column, “M” refers to the MeSH or M-list, “C” refers to the list in Cockburn et al. (2018) (C-list), “Scopus” refers to “similar” terms to core terms in M-C lists retrieved in Scopus, “NIH” or “NSF” mean quasi-synonyms retrieved from the databases. The column on “Term Status” refers to how each term has been treated for analysis and retrieval in each corresponding database. “CORE” means that the term is used as a core AI term and not penalised for potential ambiguity, while “non-core” means the term is used but partly penalising when deciding which documents to selected.

Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018; on NIH RePORTER; and on NSF Award Search