U.S. Patent and Trademark Office
OFFICE OF THE CHIEF ECONOMIST
IP DATA HIGHLIGHTS NO. 5

Supplementary Material

Inventing AI: Tracing the diffusion of artificial intelligence with U.S. patents

Project Team:
Andrew A. Toole, Nicholas A. Pairolero,
Alexander V. Giczy, James Q. Forman, Jesse Frumkin, David B. Orange,
Anne Thomas Homescu, Steve Melnick, Christyann Pulliam, Matthew Such,
Kakali Chaki, Eric Nilsson, Ying Yu Chen, Vincent M. Gonzales, Ben M. Rifkin,
and Christian Hannon

October 2020

The views expressed are those of the individual authors and do not necessarily reflect official positions of the Office of the Chief Economist or the U. S. Patent and Trademark Office. USPTO Economic Working Papers are preliminary research being shared in a timely manner with the public in order to stimulate discussion, scholarly debate, and critical comment. For more information about the USPTO's Office of the Chief Economist, visit www.uspto.gov/economics.

Acknowledgements: The project team thanks Sandy Phetsaenngam, Teresa Verigan, and Melissa Harvey for their assistance in this report.

UNITED STATES
PATENT AND TRADEMARK OFFICE
uspto

Contents

INTRODUCTION

This supplement describes the methodology and data used in "Inventing AI: Tracing the diffusion of artificial intelligence with U.S. patents" published in October 2020 (hereinafter "Inventing AI").

BACKROUND

Scope of analysis

The patent landscape of "Inventing AI" encompasses publically available granted U.S. patents and U.S. patent application pre-grant publications (PGPub) published from 1976 through 2018. We employ a neural network (machine learning) classification model to identify patent documents in this "patent universe" that are relevant to AI. We then analyze this resulting AI patent landscape by examining patenting trends and diffusion across technologies, inventor-patentees, organizations, and geography.

Definition of patent document and related terms

We use the term "patent document" to be either a U.S. PGPub or a granted U.S. patent. We use the word "patent" by itself to mean a granted U.S. patent.[1] We refer to a U.S. non-provisional application for a patent as a "patent application." A "public patent application," then, is a patent application that is made available to the public, either by being published as a PGPub or, if there is no PGPub,[2] being published as a granted patent.

Each patent document has a publication date. For patents, the publication date is the date the patent was granted, i.e., the "grant date." For PGPubs, the publication date is the date that the PGPub was published. Additionally, we define the term "earliest U.S. publication date" to mean the first U.S. publication of a patent application, i.e., the publication date of the PGPub or of the granted patent, whichever is earlier. The "earliest U.S. publication date" excludes publication of any related international, foreign, or domestic applications.

[1] As described in the Scope of Analysis, the phrase "patent landscape" includes granted patents and PGPubs. The phrase "patent landscape" is a term of the art; see Trippe (2015).

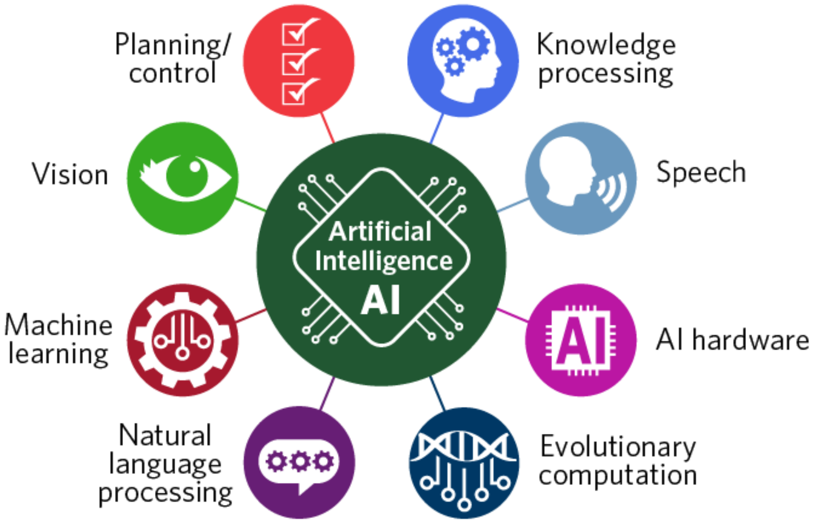
[2] The PGPub may not yet have been published, or the patent applicant submitted a complaint nonpublication request; see USPTO Manual of Patent Examination Procedures (MPEP) § 1112.

Each patent application has a filing date. For purpose of this study, we use the actual filing date of the application and not the "effective filing date," which would consider domestic and foreign priority. For international applications filed under the Patent Cooperation Treaty (PCT) and entering the U.S. national stage (i.e., a "371 national stage entry"), by treaty and by law the international filing date becomes the U.S. filing date. Thus, we use the term "filing or 371 date" to refer to the filing date for non-371 national state entry applications and the date on which an international application entered the U.S. national stage, i.e., the "371 date."

Definition of AI

A definition of "AI" is foundational to our analysis. We arrive at one comprising eight component technologies (Figure 1) after a literature review and discussions with USPTO patent examiners who review AI patent applications. See "Inventing AI" for definitions and examples of each component. These eight AI components are non-exclusive—a single patent document may be categorized in more than one. Our definition is also intentionally broad in that we do not limit our analysis to specific methods, such as "deep learning" or "neural networks," that have gained recent prominence.[3]

Figure 1: AI component technologies



Source: USPTO analysis.

[3] See Krohn, Beyleveld, and Bassens (2020), 3-19.

Analytic approach

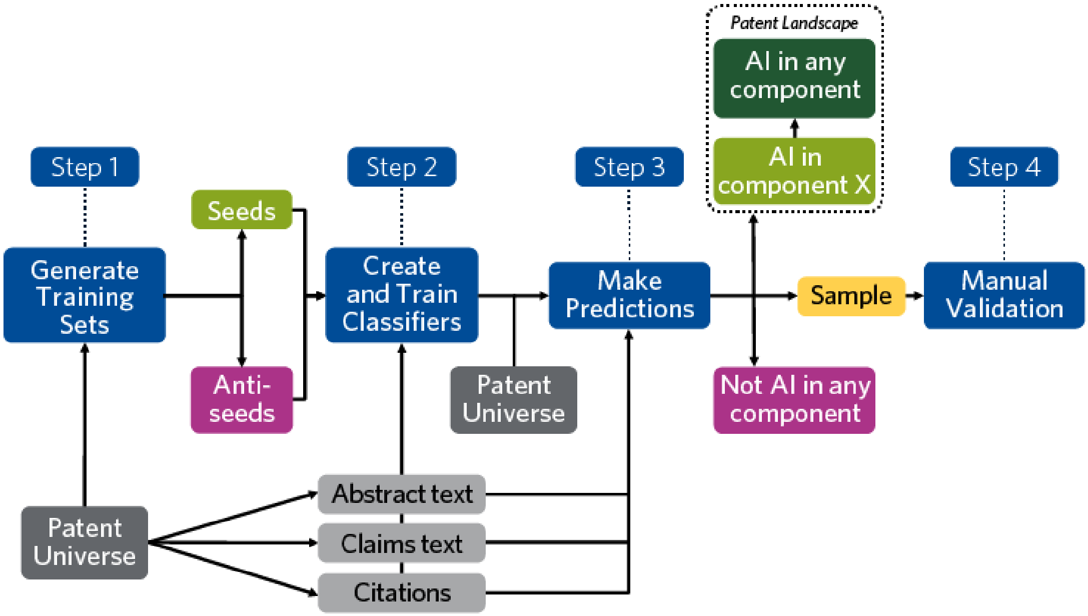
Our patent landscape methodology is summarized in Figure 2 (see also discussion in the Appendix of "Inventing AI").

The machine learning process of Steps 1-3 mirror the automated patent landscaping approach described by Abood and Feltenberger (2018). From a body of patent documents, a seed set providing positive examples of AI and an anti-seed set providing negative examples is constructed (Step 1) to train a classification model (Step 2) which is then used to classify the documents of the "patent universe" dataset (Step 3). As seen in Figure 2, we use the text of patent document abstracts and claims, in addition to patent citations, for the model. Since we have eight AI technology components, we create eight models, one for each component.

We also add a manual validation step (Step 4) by evaluating a random sample of model predictions against assessments made by experienced USPTO patent examiners. This validation allows us to assess the accuracy of our methodology.

Finally, we analyze relevant bibliographic and metadata, such as published date, patent classification, inventor-patentees, and owners-at-grant, of the resulting AI patent landscape to identify trends and characteristics of U.S. AI patenting.

Figure 2: AI patent landscape methodology overview



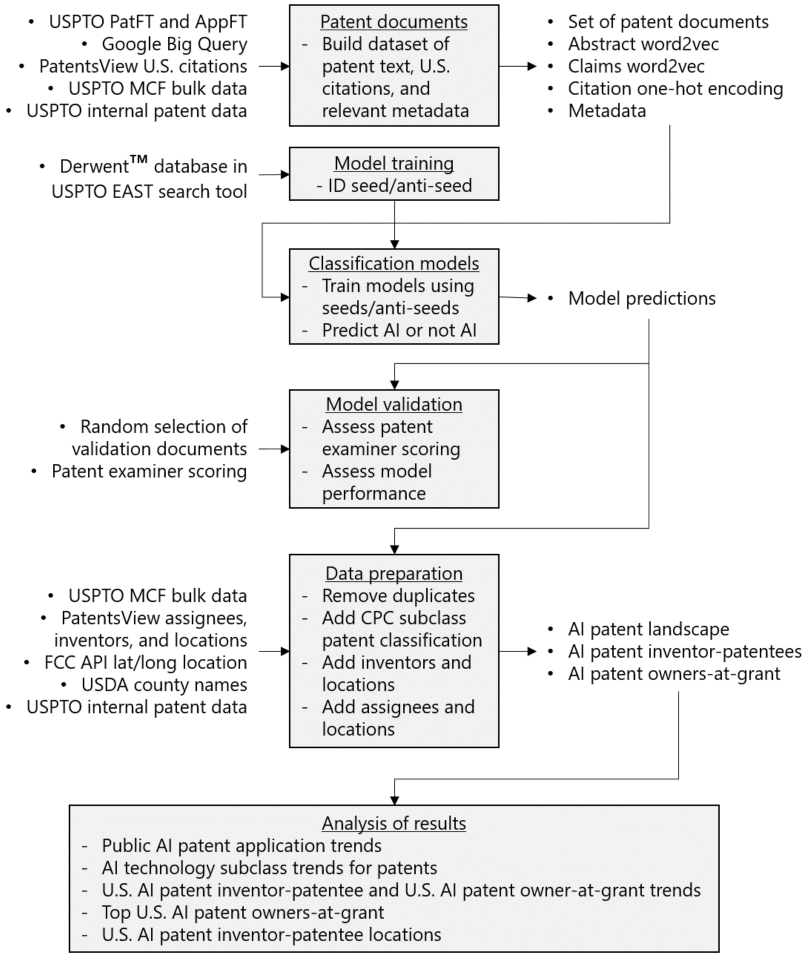
Source: USPTO based on methodology in Abood and Feltenberger (2018).

DATA CONSTRUCTION AND DESCRIPTION

Overview

Our data consists of patent document text, citations, classification, inventor-patentees and their locations, owners-at-grant and their locations, and other characteristics. The result is a dataset comprising 11,723,984 individual patent documents published between January 1976 and February 2019.[4] Figure 3 below provides an overview of its construction.

Figure 3: Data construction overview


• USPTO PatFT and AppFT
• Google Big Query
• PatentsView U.S. citations
• USPTO MCF bulk data
• USPTO internal patent data
Patent documents
- Build dataset of patent text, U.S. citations, and relevant metadata
• Set of patent documents
• Abstract word2vec
• Claims word2vec
• Citation one-hot encoding
• Metadata
• Derwent™ database in USPTO EAST search tool
Model training
- ID seed/anti-seed
Classification models
- Train models using seeds/anti-seeds
- Predict AI or not AI
• Model predictions
• Random selection of validation documents
• Patent examiner scoring
Model validation
- Assess patent examiner scoring
- Assess model performance
• USPTO MCF bulk data
• PatentsView assignees, inventors, and locations
• FCC API lat/long location
• USDA county names
• USPTO internal patent data
Data preparation
- Remove duplicates
- Add CPC subclass patent classification
- Add inventors and locations
- Add assignees and locations
• AI patent landscape
• AI patent inventor-patentees
• AI patent owners-at-grant
Analysis of results
- Public AI patent application trends
- AI technology subclass trends for patents
- U.S. AI patent inventor-patentee and U.S. AI patent owner-at-grant trends
- Top U.S. AI patent owners-at-grant
- U.S. AI patent inventor-patentee locations

Notes: API = application programming interface; AppFT = Application Full Text data; CPC = Cooperative Patent Classification; EAST = Examiner Automated Search Tool; FCC = Federal Communications Commission; FIPS = Federal Information Processing System; MCF = Master Classification File; PatFT = Patent Full Text data; USDA = U.S. Department of Agriculture.

[4] As later discussed, our analysis of the resulting patent landscape is based on full calendar years and hence ends in December 2018.

Patent document dataset construction

Patent documents

Our machine learning classification algorithm is trained on patent document text and citations and makes predictions using the same. We limit the patent documents to those having text publically available through the USPTO Bulk Data Storage System (BDSS).[5] For patents, we use the Patent Grant Full Text Data (PatFT), which starts in January 1976. For PGPubs, we use the Application Full Text Data (AppFT), which starts on March 15, 2001.[6]

Hence, our data set starts in January 1976 when the full text of granted patents is available. It ends in February 2019 and contains a total of 11,723,984 patent documents: 6,208,365 patents (53.0%) and 5,515,619 PGPubs (47.0%).

In addition to the text and citation relationships, the patent document dataset includes additional metadata, such as patent families and classification, which we use to create the anti-seed set of negative examples for training. We source this metadata from PatentsView,[7] the USPTO BDSS Master Classification File (MCF),[8] and internal USPTO databases.

Patent document text, pre-processing, and word2vec embedding

We use the text from patent document abstracts and claims. The abstract provides a short summary of the invention and what is new in the art.[9] The claims "particularly point out and distinctly claim the subject matter which the inventor or joint inventor regards as his or her invention"[10] and thus establish the technical and legal bounds of the patent. Abood and Feltenberger's (2018) implementation focuses on abstract text, but they state parts of the patent document may be used, such as patent classification and claim text.[11] Our decision to include claims text enables us to consider the precise technical and legal scope of the invention.

[5] See https://bulkdata.uspto.gov/

[6] The publication of patent applications as PGPubs began with the American Inventors Protection Act (APIA), enacted November 29, 1999.

[7] See https://www.uspto.gov/ip-policy/economic-research/patentsview and data downloads at https://www.patentsview.org

[8] See https://bulkdata.uspto.gov/data/patent/classification/cpc/

[9] See MPEP § 608.01(b).

[10] MPEP § 608.01(k); see also §§ 608.01(i)-(o).

[11] Abood and Feltenberger (2018), 119-21.

For the claims text we use PatFT and AppFT for granted patents and PGPubs, respectively. For abstract text we use PatFT for patents and Google Big Query[12] for PGPubs. We do not use AppFT for the PGPub abstract text due to internal resource constraints at the time of processing the data.

We pre-process claims text and abstract text in the same manner. Pre-processing includes: lowercasing text; removing starting numbers, symbols, and formulas; cleaning special characters; and removing extra spaces. Additionally we remove parenthetical text—numerals referencing items in figures are placed in parenthesis, for example, and the current status of claims in PGPubs may also be within parentheses at the beginning of the claim. We also remove cancelled claims.[13] Finally, we concatenate all the claims of a patent document into a single string of text. We similarly concatenate the sentences of the abstract into a single text string.

Following pre-processing, we use word2vec to separately encode the text of the abstracts and of the claims. Each word is encoded as a 300-dimension vector.[14]

Citations and one-hot encoding

In addition to text from patent document abstracts and claims, our machine learning algorithm also uses patent citations:[15] both backward citations (i.e., the references a given patent document cites) and forward citations (i.e., the documents citing a given patent document). We use only citations to U.S. patent documents—granted U.S. patents and U.S. PGPubs.

We use PatentsView to get the citations that are listed on granted patents. PGPubs do not list citations and thus citation data is not directly available. If a PGPub was granted as a patent, we use the citations on the granted patent for that PGPub. If a PGPub was not granted as a patent (i.e., it was abandoned or is still under review), then no citations are used for that PGPub.

From these data we construct citation relationships. We then use one-hot encoding as per Abood and Feltenberger (2018)[16] to capture those relationships in a machine-readable format—

[12] See Wetherbee at https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data. We also download the abstract text for granted patents from Google Big Query; this text should be equivalent to the patent abstract text in PatFT.

[13] The claims of a patent application may change during its examination to address rejections over the prior art, other rejections, and informalities as made by the patent examiner; see MPEP § 706.

[14] Abood and Feltenberger (2018), 116-7; our word2vec approach uses code from Persiyanov (2018), see https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/Any2Vec_Filebased.ipynb

[15] Abood and Feltenberger (2018) also use patent citations or references (117-8).

[16] Abood and Feltenberger (2018), 117-8.

the list of all citations is assigned a positon in a vector, and for each patent document we create a citation vector having the value of "1" in the position corresponding to its citations (and "0" otherwise). Given the large number of possible citations, the citation vectors are very sparse.

Additional metadata

The creation of the training data per Abood and Feltenberger's (2018) approach finds patent documents similar to the seed set of positive examples, called "expansion" documents. These expansion documents are excluded from selection of the anti-seed of negative examples. Similarity is based on patent families and patent classification.[17] Thus, we add this metadata to our patent document dataset.

For patent families we use an internal USPTO data set. For patent classification we use the USPTO Bulk Data Storage System (BDSS) Master Classification File (MCF).[18] The MCF contains two files, one for PGPubs and one for granted patents, and lists the classifications of each document per the CPC system. Each patent document may have several CPC codes, and we use the "CPC First" code since it best represents the overall invention.[19]

Patent landscape analytic data construction

We refer to "analytic data" as the predictions from the AI classification models forming our AI patent landscape plus additional patent document bibliographic and metadata to allow us to contextualize and analyze the AI patent landscape. These additional patent document data includes inventor-patentees, owners-at-grant, and their locations. In the following sections we describe these data; but first, we discuss the preparation steps necessary to analyze the resultant landscape.

Data preparation

First, we to drop all reissue patents to ensure we are analyzing only utility patent grants and PGPubs.[20]

Second, we remove duplicate patent documents. Since the claims may change during the examination of the patent application, the claims of the PGPub may differ from those of the

[17] Abood and Feltenberger (2018), 109-15.

[18] See https://bulkdata.uspto.gov/data/patent/classification/cpc/

[19] See MPEP § 905.03(a).

[20] Reissue patents may include non-utility patents, and we drop the 17,351 reissue patents in our dataset. Since our analysis is focused on the earliest U.S. publication date of patent documents, the loss of utility reissue patents is not significant.

granted patent. Our classification model includes claims text, and thus it is appropriate to include both types of patent documents in the patent document dataset for machine learning. In the analysis of the resultant patent landscape, however, we carry forward only one document to avoid double counting of patent applications. Additionally, a small number of patent applications may contain more than one PGPub, e.g., a corrected or a republished PGPub. Moreover, due to data errors, a single application may have more than one granted patent. To remove duplicates, we use the following rules for each patent application:

- If a granted patent does not have PGPub, keep the patent and its publication date, which is the earliest US publication date of that patent application.
- If a PGPub and a granted patent exist for a single application, keep the granted patent and use the AI model prediction of the granted patent as well as its associated bibliographic data. Additionally, keep the published date of the earlier PGPub as the earliest U.S. publication date of the patent application.
- If a PGPub does not result in a granted patent,[21] keep the PGPub and its publication date, which is the earliest US publication date of that patent application.
- If a PGPub without a granted patent has more than one PGPub, keep the most recently published PGPub,[22] using the AI model prediction and associated bibliographic data of that most recent PGPub. Additionally, keep the published date of the first PGPub as the earliest U.S. publication date of the patent application.
- If a single patent application has more than one granted patent associated with it, keep all the granted patents since, by definition, each granted patent is a novel and non-obvious invention.[23]

Third, we address the multiple dates associated with a patent application. As noted above, we keep the "earliest U.S. publication date" when removing duplicate documents. For national stage entry applications under the PCT, we also include the "371 date."

The end result is a total of 8,444,624 patent documents in our analytic data, comprising 6,191,014 patents (73.3%) and 2,253,610 PGPubs (26.7%).

[21] This scenario exists if a patent application was abandoned or the patent application was still under review as of the construction of our patent document dataset.

[22] A total of 8,130 patent applications have multiple PGPubs; 4,031 distinct applications have multiple PGPubs and no granted patent.

[23] A total of 1,670 patent applications in our data have two patents.

CPC technology classification

For CPC technology classifications in the analytic data we use the MCF[24] augmented by internal USPTO data from the USPTO Patent Application Locating and Monitoring (PALM) system.[25]

We use CPC classification only to the CPC subclass level, e.g., G06N for computer systems based on specific computational models. As noted above, a patent document may be classified in multiple CPC areas, i.e., "CPC First" pertaining to focus of the invention as a whole, "CPC Inventive" for additional inventive subject matter, and "CPC Additional" capturing other technical material.[26] We use only the "CPC First" classification in our analysis.

As discussed in the "Findings: Methodology and Results" section below, we analyze the technology classification of only granted patents. The number of granted patents missing CPC First subclass is 1,945 (0.02% of patents).

Inventor-patentees and owners-at-grant

We use PatentsView to add inventor-patentees (i.e., inventors who seek patents) and owners-at-grant (i.e., assignees at patent grant), plus their locations. The advantage of using PatentsView is that it contains disambiguated names and locations, which helps overcome small differences in the same entity that may be on different printed patent documents.[27] PatentsView, however, does not include PGPubs,[28] and hence our analysis of inventor-patentees and owners-at-grant is limited to granted patents.

PatentsView also provides disambiguated locations for inventor-patentees and owners-at-grant. If U.S. county Federal Information Processing System (FIPS) codes are missing, we use the location latitude and longitude from PatentsView (if available) to query an API hosted by the Federal Communications Commission (FCC) to look up the FIPS code. If there is consistency between the state from the FCC API query[29] and the state in PatentsView we augment the location data. Finally, we use a look-up table of FIPS codes and U.S. county names from the U.S. Department of Agriculture (USDA) Natural Resources Conservation Service (NRCS).[30]

[24] Although we added CPC classification to the patent document data set so as create the training data set, CPC classification was not carried forward in the analysis. Hence, we need to re-add the classification following prediction.

[25] All but a handful of data is from the MCF; a total of 2,629 (0.03%) of patents have missing CPC data.

[26] See MPEP § 905.03(a).

[27] E.g., "Incorporated" on one document but "Inc." on another.

[28] As of the date of this supplemental, PatentsView is in the process of adding PGPub data.

[29] See https://geo.fcc.gov/api/census/

[30] See https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697

As discussed in "Inventing AI," we use assignees at patent grant as provided on the granted patent for owners-at-grant. In addition to a lack of quality data for patent reassignments, focusing on the patent owners at the time of patent grant is consistent with our analysis that uses the patent grant date. Additionally, we do not include inventors who have not transferred their rights prior to patent grant, nor do we include non-inventor patent applicants. We do not include these patent owners since PatentsView does not extend its assignee disambiguation to inventors and to non-inventor applicants.

Of the granted patents in our analytic data, 773 (0.01%) are missing a disambiguated inventor-patentee and 773,238 (12.5%) are missing a disambiguated owner-at-grant. The analytic data contains 1,708,335 unique U.S. inventor-patentees; of these, 11,243 (0.7%) are missing a county FIPS code. It also contains 219,722 unique U.S. owners-at-grant; of these, 2,088 (1.0%) are missing a county FIPS code.

PATENT LANDSCAPE METHODOLOGY

We now discuss the methodology we use to create the AI patent landscape, moving through each of the steps in Figure 2 above: generating the training set, creating and training the classifiers, making predictions on whether each patent document contains each AI component technology, and manually validating the classifiers to compare our predictions with assessments of experienced patent examiners. With the exception of the last step, manual validation, our methodology is based on the automated patent landscaping process in Abood and Feltenberger (2018). We also leverage code implementing the process as posted by Feltenberger on GitHub.[31]

Generate training sets (step 1)

Supervised machine learning classification models require a set of positive seed and negative anti-seed examples to train the models. Abood and Feltenberger (2018) start by generating a narrow seed set, expand the seed set using patent family citations and classification codes (called Level 1 and 2, or L1 and L2, expansions), and finally select the anti-seed set from remaining patent documents outside the seed, L1, and L2 documents.[32] L1 and L2 are less likely to be related to the topic of interest (L2 less likely than L1), and excluding both from the anti-seed set increases the chance the negative examples are not related to the topic of interest.[33]

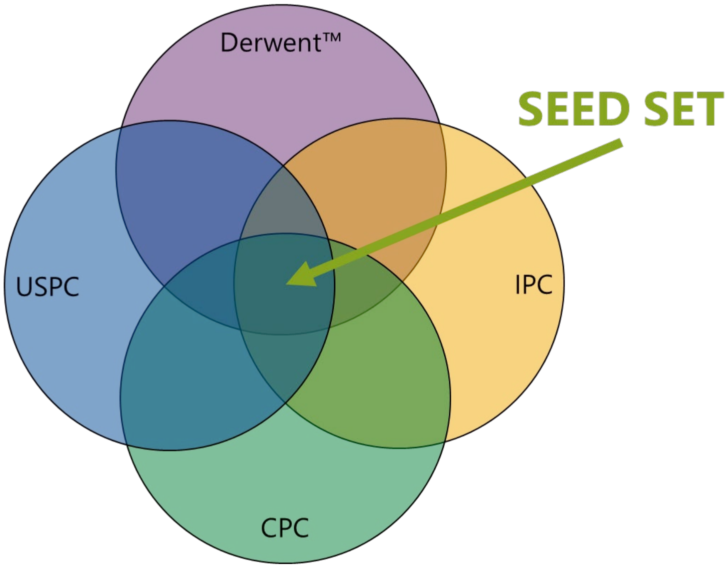
[31] See https://github.com/google/patents-public-data/tree/master/models/landscaping.

[32] Abood and Feltenberger (2018), 105 and 109.

[33] Abood and Feltenberger (2018), 115-6.

Since we have eight AI technology components, we have eight classification models and hence eight individual seed sets. We generate these seed sets using the USPTO [Patent] Examiner Automated Search Tool (EAST)[34] to query the Clarivate Derwent World Patent Index™[35] for patent documents in classifications relevant to the AI technology component. In general, the seed set documents are at the intersection of the CPC system, the International Patent Classification (IPC) system, the U.S. Patent Classification (USPC) system, and Derwent's patent index, as illustrated in Figure 4 below. Further, the seed set documents are limited to U.S. patent documents. Appendix I details the specific queries we use for each seed set.[36]

Figure 4: General process to generate seed sets



We then perform the L1 and L2 expansions from each seed set as follows:[37]

- For L1: First, determine the family members of each of the patent documents in the seed set, find the backward and forward citations of those family members, and determine the family members of those citations ("family-citation-family expansion"). Second, determine the patent document share for each CPC code comprising the seed set documents, and for each CPC code that is 50 times the corresponding share in the patent document dataset, determine the patent documents within that CPC ("code

[34] See description the USPTO Public Search Facility webpage: https://www.uspto.gov/learning-and-resources/support-centers/public-search-facility/public-search-facility and MPEP § 902.03(e).

[35] See https://clarivate.com/derwent/solutions/derwent-world-patent-index-dwpi/.

[36] The seed set for AI hardware is the furthest departure from this intersection approach.

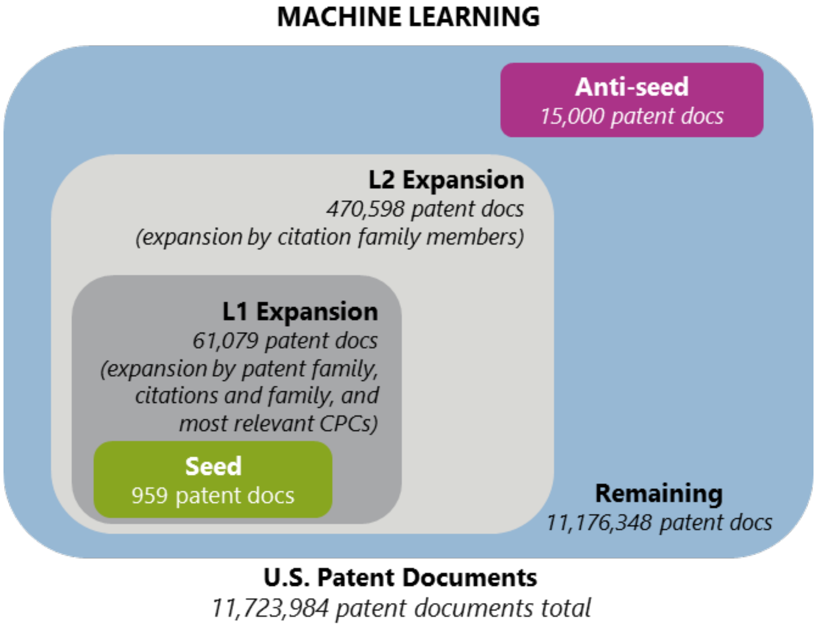
[37] See Abood and Feltenberger (2018), 109-15.

expansion"). Third, the L1 expansion comprises all the patent documents in the family-citation-family expansion or the code expansion.

- For L2: Determine the forward and backward citations of each of the L1 patent documents and their family members. The L2 expansion comprises all the citation and family member patent documents of L1.

The anti-seed set for an AI technology component is a random sample from the patent documents that are not in the seed set, L1 expansion, and L2 expansion for that component. We select 15,000 patents documents to be the size for each AI component anti-seed set. Figure 5 illustrates the results for the machine learning component. Results for all AI components are presented in Table 1.

Figure 5: Seed, L1, L2 and anti-seed generation for machine learning component



Source: USPTO analysis based on methodology in Abood and Feltenberger (2018).

Table 1: Number patent documents in each group by AI component

AI Component	Seed	L1 Expansion	L2 Expansion	Anti-seed	Remaining	Total
Machine learning	959	61,079	470,598	15,000	11,176,348	11,723,984
Evolutionary computation	82	59,316	349,570	15,000	11,300,016	11,723,984
Natural language processing	1,084	82,762	396,564	15,000	11,228,574	11,723,984
Speech	763	92,346	427,397	15,000	11,188,478	11,723,984
Vision	803	166,434	629,961	15,000	10,911,786	11,723,984
Knowledge processing	661	89,419	518,719	15,000	11,100,185	11,723,984
Planning/control	1,451	179,753	799,828	15,000	10,727,952	11,723,984
AI hardware	2,659	117,056	838,484	15,000	10,750,785	11,723,984

Source: USPTO analysis.

From Table 1 we see an average of about 1,000 documents for the seed set (ranging from a low of 82 for evolutionary computation to a high of 2,659 for AI hardware). The seed and anti-seed are presumed to be "gold standards" for positive and negative patent documents, respectively. The other patents documents—those in the L1, L2 and remaining groups—are unknowns.[38] In the next step, we create and train machine learning models to classify those unknowns.

Create and train classifiers (step 2)

We use the corresponding seed and anti-seed sets to train eight machine learning classifier models, one model for each AI component. Specifically, we used the abstract text, claims text and citations of the seed and anti-seed patent documents. The model architectures consist of long short-term memory (LSTM) neural networks as per Abood and Feltenberger.[39] As previously described, to help capture contextual information in the abstract and claims text we

[38] While we have a degree of confidence the L1 and L2 patent documents may be representative of the AI component, the primary purpose of the expansions are to create the anti-seed.

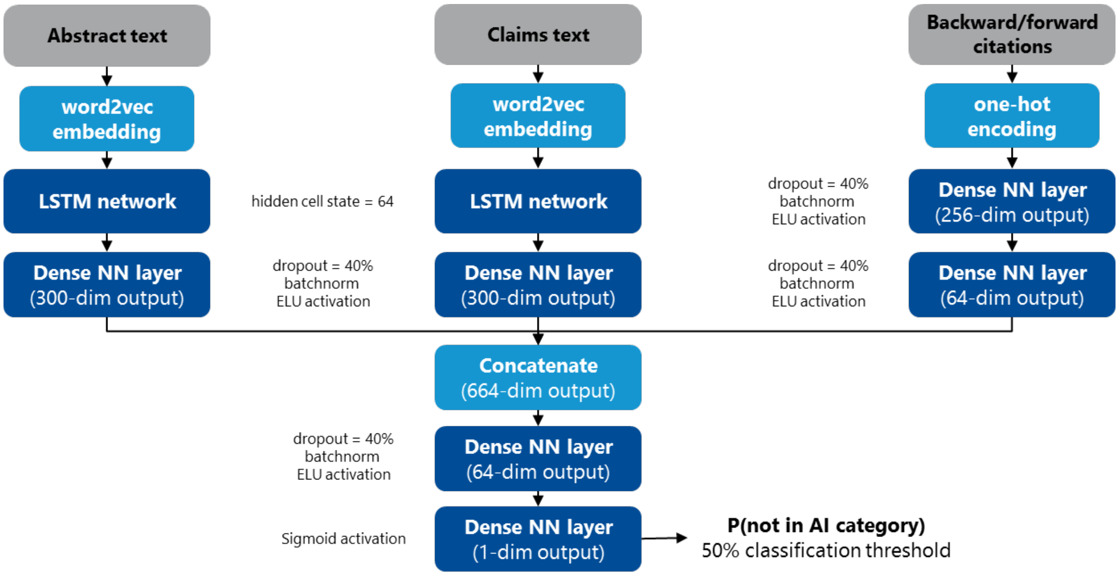
[39] Hence subtracting the component model score for each patent document results in a probability of the document being AI in that component.

encode this text using word2vec, one embedding for abstracts and another for claims, trained on the abstracts and claims from the entire patent document dataset, respectively.

Figure 6 provides an overview of the model architecture. Each word of the abstract text and the claims text of a patent document is translated into its 300-dimension word2vec embedded vector and input into a separate LSTM neural network. Use of two networks allows the model to consider abstract and claims text separately from each other. As these neural networks process each word they pass a 64-dimension vector (hidden state) from the output of one word to the input of the next word—this internal structure allows the LSTM networks to consider the sequence of words in the abstract and claims.[40] Meanwhile, the forward and backward citations of the patent document are one-hot encoded and input into two dense neural network layers. The outputs of the abstract LSTM network, claims LSTM network, and citation dense network are concatenated into a 664-dimension vector that is input into a 64-neuron dense neural network. The output of this dense layer is finally input into a single neural network layer having a sigmoid activation function. This functions produces a number between 0.0 and 1.0 interpreted as the probability of the patent document not being in that AI technology component.

To implement the models we use code posted by Feltenberger on GitHub,[41] modifying it to also include claims text.

Figure 6: Overview of classification model architecture


Abstract text
Claims text
Backward/forward citations
word2vec embedding
word2vec embedding
one-hot encoding
LSTM network
hidden cell state = 64
LSTM network
dropout = 40% batchnorm ELU activation
Dense NN layer (256-dim output)
Dense NN layer (300-dim output)
dropout = 40% batchnorm ELU activation
Dense NN layer (300-dim output)
dropout = 40% batchnorm ELU activation
Dense NN layer (64-dim output)
Concatenate (664-dim output)
dropout = 40% batchnorm ELU activation
Dense NN layer (64-dim output)
Sigmoid activation
Dense NN layer (1-dim output)
P(not in AI category)
50% classification threshold

Source: Abood and Feltenberger (2018), modified by USPTO.
Notes: The model is trained on the "not AI" anti-seed category. Key: dim = dimension; NN = neural network; LSTM = long short-term memory; ELU = exponential linear unit.

[40] See Krohn, Beyleveld, and Bassens (2020), 244-7.

[41] See https://github.com/google/patents-public-data/tree/master/models/landscaping

Make predictions (step 3)

Once the models were trained on the seed and anti-seed sets they were used to predict whether each document in the patent document dataset contains each AI component technology. Each model outputs a number between 0 and 1 (or 0% to 100%), interpreted as the probability of being in that particular AI component technology. We use a 50% threshold to determine whether a given patent document is in the AI component—those equal to or above the threshold are in the technology, and those below are not.[42] Additionally, we consolidate the results from the eight models such that if one model predicts AI in a component technology, then the patent document is labeled as having "any AI."

The prediction results are summarized in Tables 2 and 3 below, both as the number and percentage of patent documents by AI component and by "any AI."

[42] The classification models output the probability of a document not being AI; for convenience we refer to the models as predicting "AI," which is 1.0 – p(not AI).

Table 2: Model predictions—number of AI and non-AI patent documents by AI component

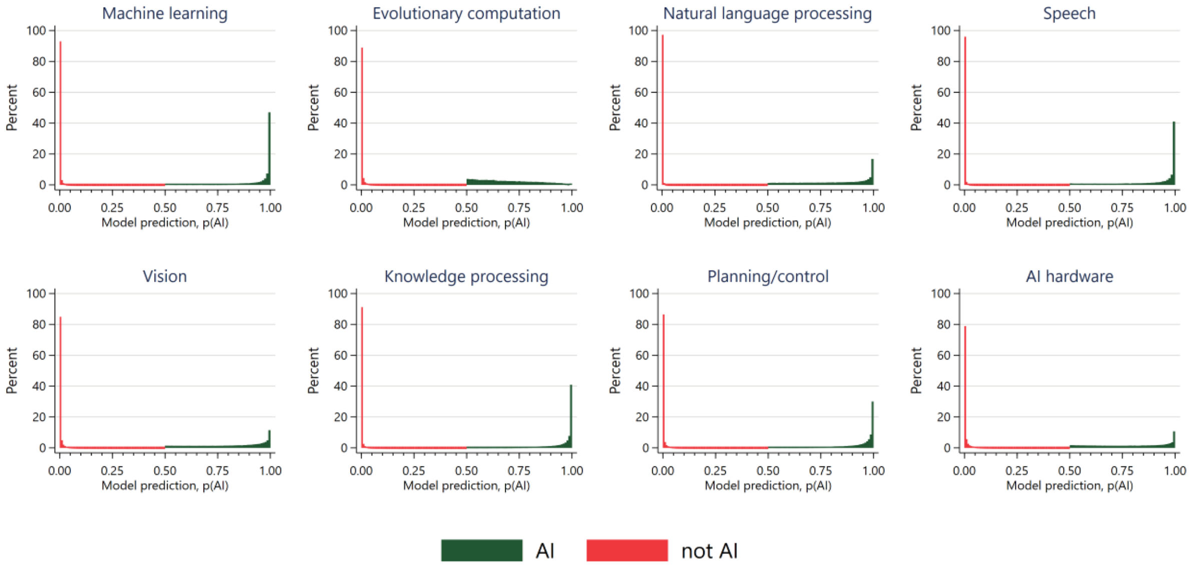
AI Component	Predict	Seed	Anti-seed	L1 Expansion	L2 Expansion	Remaining	Total
Machine learning	AI	922	57	25,530	37,114	85,655	149,278
	Not AI	37	14,943	35,549	433,484	11,090,693	11,574,706
Evolutionary computation	AI	66	12	5,400	8,243	25,534	39,255
	Not AI	16	14,988	53,916	341,327	11,274,482	11,684,729
Natural language processing	AI	1,056	23	45,316	45,138	40,390	131,923
	Not AI	28	14,977	37,446	351,426	11,188,184	11,592,061
Speech	AI	739	6	39,941	17,969	15,993	74,648
	Not AI	24	14,994	52,405	409,428	11,172,485	11,649,336
Vision	AI	766	75	90,776	96,455	126,503	314,575
	Not AI	37	14,925	75,658	533,506	10,785,283	11,409,409
Knowledge processing	AI	653	417	68,056	188,191	447,695	705,012
	Not AI	8	14,583	21,363	330,528	10,652,490	11,018,972
Planning/ control	AI	1,390	279	12,8545	270,070	368,887	769,171
	Not AI	61	14,721	51,208	529,758	10,359,065	10,954,813
AI hardware	AI	2,118	109	50,829	154,333	193,158	400,547
	Not AI	541	14,891	66,227	684,151	10,557,627	11,323,437
Any AI	AI	7,299	3,670	350,091	460,585	440,396	1,262,041
	Not AI	346	99,105	166,749	1,136,433	9,059,310	10,461,943

Table 3: Model predictions—percent of AI and non-AI patent documents by AI component

AI Component	Predict	Seed	Anti-seed	L1 Expansion	L2 Expansion	Remaining	Total
Machine learning	AI	96.1%	0.4%	41.8%	7.9%	0.8%	1.3%
	Not AI	3.9%	99.6%	58.2%	92.1%	99.2%	98.7%
Evolutionary computation	AI	80.5%	0.1%	9.1%	2.4%	0.2%	0.3%
	Not AI	19.5%	99.9%	90.9%	97.6%	99.8%	99.7%
Natural language processing	AI	97.4%	0.2%	54.8%	11.4%	0.4%	1.1%
	Not AI	2.6%	99.8%	45.2%	88.6%	99.6%	98.9%
Speech	AI	96.9%	0.0%	43.3%	4.2%	0.1%	0.6%
	Not AI	3.1%	100.0%	56.7%	95.8%	99.9%	99.4%
Vision	AI	95.4%	0.5%	54.5%	15.3%	1.2%	2.7%
	Not AI	4.6%	99.5%	45.5%	84.7%	98.8%	97.3%
Knowledge processing	AI	98.8%	2.8%	76.1%	36.3%	4.0%	6.0%
	Not AI	1.2%	97.2%	23.9%	63.7%	96.0%	94.0%
Planning/ control	AI	95.8%	1.9%	71.5%	33.8%	3.4%	6.6%
	Not AI	4.2%	98.1%	28.5%	66.2%	96.6%	93.4%
AI hardware	AI	79.7%	0.7%	43.4%	18.4%	1.8%	3.4%
	Not AI	20.3%	99.3%	56.6%	81.6%	98.2%	96.6%
Any AI	AI	95.5%	3.6%	67.7%	28.8%	4.6%	10.8%
	Not AI	4.5%	96.4%	32.3%	71.2%	95.4%	89.2%

Since the models produce a probability of each patent document being in the AI component or not, we plot the distributions of the "AI" and "not AI" predictions to see how differentiated they are. Figure 7 provides such plots for each AI component. In each sub-figure, the red histogram plots the distribution of predictions for patent documents that were predicted not to be in the AI component, and the green histogram plots the predictions for those predicted to be in the AI component. We see the probability distribution for "not AI" (red histogram) spikes close to 0.00 for all models. For the probability distribution for "AI" (green histrogram), all models except for evolutionary computation spike near 1.0. These spikes at 1.0 and 0.0 indicate that most of the models are highly certain about their predictions. Regarding evolutionary computation, the p(AI) distribution is relatively flat, with most of the positive predictions being close to 0.5, indicating that this model is highly uncertain about its predictions.

Figure 7: Distribution of model predictions for AI and not AI



Note: Each graph plots two separate distributions, one for AI and on for not AI.

To further assess our classifications models and the machine learning process used in patent landscaping we added a manual validation step, which is discussed in the next section.

Manual validation (step 4)

Methodology

To explore all aspects of the machine learning patent landscape process, we randomly selected documents from the seed set, anti-seed set, and the combined L1, L2, and remaining sets of patent documents.

Sampling from the seed and the anti-seed enables us to gain insight into the generation of these sets as well as model training. However, to characterize predictive performance we used only the samples from the combined L1, L2, and remaining sets (i.e., the patent documents not used to train the model).

Manual validation is labor-intensive, and due to resource limitations we select patent documents at a "consolidated group" level to enable us to assess the patent landscaping process as a whole, as opposed to the performance of individual classification models.[43] If a patent document is used in the seed set for any classification model, then it is placed in the consolidated seed group; if used in any L1 expansion, then it is placed in the consolidated L1 group; and so on for L2 and anti-seed.[44] We then randomly select 216 documents each from the consolidated seed and anti-seed sets, and 368 documents from the combined L1, L2, and remaining groups.

We use a total of four experienced patent examiners plus a fifth patent examiner adjudicator. Each patent document is reviewed by two experienced patent examiners and annotated as being AI or not in each of the eight AI technology components; a document may be annotated as AI in multiple components.

We divide the 800 patent documents such that each patent examiner reviews about 400. The patent documents in each consolidated group (seed; anti-seed; and L1, L2, and remaining) are allocated approximately evenly among each of the six pairs of patent examiners.[45] In the same manner as the model predictions, we integrate the eight AI technology component annotations into a single "any AI" annotations. If the resulting "any AI" conclusion of the two patent examiner annotators disagree, then the difference is resolved by the adjudicating patent examiner. Adjudication is performed for the overall "any AI" conclusion and not for individual AI technology components.

Before comparing the results of this manual validation of patent documents with model predictions, we first assess how well the patent examiner annotators agree with each other. This assessment provides some information about how challenging the problem of identifying AI in patent documents is. Confusion matrixes provide one approach to look at agreements and disagreements.

[43] Measuring the performance of the eight individual models would have required samples for each model, and we did not have the necessary labor-intensive resources.

[44] The resulting order is thus: seed set, L1 expansion, L2 expansion, anti-seed set, and remaining. An improved order would be seed, anti-seed, L1, L2, and remaining so as to avoid an anti-seed document in one model from being placed in the L1 or L2 consolidated group. In our case, the improved order would have placed 16,511 more documents in the consolidated anti-seed group, 2,798 fewer documents in L1, and 13,713 fewer documents in L2. Of the 800 randomly selected documents in the manual validation, however, only 1 document would have been moved from L2 to anti-seed. Hence the impact to our analysis is minimal.

[45] The pairs are patent examiners 1-2, 1-3, 1-4, 2-3, 2-4, and 3-4. Each pair reviewed 36 patent documents in the consolidated seed group (216 total), 36 patent documents in the consolidated anti-seed group (216 total), and 61 or 63 patent documents in the consolidated L1, L2, and remaining group (368 total).

Analysis of annotator agreement: confusion matrix

There are two methods to analyze the annotation results using confusion matrixes (also known as contingency tables). The first is to compare whether the two patent examiners reviewing a single document agree or disagree. The second is to incorporate the adjudicating patent examiner: for each document having a disagreement, the adjudicator agrees with one of the reviewing patent examiners and disagrees with the other (if the two reviewing patent examiners agree, then no adjudication occurs; we assume the adjudicating patent examiner would agree with the two reviewing patent examiners). Tables 4 and 5 provide results for each method.

Table 4: Confusion matrix and metrics for Annotator A vs. Annotator B (first method)

Confusion Matrixes	Seed		Anti-seed		L1, L2, and remaining	
	B: any AI	B: not AI	B: any AI	B: not AI	B: any AI	B: not AI
A: any AI	185	7	14	26	31	43
A: not AI	14	10	11	165	41	253
Metrics						
# documents	216		216		368	
Precision	0.9635		0.9375		0.4189	
Recall	0.9296		0.8639		0.4306	
Accuracy	0.9028		0.8287		0.7717	
F1 score	0.9463		0.8992		0.4247	

Note: Analysis compares resulting "any AI" annotation of patent examiner A (first examiner reviewing a document) compared to patent examiner B (second examiner reviewing the same document). For the anti-seed set, the metrics are calculated such that "not AI" is the positive result. See discussion for metric definitions.

Table 5: Confusion matrix and metrics for Annotators A and B with adjudication (second method)

Confusion Matrixes	Seed		Anti-seed		L1, L2, and remaining	
	Adj: any AI	**Adj: not AI**	**Adj: any AI**	**Adj: not AI**	**Adj: any AI**	**Adj: not AI**
A/B: any AI	199	7	16	35	40	75
A/B: not AI	14	17	2	200	9	328
			Metrics			
# documents	237		253		452	
Precision	0.9660		0.9901		0.3478	
Recall	0.9343		0.8511		0.8163	
Accuracy	0.9114		0.8538		0.8142	
F1 score	0.9499		0.9153		0.4878	

Note: Analysis compares resulting "any AI" annotation of patent examiner A (first examiner reviewing a document) compared to patent examiner B (second examiner reviewing the same document) plus adjudication. The confusion matrix reflects the results of examiner A and B (rows) and the results of adjudication (columns, where "Adj" is shorthand for adjudicator in the column heading). If there is no disagreement between A and B, we assume the adjudicator would agree with A and B. If there is a disagreement, the adjudicator would agree with one of A or B and disagree with the other of B or A. Hence, the total number in each confusion matrix differs for that of Table 4 without adjudication. For the anti-seed set, the metrics are calculated such that "not AI" is the positive result. See discussion for metric definitions.

The tables in include the number of documents or observations included in the confusion tables and metrics. The number of documents is greater using the second method since adjudication results in two observations for each disagreement (as discussed above). "Precision" is the number of true positives divided by the number of predicted positives. For the first method, it does not matter if reviewing patent examiner "A" or "B" is assume to be "true." For the second method, the result of adjudication is assumed to be true. For both, "positive" is AI for the seed and L1, L2, and remaining sets and "not AI" for the anti-seed set. "Recall" is the number of true positives divided by the number of actual positives. "Accuracy" is the number of true positives and true negatives divided by the total number of documents or observations. The "F1 score" is a combination of precision and recall metrics using the harmonic mean.

This analysis gives us a baseline of how well experienced human raters may classify patent documents as AI or not using our AI technology component definitions (see discussion in "Inventing AI") and instructions. The results for the L1, L2, and remaining sets are the most meaningful to compare with our model results and with other studies (discussed in the sections below) since they are documents that are not used in training our models. In general, the results indicate that identifying AI in patent documents is not easy, even for experts in the field. For the

L1, L2 and remaining sets, the examiners identified 82 percent of the AI documents as AI, but out of all the documents they labeled AI, only 35 percent were correctly labeled.

Comparison of examiner annotation to seed and anti-seed set generation

The patent examiner annotations of the seed and anti-seed sets allows us to assess the process we used to generate those training data sets, i.e., performing a narrow search to identify seed set documents and using L1 and L2 expansions to generate the anti-seed set per Abood and Feltenberger (2019). If we assume all patent documents in the consolidated seed set are AI and all the patent documents in the anti-seed set are not AI—which is consistent with how they would be used in the classifications models[46]—we can compare each document to how they were annotated by the patent examiners (to include adjudication). This analysis assumes the patent examiner annotations is "truth" and is presented in Table 6 below.

The results indicate that the seed and anti-seed generation process is very good—accuracy is 92%. As previously discussed, disagreements between patent examiners exist: considering adjudication, examiner accuracy is 91% for the seed set and 85% for the anti-seed set (see Table 5, bottom half). The two results are similar. Thus we can conclude that the automated seed and anti-seed generation process we used produces results that are as good as a more labor-intensive process of human review.

The metrics in Table 6 are carried forward to "USPTO Model Seed/Anti-seed Generation" seed and anti-seed columns in Table A1 of "Inventing AI."[47]

[46] As previously discussed, the designation of "seed" and "anti-seed" is based on a consolidated group. This consolidation may result in a patent document identified as "seed" in the consolidated group (due to its use in the seed set of at least one model) that is also used as the anti-seed in another model. Since in our case only one annotated patent document in the manual validation random sample falls in this dual role (seed set in two models and anti-seed in one model), the impact on the confusion matrix analysis would be minimal.

[47] We do not carry forward the seed and anti-seed analysis from the examiner agreement annotations (Tables 4 and 5) since that analysis offers a different perspective, i.e., whether patent examiners agree as to whether the seed and anti-seed documents are AI or not. Likewise, we do not compare model predictions to patent examiner annotations for the seed and anti-seed sets; such a comparison is akin to metrics against the training set during model training, which provides a different perspective.

Table 6: Confusion matrix and metrics for seed and anti-seed generation

Confusion Matrixes	Seed set generation		Anti-seed set generation	
	Examiners: any AI	**Examiners: not AI**	**Examiners: any AI**	**Examiners: not AI**
AI (seed)	199	17	0	0
Not AI (anti-seed)	0	0	16	200
Metrics				
# documents	216		216	
precision	0.9213		0.9259	
recall	1.0000		1.0000	
accuracy	0.9213		0.9259	
f1 score	0.9590		0.9615	

Note: Analysis compares patent examiner annotation scoring, which is assumed to be "truth," to the assumption that seed and anti-seed documents are all AI and all not-AI, respectively. Patent examiner annotation includes adjudication to resolve differences. For the anti-seed set, the metrics are calculated such that "not AI" is the positive result. See discussion above for metric definitions.

Comparison of examiner annotation to model predictions

To assess our USPTO model predictions we compare the results of the patent examiner annotations, including adjudication, with the AI vs. not AI prediction results of our models, consolidated for "any AI". The analysis is restricted to the L1, L2, and remaining set of documents since these were not used to train the classification models.[48] This analysis is summarized in Table 7 below.

Comparing the results of Table 7 to the human rater classification (L1, L2, remaining set columns in Table 5 above, i.e., with adjudication) we see the model precision and accuracy is higher than human raters, but recall is much lower (0.3750 for the model vs. 0.8142). The recall score indicates the model predicts fewer true positives than patent examiners. The F1 score is also lower than human raters (0.3896 for the model vs. 0.4878), but it remains comparable when compared to other studies (see discussion below).

Table 7 is carried forward as the "USPTO Model" column in Figure A1 of "Inventing AI."

[48] The exception, as previously discussed, is one patent document what is included in L2 when it was used as an anti-seed for one model; the impact on our analysis should be minimal.

Table 7: Confusion matrix and metrics for model predictions

Confusion Matrix	Model predictions (L1/L2/remaining set)	
	Examiners: any AI	Examiners: not AI
Model: AI	15	22
Model: not AI	25	306
Metrics		
# documents	368	
precision	0.4054	
recall	0.3750	
accuracy	0.8723	
f1 score	0.3896	

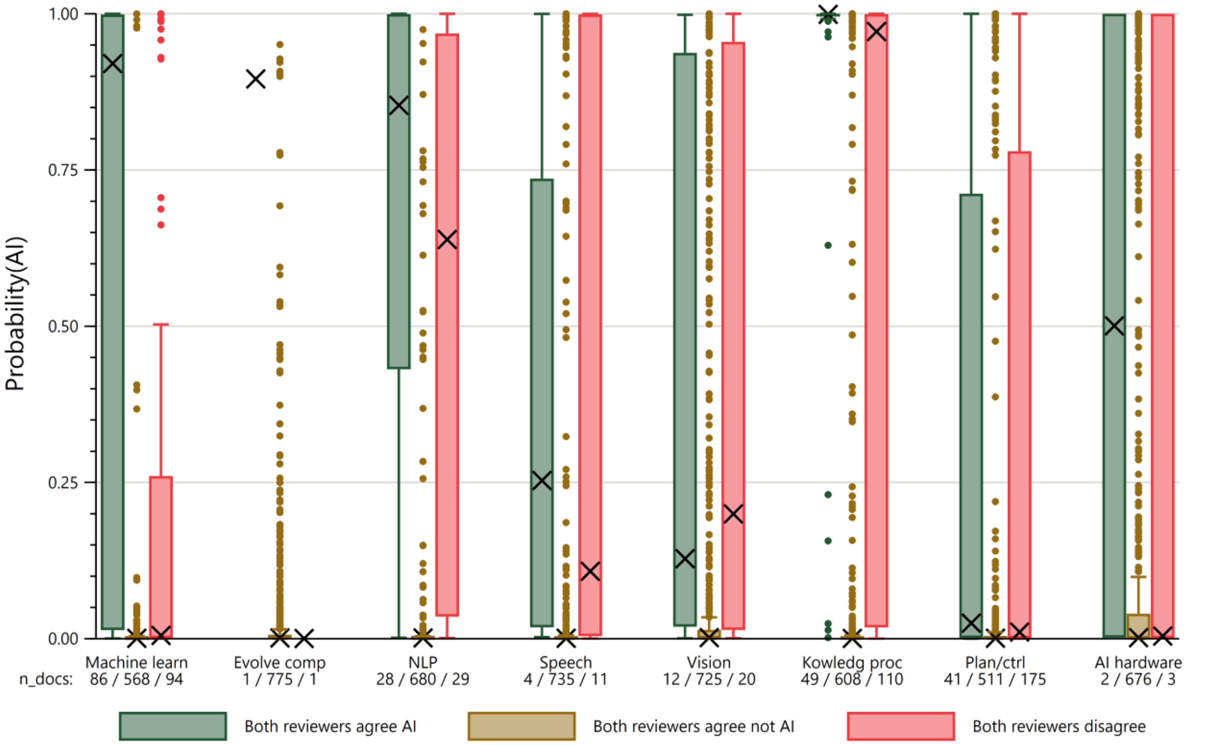
Note: Analysis compares patent examiner annotation scoring, which is assumed to be "truth," to USPTO model prediction results for "any AI." Patent examiner annotation includes adjudication to resolve differences. For the anti-seed set, the metrics are calculated such that "not AI" is the positive result. See discussion above for metric definitions.

We also use the manual validation to further analyze model results by comparing model predictions for patent documents in the consolidated L1, L2, and remaining set for each of the eight AI technology components to the results of patent examiner annotations. We can plot the distribution of model predictions, i.e., the probability of AI, in three broad groups: (1) model predictions of patent documents in which both reviewing patent examiners agree are AI, (2) model predictions of patent documents in which both reviewing patent examiners agree are not AI, and (3) model predictions of patent documents in which both reviewing patent examiners disagree.[49] We would expect model predictions for documents having AI agreement to have a high p(AI), model predictions for documents having not AI agreement to have a low p(AI), and model predictions for documents having disagreement to be somewhere in the middle. Unlike the confusion matrix analysis above, this analysis examines results from each of the eight AI component classification models.

Figure 8 illustrates the results using box plots for the prediction distributions. Since we did not randomly select the patent by individual AI technology components, the number of patent documents in each analysis varies, and in some instances is too small to draw conclusions.

[49] This analysis excludes adjudication since the third groups would be more difficult to define.

Figure 8: Distribution of model predictions by examiner annotation results



Notes: Analysis includes only patent documents in the consolidated L1, L2, and remaining set. The box plots illustrate the following:[50] *the left and right sides of the box indicate the 25th and 75th percentiles, respectively, and the median is a vertical line extending within the box; the whiskers indicate adjacent values; and solid circles outside the whiskers, if any, are outside values. An "X" marks the median of the p(AI) distribution. The number of documents in each AI technology component and distribution group is listed below the AI component label, ordered as per the legend (i.e., both reviewers agree AI in the left number, both reviewers agree not AI in the middle number, and reviewers disagree in the right number). The analysis does not include the "any AI" category since calculating p(AI) is not straight-forward.*

We immediately see that the model predictions are close to zero for the patent documents which both reviewing patent examiners agree are not AI—the box plots are barely visible, and the median is very close to 0.00 (brown box plot). For the patent documents that both reviewing patent examiners agree are AI the results are mixed: the prediction distribution for machine learning is high (median above 0.90), and also high for knowledge processing and natural language processing (NLP) (although the number of documents is smaller). For planning/control, however, the median is close to zero but the distribution extend to a 75th percentile of around 0.70. For the remaining components in the "both reviewers agree AI" the number of documents is very few to draw strong conclusions. Regarding the documents in the last group—both reviewing patent examiners disagree—the distribution of model predictions shows a lot of variance (i.e., large boxes) and the medians are close to zero, indicating the

[50] See Stata Corporation. "Graph box." Stata 13 online manual. https://www.stata.com/manuals13/g-2graphbox.pdf. Discussion of adjacent values may be found at Cox (2004).

models favor predicting not AI over AI. The notable exceptions are for NLP, which has a median close to the middle (which would be expected for uncertain results), and knowledge processing, which has median close to 1.0 (which would favor predicting AI over not AI).

These results indicate the planning/control classification model may have a high false negative error rate (since most of the AI documents agreed by both reviewing patent examiners is predicted as being not AI) and that the knowledge processing classification model may overly favor predicting AI. However, since reviewing patent examiner disagreements far outnumber agreements for AI in these two components, these results may indicate confusion an interpretation uncertainty between the reviewers.

Comparison to other studies

We also compare our model results to other studies by recreating the results from Cockburn et al. (2019) and from WIPO (2019), in addition to a naive case where all patent documents are presumed to be not AI.

To recreate Cockburn, we similarly query USPC class 901 and 706 combined with a patent title keyword query of patent titles (see Appendix II for query details) using the USPTO EAST patent search tool. Cockburn limited the study to patents between 1990 and 2014, inclusive. For the classification query using Cockburn's methodology we get 8,871 patents versus Cockburn's 8,640. Combining this result with the keyword title query and de-duplicating results yields 15,004 patents versus Cockburn's 13,615. Since our analysis uses a longer time period and includes PGPubs, we expand the queries to remove the time constraint and include both U.S. Patent and U.S. PGPub EAST databases. The result is a total of 52,442 patent documents (following de-duplication). Merging this result with the 800 randomly selected patent documents in our manual validation results in 57 patent documents that are "AI." The remaining 743 documents (out of the 800 in our random sample) are set to "not AI" since they do not come up using Cockburn's methodology.

WIPO uses a more complex combination of patent classifications and keywords. [51] We replicate these queries using EAST, excluding queries to Japanese patent applications since our analysis is limited to U.S. patent documents. We also include U.S. PGPubs (see Appendix II). The result is 294,470 patent documents. Merging this result with the 800 randomly selected patent documents in our manual validation results in 143 patent documents that are "AI": 136 from the seed set, 1 from the anti-seed set, and 6 from the combined L1, L2, and remaining set. The remaining 657 documents are treated as being "not AI."

[51] WIPO (2019), *Data collection method and clustering* background paper.

We also create a naive case in which all 800 randomly selected patent documents are set to be "not AI." This case replicates what would happen if our models were to default to predicting the negative class, which may happen in poorly trained models.

Table 8: Confusion matrixes and metrics for other studies

Confusion Matrixes	Cockburn (recreated)		WIPO (recreated)		Naive (all not AI)	
	Examiners: any AI	Examines: not AI	Examiners: any AI	Examines: not AI	Examiners: any AI	Examines: not AI
Study: any AI	0	0	4	2	0	0
Study: not AI	40	328	36	326	40	328
Metrics						
# documents	368		368		368	
Precision	0.0000		0.6667		0.0000	
Recall	0.0000		0.1000		0.0000	
Accuracy	0.8913		0.8967		0.8913	
F1 score	0.0000		0.1739		0.0000	

Note: Analysis compares patent examiner annotation scoring, which is assumed to be "truth," to other studies. Only patent documents corresponding to the consolidated L1, L2, and remaining set reviewed by the patent examiners, with adjudication, are included. Cockburn and WIPO results are recreated; naive results are based on the assumption that all patent document are predicted as being "not AI." See discussion for metric definitions.

The resulting confusion matrixes and metrics are presented in Table 8. To compare these other studies with our human annotations and model results discussed above, the table is limited to those patent documents in the L1, L2, and remaining set. Since our recreation of Cockburn does not result in any of the 800 randomly selected documents being in this set, precision, recall, and the F1 score for Cockburn equal zero. We note the accuracy of these other studies is about the same as the accuracy of our classification model. However, our recall and F1 score is larger, indicating the ability of the machine learning approach we used to find a broader set of AI patents as compared to the query-based approaches of Cockburn and WIPO.

Table 8 is carried forward as the last three columns in Table A1 of "Inventing AI."

FINDINGS: METHODOLOGY AND RESULTS

We now discuss the methodology we use to analyze the results of the AI classification models, looking at trends in AI patenting, the diffusion of AI patents to different technology classifications, the growth of AI inventor-patentees and owners-at-grant, the top 30 U.S. owners-at-grant, and the diffusion of AI patents by U.S. inventor-patentee location.

Volume and share of AI

Our first analysis of our post-prediction AI patent landscape involves graphing the volume (number) and share (percentage) of AI patent document over time. Since we de-duplicate patent documents (see Data Construction discussion above), each document represents a patent application—either granted as a patent (wherein any associated PGPub has been removed in the de-duplication process), published as a PGPub and abandoned, or published as a PGPub and still under evaluation. The number of AI patent applications is then a count by year, and the percentage is the number of AI patent applications divided by all patent applications by year. The methodology is the same for AI as a whole and for each AI component technology.

The remaining question, however, is what date to use for time: the publication date or the patent application filing date?

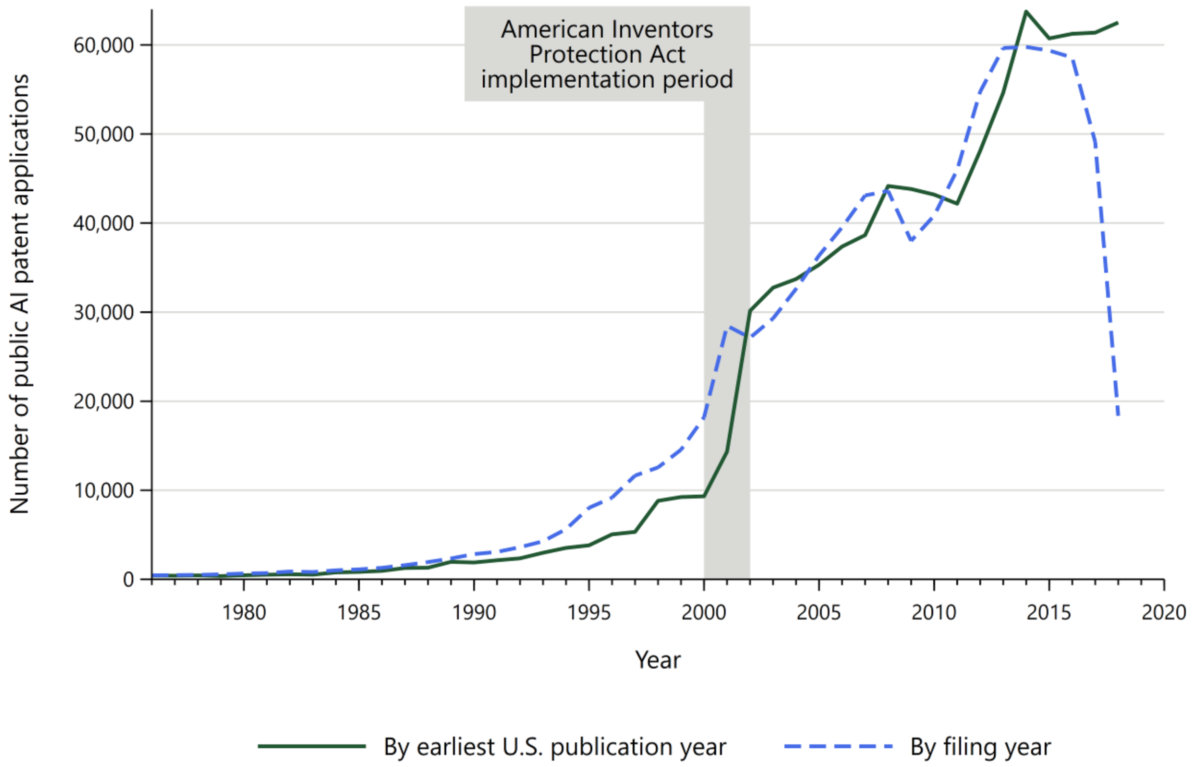
We choose the publication date, and more specifically the "earliest U.S. publication date",[52] since this is the date on which the patent application became known to the public. Use of the earliest U.S. publication date also precludes censoring problems with the use of the patent application filing date. Since a PGPub is published 18 months after its effective filing date,[53] our analysis would have to be truncated for recent years.

Figure 9 below illustrates the effect of data censoring caused by the lag between the filing of a patent application and its publication as a PGPub. The collapse in recent years for AI applications by filing year occurs since many of these applications were not yet published by the end of our data sample. Using the earliest U.S. publication year avoids this problem.

[52] The earliest U.S. publication date incorporates the PGPub publication date, if any, for a granted patent; see discussion in the Background section.

[53] MPEP § 1120.I. The "effective filing date" considers the benefit of an earlier filed patent application.

Figure 9: Number of patent applications by earliest U.S. publication vs. filing year, 1976-2018



Volume by AI technology component

We also analyze the volume of AI patents by each of the AI technology components (Figure 2 of "Inventing AI"). Planning/control and knowledge processing has the largest number of AI patents. As discussed in "Inventing AI," these two components are broad. Since our analysis allows a patent document to be classified in more than one AI technology component, the results may reflect overlapping component predictions, as presented in Table 9.

Table 9: Overlap of AI technology component predictions

Percent AI patents classified in component X (row) and component Y (column)

(X ∩ Y) / (X ∪ Y)

X (row) Y (column)	Machine learning	Evolutionary computation	Natural language processing	Speech	Vision	Knowledge processing	Planning/ control	AI hardware
Machine learning		13.58%	9.93%	8.06%	15.64%	17.62%	13.53%	18.07%
Evolutionary computation	13.58%		5.14%	4.10%	5.62%	5.20%	4.38%	6.10%
Natural language processing	9.93%	5.14%		27.70%	12.42%	14.09%	13.67%	18.05%
Speech	8.06%	4.10%	27.70%		9.76%	6.55%	6.08%	9.50%
Vision	15.64%	5.62%	12.42%	9.76%		16.84%	12.25%	15.30%
Knowledge processing	17.62%	5.20%	14.09%	6.55%	16.84%		53.50%	29.88%
Planning/ control	13.53%	4.38%	13.67%	6.08%	12.25%	53.50%		25.72%
AI hardware	18.07%	6.10%	18.05%	9.50%	15.30%	29.88%	25.72%	

Percent AI patents classified in component X (row) also classified in component Y (column)

(X ∩ Y) / (X)

# AI patents X	105,334	27,577	91,475	51,405	209,489	491,487	543,892	268,845
X (row) Y (column)	Machine learning	Evolutionary computation	Natural language processing	Speech	Vision	Knowledge processing	Planning/ control	AI hardware
Machine learning		57.63%	19.43%	22.74%	20.32%	18.19%	14.23%	21.30%
Evolutionary computation	15.09%		6.36%	6.06%	6.03%	5.22%	4.41%	6.34%
Natural language processing	16.88%	21.09%		60.29%	15.87%	14.65%	14.05%	20.49%
Speech	11.10%	11.29%	33.88%		11.07%	6.79%	6.27%	10.33%
Vision	40.42%	45.78%	36.34%	45.12%		20.56%	15.12%	23.61%
Knowledge processing	84.88%	92.98%	78.73%	64.95%	48.22%		66.35%	65.06%
Planning/ control	73.47%	87.04%	83.55%	66.34%	39.26%	73.43%		61.84%
AI hardware	54.37%	61.79%	60.23%	54.04%	30.30%	35.59%	30.57%	

There are two ways overlap may be calculated. In the first, represented by the top half of Table 9, the percent overlap is the number of AI patents classified in both AI components divided by the total number of AI patents in either component, i.e., the intersection of component X and component Y divided by the union of component X and component Y. This calculation is symmetric in that it does not matter which component is designated as "X" and which as "Y." As seen in the top half of Table 9, planning/control and knowledge processing have slightly over 50% of common.

In the second way to calculate overlap, represented by the bottom half of Table 9, the number of AI patents classified in both AI components is divided by the number of AI patents in the first component, i.e., the intersection of component X and component Y divided by the number of AI patents in component X (and hence is not symmetric). This calculation is interpreted as the percentage of AI patents in component X that are also classified in component Y. As seen in the bottom half of Table 9, 66% of planning/control is also classified in knowledge processing, and 73% of knowledge processing is also classified in planning/control. We also see that overlap with planning/control and knowledge processing make up a significant percentage of other AI technologies components. For example, 73% of machine learning is also classified in planning/control, and 85% also classified in knowledge presentation.

The bottom half of Table 9 also gives us insight into the interrelationships between the AI component technologies. AI hardware plays a significant part in machine learning (54% of machine learning is also in AI hardware), and a large part of machine learning is geared toward vision (40% of machine learning is also in vision). Not surprising, speech and natural language processing share a large percentage of common patents. Speech and natural language processing also share a large percentage of common patents with vision—a perhaps counterintuitive result that may be explained by natural language processing using the successful deep learning techniques of vision.[54]

Diffusion of AI across technologies

To show how AI may be diffusing across technologies, we use the technology category in which a patent is classified under the Cooperative Patent Classification (CPC) system. CPC is a classification system jointly established by the UPSTO and European Patent Office (EPO) and based on the International Patent Classification (IPC) system. The UPSTO began using CPC in 2013;[55] older U.S. patents were reclassified under CPC.

Specifically, we use the CPC subclass. For the purpose of looking into technology diffusion, the subclass provides a sufficient level of detail without being too specific. For example,[56] class "B06" pertains for *vehicles in general*, while its subclasses provide greater detail, e.g., subclass "B60B" for *vehicle wheels* and "B06G" for *vehicle suspension arrangements*. The next level down, such as main group "B60B 1/00" for *spoked wheels* and "B60G 13/00" for *resilient suspensions characterized by arrangement, location or type of vibration dampers*, are too detailed for our

[54] Krohn, Beyleveld, and Bassens (2020), 25.

[55] See USPTO, "Patent Classification" webpage, https://www.uspto.gov/patents-application-process/patent-search/classification-standards-and-development; see also MPEP § 905.

[56] See CPC Scheme, https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html

purpose. We also use the CPC First of a patent's classification. CPC First is "the inventive classification symbol which most adequately represents the invention as a whole for the patent family."[57]

We limit the analysis to granted patents for two reasons. First, the primary classification of a patent application may change from its publication as a PGPub to its publication as a granted patent given the judgement of the patent examiner. Second, our analysis of inventor-patentee and owner-at-grant diffusion is limited to granted patents due to our use of PatentsView, and thus use of patents for technology diffusion provides a consistent perspective.

Technology diffusion is measured by dividing the number of CPC subclasses of AI patents by the total number of AI subclasses of all patents for a given year. The denominator thus varies year-to-year; it averages 609.2 subclasses per year from 1976-2018, with a standard deviation of 4.3 subclasses. For the numerator, we chose to count a CPC subclass as having AI patents if there is more than one AI patent (i.e., at least two AI patents) in that CPC subclass in that year. See the Robustness Analysis section for additional discussion regarding this threshold, as well as discussion regarding our CPC data construction.

As discussed above, there is significant overlap in the AI technology components. Overlap also occurs in our AI component technology subclasses. Table 10 summarizes this overlap using the same methodology as before. A large percentage of CPC subclasses having knowledge processing also contains CPC subclasses having planning/control and vice versa (as seen in both the top and bottom halves of the Table 10). As discussed in "Inventing AI," the interdependence of AI technology components helps explain the distinct clusters in Figure 3 of "Inventing AI." This interdependence is evident in the overlaps in Table 10.

[57] MPEP § 905.03(a)III.A.(c).

Table 10: Overlap of AI technology component by technology subclasses

<table>
<tr><th colspan="9">Percent subclasses having AI patents classified in component X (row) and component Y (column)
(X ∩ Y) / (X U Y)</th></tr>
<tr><th>X (row)
Y (column)</th><th>Machine learning</th><th>Evolutionary computation</th><th>Natural language processing</th><th>Speech</th><th>Vision</th><th>Knowledge processing</th><th>Planning/ control</th><th>AI hardware</th></tr>
<tr><td>Machine learning</td><td></td><td>48.69%</td><td>26.36%</td><td>28.70%</td><td>54.03%</td><td>70.69%</td><td>67.36%</td><td>62.74%</td></tr>
<tr><td>Evolutionary computation</td><td>48.69%</td><td></td><td>18.54%</td><td>20.40%</td><td>39.47%</td><td>48.93%</td><td>48.17%</td><td>45.27%</td></tr>
<tr><td>Natural language processing</td><td>26.36%</td><td>18.54%</td><td></td><td>47.65%</td><td>30.43%</td><td>33.45%</td><td>32.46%</td><td>35.80%</td></tr>
<tr><td>Speech</td><td>28.70%</td><td>20.40%</td><td>47.65%</td><td></td><td>33.41%</td><td>30.23%</td><td>29.32%</td><td>35.77%</td></tr>
<tr><td>Vision</td><td>54.03%</td><td>39.47%</td><td>30.43%</td><td>33.41%</td><td></td><td>58.87%</td><td>53.73%</td><td>54.72%</td></tr>
<tr><td>Knowledge processing</td><td>70.69%</td><td>48.93%</td><td>33.45%</td><td>30.23%</td><td>58.87%</td><td></td><td>88.49%</td><td>61.18%</td></tr>
<tr><td>Planning/ control</td><td>67.36%</td><td>48.17%</td><td>32.46%</td><td>29.32%</td><td>53.73%</td><td>88.49%</td><td></td><td>57.42%</td></tr>
<tr><td>AI hardware</td><td>62.74%</td><td>45.27%</td><td>35.80%</td><td>35.77%</td><td>54.72%</td><td>61.18%</td><td>57.42%</td><td></td></tr>
<tr><th colspan="9">Percent subclasses having AI patents classified in component X (row) also in component Y (column)
(X ∩ Y) / (X)</th></tr>
<tr><td># CPC in X</td><td>443</td><td>311</td><td>233</td><td>225</td><td>429</td><td>559</td><td>572</td><td>406</td></tr>
<tr><th>X (row)
Y (column)</th><th>Machine learning</th><th>Evolutionary computation</th><th>Natural language processing</th><th>Speech</th><th>Vision</th><th>Knowledge processing</th><th>Planning/ control</th><th>AI hardware</th></tr>
<tr><td>Machine learning</td><td></td><td>71.70%</td><td>51.93%</td><td>57.78%</td><td>62.47%</td><td>71.20%</td><td>68.18%</td><td>73.40%</td></tr>
<tr><td>Evolutionary computation</td><td>50.34%</td><td></td><td>28.33%</td><td>31.56%</td><td>41.96%</td><td>49.19%</td><td>48.25%</td><td>48.28%</td></tr>
<tr><td>Natural language processing</td><td>27.31%</td><td>21.22%</td><td></td><td>58.67%</td><td>31.00%</td><td>33.45%</td><td>32.52%</td><td>36.95%</td></tr>
<tr><td>Speech</td><td>29.35%</td><td>22.83%</td><td>56.65%</td><td></td><td>34.03%</td><td>30.23%</td><td>29.37%</td><td>36.21%</td></tr>
<tr><td>Vision</td><td>60.50%</td><td>57.88%</td><td>57.08%</td><td>64.89%</td><td></td><td>59.39%</td><td>54.20%</td><td>62.81%</td></tr>
<tr><td>Knowledge processing</td><td>89.84%</td><td>88.42%</td><td>80.26%</td><td>75.11%</td><td>77.39%</td><td></td><td>90.03%</td><td>84.24%</td></tr>
<tr><td>Planning/ control</td><td>88.04%</td><td>88.75%</td><td>79.83%</td><td>74.67%</td><td>72.26%</td><td>92.13%</td><td></td><td>81.03%</td></tr>
<tr><td>AI hardware</td><td>67.27%</td><td>63.02%</td><td>64.38%</td><td>65.33%</td><td>59.44%</td><td>61.18%</td><td>57.52%</td><td></td></tr>
</table>

Note: Technology subclasses based on having more than one AI technology component patent in a CPC subclass in a given year.

Diffusion of AI across U.S. inventor-patentees and patent owners

We analyze the diffusion of AI across inventor-patentees and patent owners by calculating the percent of U.S. inventor-patentees and of U.S. owners-at-grant having U.S. AI patents. We take advantage of PatentsView data[58] that contains disambiguated IDs for each as well as disambiguated locations. The disambiguation enables us to group together the same inventor-patentee and separately the same owner-at-grant despite differences in the raw names on the face of the patent document.

As noted in "Inventing AI" (footnote 21), for owners-at-grant we use both organizations and individuals as identified as the assignee on the patent at grant. We do not have comprehensive data on the reassignment of patents following grant. Additionally, if a patent was not reassigned prior to grant, then the inventor or non-inventor applicant, as applicable, is the patent owner. We do not include these inventor owners and non-inventor applicant owners in our analysis since the PatentsView disambiguation algorithm does not extend between inventors, non-inventor applicants, and assignees.

Our analysis calculates the percentage of U.S. unique inventor-patentees and U.S. unique owners-at-grant [59] having AI patents in each year, where unique entities are identified by their disambiguated ID and "year" is the patent grant calendar year. The algorithm divides the number of unique AI inventor-patentees in each year by the total number of inventor-patentees in that year. Thus, if a U.S. inventor-patentee has at least one AI patent in a year, then they are a U.S. "AI inventor-patentee" in that year. Similarly for U.S. owners-at-grant. A patent having multiple inventor-patentees and/or multiple owners-at-grant is attributed to each.

One of the conclusions we draw in "Inventing AI" is that AI diffusion is that "more and more inventor-patentees within organizations are adopting AI in their work."[60] This conclusion is supported by looking at changes in the average percentage of AI inventor-patentees by unique owners-at-grant, as shown in Figure 10. We calculate this average percentage by using only patents having one owner-at-grant.[61] For each owner and year, we calculate the percentage of

[58] See www.patentsview.org

[59] To identify U.S. inventors and assignees, we use the country code in PatentsView. U.S. territories may be identified either by a U.S. state code or by a unique country code. We exclude all patents that do not have "US" as their country code.

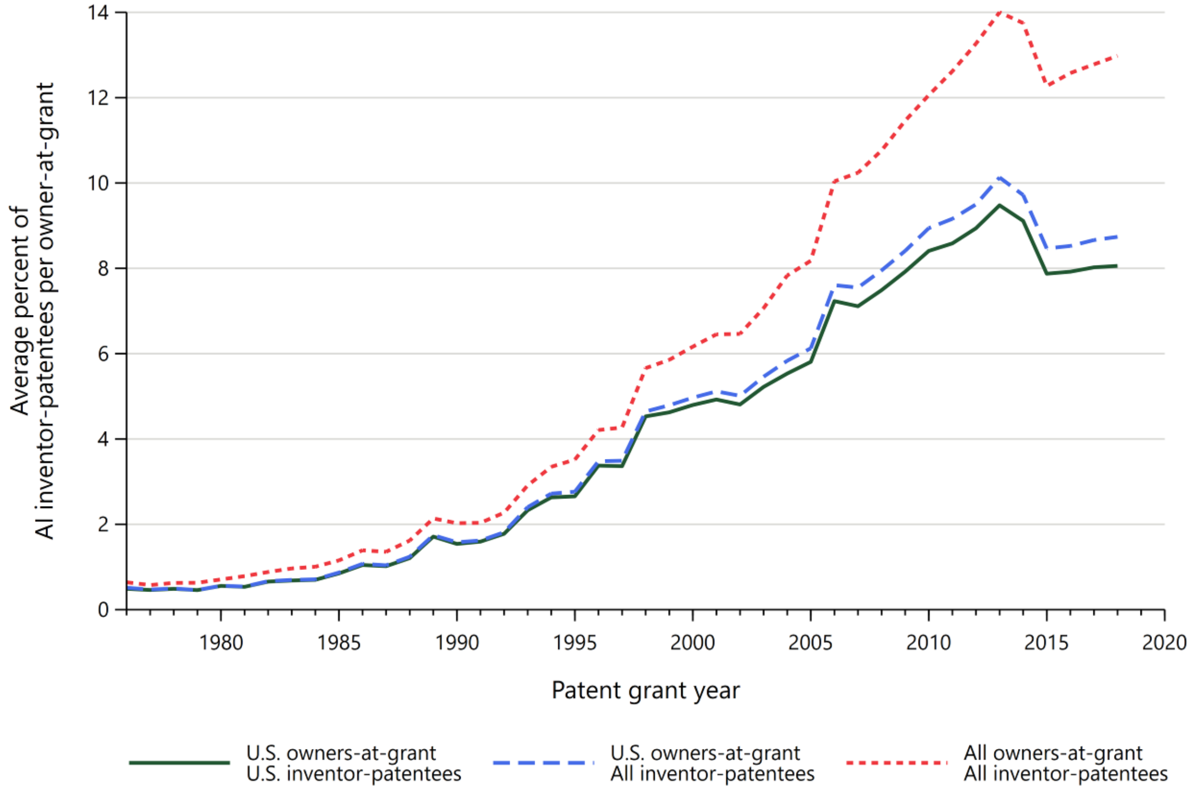
[60] Toole et al. (2020), 9.

[61] We keep only patents having one assignee since we do not have data that links specific inventors to specific assignees.

AI inventor-patentees out of all inventor-patentees at that owner, and then average across all owners for each year.

Regardless of whether we consider U.S. inventor-patentees and U.S. owners-at-grant only (green solid line) or consider broader categories (other lines), the trend of the average percentage increases over time as seen in Figure 10. This result indicates year-after-year there are more AI inventor-patentees compared to all inventor-patentees in an average owner organization. Hence, AI is diffusing among inventor-patentees within organizations.

Figure 10: Average percentage of AI inventor-patentees by unique owners-at-grant, 1976-2018



Notes: Excludes patents having more than one owner-at-grant and those missing owners-at-grant. "AI inventor-patentee" refers to an inventor-patentee having at least one AI patent for a unique owner-at-grant in a given year.

Top U.S. AI patent owners-at-grant

We determine the top 30 U.S. owners-at-grant by ranking the total number of patents assigned to each unique U.S. patent owner-at-grant between 1976 and 2018, inclusive. A number of companies, while unique business entities,[62] are nevertheless similar enough that we consider

[62] Either due to their being different legal entities or shortcoming in the PatentsView disambiguation algorithm.

them for the purpose of our analysis to be a single company. For example, we combined "Microsoft Corp.," which has 16,611 AI patents in our data, with "Microsoft Technology Licensing LLC," which has 5,444 AI patents (plus additional minor variants of these company names). We perform this consolidation by performing a simple keyword search of the top 30 assignees, manually editing the results, and adding the companies.

We do not combine companies based on mergers, acquisitions, divestitures, etc. For example, Sun Microsystems Inc. (#14) was acquired by Oracle Corp. (#7) in 2010,[63] and Lucent Technologies Inc. was divested by AT&T Corp. in 1996, merged with Alcatel SA of France in 2006 to form Alcatel-Lucent, and became part of Nokia (Finland) in 2016.[64] Such modifications would not reflect our approach of using the "assignee-at-grant."

We also make administrative changes to company names for simplicity. For example, "Hewlett-Packard Development Company, L.P." is change to "Hewlett-Packard Co." and "AT&T Intellectual Property I, L.P." is changed to "AT&T Corp."

Diffusion of AI across geography

We capture the diffusion of AI patents across geography by examining the number of patents based on inventor-patentee location by U.S. county in two time periods: from 1976 through 2000 (25 years, inclusive) and from 2001 through 2018 (18 years, inclusive), chosen as a convenient millennial division. We produce the county maps based on the FIPS codes in our data using the Stata "maptile" program[65] with the 2014 U.S. county map geography template.[66] Use of the 2014 template for both maps enables direct comparisons between the two time periods. We do not attempt to correct for any changes in U.S. counties before or after 2014.

For the purpose of analysis, we assume a total of 3,142 counties, which is the number of county FIPS codes in the U.S. Department of Agriculture's Natural Resources Conservation Services website, excluding U.S. territories (but including the District of Columbia).[67] Our data has only 2,025 codes; hence the remainder are included as the "none or no data" category.

[63] Wikipedia, "Sun Microsystems", https://en.wikipedia.org/wiki/Sun_Microsystems

[64] Wikipedia, "Lucent"; https://en.wikipedia.org/wiki/Lucent; and Wikipedia, "Nokia"; https://en.wikipedia.org/wiki/Nokia

[65] See Stepner, https://michaelstepner.com/maptile/

[66] The 2014 U.S. county template is the most recent county template; see https://michaelstepner.com/maptile/geographies/

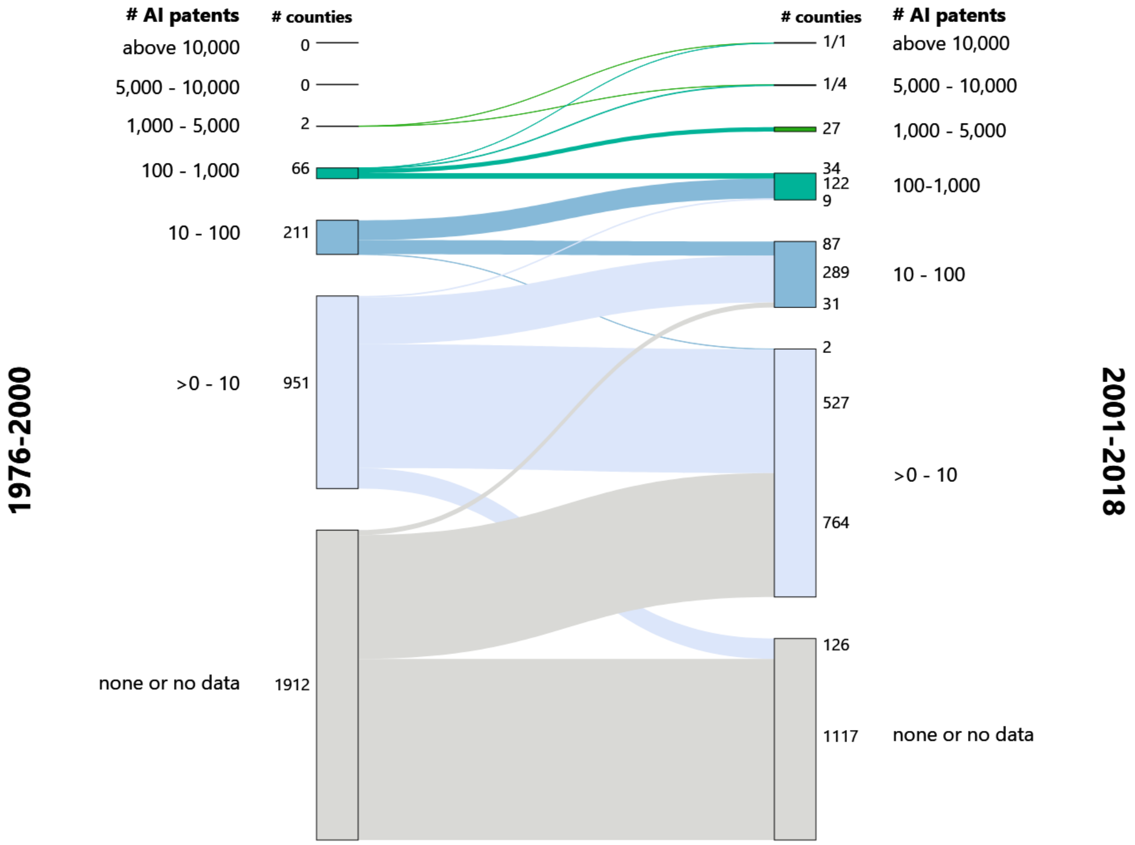
[67] See https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697

We calculate the number of patents in a county is using fractional patent counts for patents having multiple inventors—the county location of each inventor is counted as the fraction of inventors for that patent in that county (e.g., a patent with three inventors in two counties would be counted as 2/3 of a patent for one county and 1/3 for the second county). We consider locations outside the U.S. when determining the fractions.

We explicitly select the number of bins and breakpoints for the maps so as to produce visually meaningful maps. Since the data is skewed to the left, bins containing an equal number of patents would have resulted in breakpoints having a small range of patent numbers plus one bin containing a large range.

Figures 7a and 7b in "Inventing AI" illustrate the number of inventor patents by county in two time periods maps. The Sankey diagram in Figure 11 below illustrates the transition in the number of counties between the various number-of-patent bins used in the maps and provides additional insight into the changes from 1976-2000 to 2011-2018.

Figure 11: Transition of U.S. counties by each AI patent bin from 1976-2000 to 2011-2018



Notes: The number of AI patents on each side of the Sankey diagram corresponds to the number of AI patent bins in the maps of Figures 7a and 7b in "Inventing AI," and the number of counties is the count of counties in each bin of the maps.

ROBUSTNESS ANALYSIS

Finally, we provide two robustness checks. The first illustrates the impact of different prediction thresholds on our AI trend analysis. The second examines different thresholds in the technology diffusion analysis.

AI prediction discrimination threshold

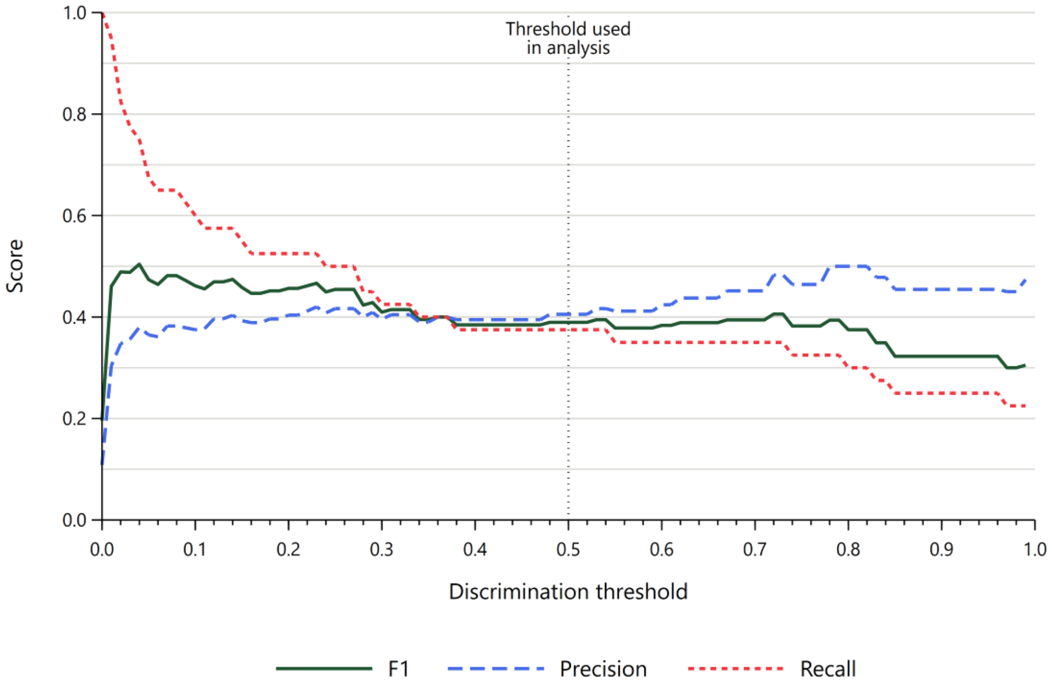
We use a 50% discrimination threshold for predicting AI from each of the eight AI component classification models. One method to determine the prediction threshold analytically is to create a validation or hold-out set from known classified data (such as from a gold standard forming the training data) and perform an "area under the receiver operating characteristics (ROC) curve" (AUC) analysis. The combined L1, L2, and remaining subset of the 800 manual validation documents annotated by experienced patent examiners may serve this function. However in our case, there is an insufficient number of documents in each AI technology component to assess each of the AI component technology classification models. In addition, we are unable to calculate a p(AI) for the consolidated "any AI" category, which would be needed for an AUC analysis, in a simple manner.[68]

What we are able to calculate, however, is a threshold plot comparing precision, recall, and the F1 score based on different "any AI" discrimination thresholds over our manual validation sample of patent documents.[69] Figure 12 illustrates this analysis. We see that as the threshold is increased, precision increases, recall decreases, and the F1 score (which is the geometric mean of precision and recall) is an upside-down U-shaped curve. Precision and recall intersect at about a 0.35 threshold, but the curves are relatively flat from between 0.3 through approximately 0.55. While this analysis does not nail down a specific discrimination threshold, it does allow us to conclude that 0.5 is reasonable.

[68] It is unlikely that the probability of a patent document being in one AI component is not independent of it being in another AI component (see discussion of AI component overlap); hence, a combined probability is not a simple calculation. Moreover, it may be argued that the probability of AI should not be dependent on a combination of individual AI component probabilities (e.g., does the probability of AI increase if one patent document receives marginal scores in several AI components?).

[69] See discussion at https://www.scikit-yb.org/en/latest/api/classifier/threshold.html

Figure 12: Discrimination threshold analysis

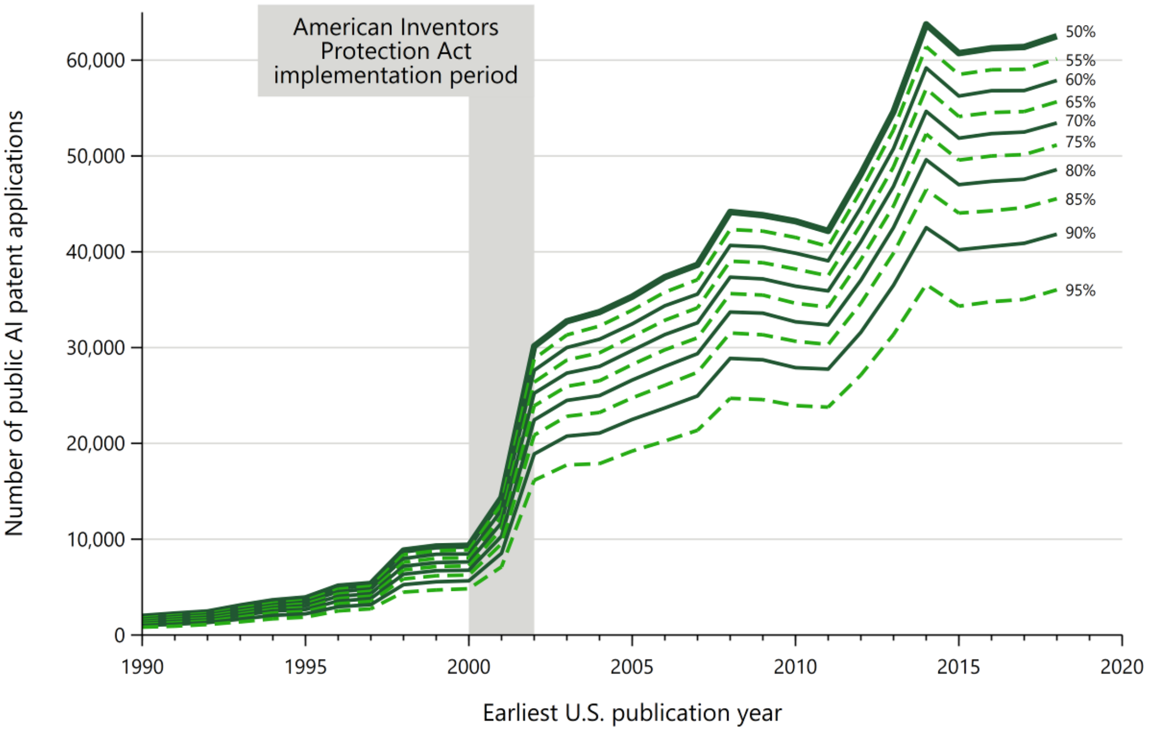


Note: Analysis based on patent examiner annotations, with adjudication, of the combined L1, L2, and remaining set and the highest p(AI) prediction of the AI component classification models for the patent documents in this set. Vertical dotted line identifies the 50% threshold used in the analysis.

We can also re-analyze the trends in public AI patent applications by using different prediction thresholds for "any AI." Figure 13 below illustrates how Figure 1 of "Inventing AI," the volume of public AI patent applications over time, would change. As we increase the threshold from 50% to 60% and more, the number of AI patent applications decreases, but the general shape of the curve remains the same. Between a threshold of 50-80%, the number of AI patent application in 2018 appears to drop approximately 2,500 applications for each 5% increase in the threshold, or about 4% of what we predicted using a 50% threshold.

Increasing the prediction threshold would result in a more conservative estimate of AI growth and diffusion, but simultaneously, potentially underestimate the nature and diffusion of AI in U.S. patents.

Figure 13: Predicted public AI patent applications, 1990-2018, by varying prediction threshold



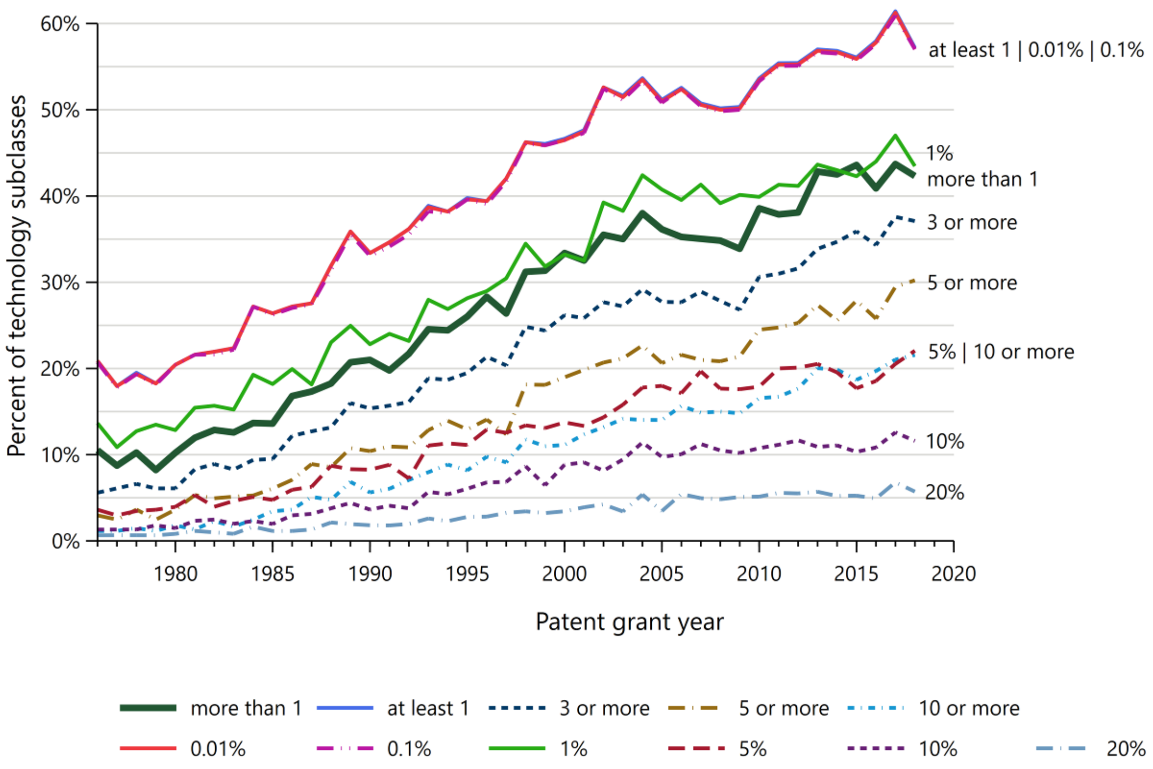
Notes: Percent labels identify the line for each AI prediction threshold, with the thick green line identifying the 50% prediction threshold used in the analysis. The graph begins at 1990 since the numbers are low before that year and the lines become indistinguishable from each other.

Technology dispersion threshold

The robustness check we perform is to vary the patent threshold for including a CPC technology subclass in the dispersion count. In “Inventing AI” we used a threshold of “more than one” AI patent. The threshold may be based on other numeric values, or it may be based on the percent of patents in that CPC subclass (e.g., we could count the CPC subclass as having AI in a given year if AI patents comprise 1% or more of the total number of patents in that CPC subclass in that year). Since our analysis is done on an annual basis, there is a distribution of numeric values for thresholds based on the percent of patents.

Figure 14, below, illustrates the impact of using different thresholds. Additionally, Table 11 summarizes the distribution of numeric values for percent thresholds.

Figure 14: Technology diffusion based on different CPC threshold values, 1976-2018



Notes: The thick green line identifies the "more than one" patent document threshold used in the analysis.

The smallest percent thresholds (0.01% and 0.1%) are essentially the same as having at least one AI patent in a CPC subclass (Figure 14 and Table 11). Our selected threshold of "more than one" AI patent in a CPC subclass is about the same as a 1% threshold. The difference between "at least one" and "more than one" can be considered to be random noise of single patents being classified in a subclass. As seen in Figure 14, this difference is on average 13.8% each year. Larger thresholds further reduce the percentage of CPC subclasses having AI patents. For example, using a 5% threshold is about the same as requiring 10 or more AI patents in a subclass and would reduce our measure of technological diffusion from 42.3% of CPC subclasses in 2018 to 21.7% in 2018.

There is no standard threshold level. Our selection of "more than one" (i.e., two or more) AI patents appears to balance between a more permissive threshold of "at least one" (i.e., any) AI patent in a subclass and the more restrictive threshold of 5% or "10 or more" AI patents.

Table 11. Distribution of patents for percent thresholds

Statistic	# patents in CPC subclass/yr	Number AI patents in CPC subclasses per year to meet threshold					
		0.01%	0.1%	1%	5%	10%	20%
mean	236.26	0.02	0.24	2.36	11.81	23.63	47.25
std dev	884.13	0.09	0.88	8.84	44.21	88.41	176.83
skewness	17.01	17.01	17.01	17.01	17.01	17.01	17.01
min	1.00	0.00	0.00	0.01	0.05	0.10	0.20
p1	1.00	0.00	0.00	0.01	0.05	0.10	0.20
p10	6.00	0.00	0.01	0.06	0.30	0.60	1.20
p25	19.00	0.00	0.02	0.19	0.95	1.90	3.80
p50	63.00	0.01	0.06	0.63	3.15	6.30	12.60
p75	186.00	0.02	0.19	1.86	9.30	18.60	37.20
p90	465.00	0.05	0.47	4.65	23.25	46.50	93.00
p99	2,749.00	0.27	2.75	27.49	137.45	274.90	549.80
max	31,277.00	3.13	31.28	312.77	1,563.85	3,127.70	6,255.40

Notes: Statistics are mean, standard deviation, skewness, minimum, percentiles (1%, 10%, 25%, 50% (median), 75%, 90%, and 99%), and maximum of all CPC subclasses; data from 1976-2018, inclusive.

APPENDIX I: QUERIES FOR SEED SET GENERATION

The table below presents the Clarivate Derwent™ database queries we use to generate the seed sets for each of the eight AI technology components. The queries include classification codes from the Cooperative Patent Classification (CPC) system (query code ".cpc."), the International Patent Classification (IPC) system (query codes ".ipcr,.cipg,cicl.cips."), and the U.S. Patent Classification (USPC) system (query codes ".clas.", ".ccls.", ".cor", or ".cas.") systems, in addition to the Clarivate Derwent World Patent Index™ (query codes ".EMCD,CMCD."). Results are limited to U.S. patent documents (query code ".pfpc.").

We note the CPC, IPC, and USPC classifications are as they existed in the Derwent™ database at the time the queries were run in the December 2018 timeframe. Additionally, wildcards are represented by "$" and "?" symbols.

Table A1: Queries for seed set generation (circa December 2018)

AI Component	Query	Discussion and Glossary
AI hardware	((((708/$ OR 712/$ OR 326/$ OR 257/$ OR 365/$ OR 711/$).COR. OR (G06N99/002 OR G06F9/$ OR G06T1/20 OR G06T1/60 OR H04N19/42$ OR H04N19/43$).cpc.) AND T01-J16$.EMCD,CMCD.) OR (G06N3/002 OR G06N3/02 OR G06N3/06$ OR G06N7/04$).cpc.) AND US.pfpc.	Intersection of processing circuitry, solid state, or memory in USPC with processing, memory, or video hardware in CPC with Derwent™ AI, which is then in union with bio-molecular computers, neural network hardware, or fuzzy logic hardware in CPC. <u>CPC and/or IPC</u> G06F9: Program control G06T1/20: Processor architectures G06T/60: Memory H04N19/42,43: Video hardware G06N3/002: Bio-molecular computers G06N3/06: Neural network hardware G06N7/04J: Fuzzy logic hardware <u>Derwent™</u> T01-J16: Artificial Intelligence <u>USPC</u> 712: Processing architectures 326: Circuitry 365: Solid-state 711: Memory

AI Component	Query	Discussion and Glossary
Evolutionary computation	(T01-J16C4$).EMCD,CMCD. AND 706/13.cor. AND US.pfpc. AND (G06N3/086 OR G06N3/12 OR G06N3/12?).cpc.	Intersection of Derwent™ genetic algorithms with genetic algorithms in USPC with genetic algorithms or genetic models in CPC. CPC and/or IPC G06N3/086 Genetic algorithms G06N3/12 Genetic models Derwent™ T01-J16C4: Genetic algorithms USPC 706/13: Genetic algorithm and genetic programming
Knowledge processing	(G06F17/3$ OR G06N5/$ OR G06F19/00 OR G06F19/24).cpc. AND (G06F17/3$ OR G06N5/$ OR G06F19/00 OR G06F19/24).ipcr,cipg,cid,cips. AND (T01-J16$).EMCD,CMCD. AND US.pfpc. AND (706/45 OR 706/46 OR 706/47 OR 706/48 OR 706/49 OR 706/5? OR 706/60 OR 706/61 OR 706/61).COR.	Intersection of information retrieval, adapted digital processing, machine learning, or knowledge-based models in CPC and in IPC with Derwent™ AI with knowledge processing in USPC. CPC and/or IPC G06F17/30: Information retrieval G06F19/00: Adapted Digital Processing G06F19/24: Machine Learning G06N5/$: Knowledge-Based Models Derwent™ T01-J16: Artificial Intelligence USPC 706/45-61: Knowledge Processing

AI Component	Query	Discussion and Glossary
Machine learning	(G06N99/005 OR G06N3/$ OR G06F15/18 OR G06F19/24 OR A61B5/7267 OR G06N7/005 G06N7/023).cpc. AND (T01-J16C1$ OR T01-J16C2$ T01-J16C4$ T01-J16C6$).EMCD,CMCD. AND US.pfpc. AND (706/12 OR 706/13 OR 706/14 OR 706/15 OR 706/16 OR 706/17 OR 706/18 OR 706/19 OR 706/2? OR 706/3? OR 706/40 OR 706/41 OR 706/42 OR 706/43 OR 706/44).cor.	Intersection of learning machines, biological computation, bioinformatics, training physiological classifiers, or neural networks in CPC with neural networks, genetic algorithms, or intelligent searching in Derwent™ with machine learning, adaptive systems, or neural networks in USPC. CPC and/or IPC G06N99/005: Learning machines G06N3: Computer systems based on biological models G06F19/24: Bioinformatics for machine learning, data mining, or biostatistics G06N7/005: Probabilistic networks G06N7/023: Parameters of a fuzzy system A61B5/7267: Classification of physiological signals involving training the classification device Derwent™ T01-J16C1: Neural networks T01-J16C2: Learning T01-J16C4: Genetic algorithms T01-J16C6: Intelligent searching USPC 706/12-44: Machine learning, adaptive system, neural networks

AI Component	**Query**	**Discussion and Glossary**
Natural language processing	(G06F17/2$ OR G06N99/005 OR G06F19/00 OR G06F19/24).cpc. AND (G06F17/2$ OR G06F19/00 OR G06F19/24).ipcr,cipg,cicl,cips. AND (T01-J16C3$ OR T01-J14$).EMCD,CMCD. AND T01-J16$.EMCD,CMCD. AND US.pfpc. AND (704/? OR 704/10 OR 706/45 OR 706/46 OR 706/47 OR 706/48 OR 706/49 OR 706/5? OR 706/60 OR 706/61).cor.	Intersection of natural language processing in CPC, IPC, Derwent™ and USPC. CPC and/or IPC G06F17/20: Handling natural language data G06F19/00: Adapted Digital Processing G06F19/24: Machine Learning G06N99/005: Learning Machines Derwent™ T01-J16C: Natural and pictorial language processing T01-J14: Language translation T01-J16: Artificial Intelligence USPC 704/1-10: Linguistics 706/45-61: Knowledge Processing
Planning/ control	(G06Q10/$ OR G05B13/$ OR G05B17/$ OR G06N3/006 OR G06N3/008).cpc. AND (G06Q10/$ OR G05B13/$ OR G05B17/$).ipcr,cipg,cicl,cips. AND (T01-J16$ OR T06-A05A$).EMCD,CMCD. AND US.pfpc.	Intersection of software/business applications or adaptive control systems in CPC and in IPC with AI or AI-based adaptive control in Derwent™ CPC and/or IPC G06Q10: Administration, management G05B13: Adaptive control systems G05B17: Models or Simulator G06N3/006: Artificial life based on virtual entities G06N3/008: Artificial life based on physical entities Derwent™ T01-J16: Artificial Intelligence T06-A05A: Artificial Intelligence-based adaptive control systems

AI Component	Query	Discussion and Glossary
Speech	(G10L15/$ OR G10L17/$ OR G10L21/$ OR G10L25/$ OR G10L13/$).cpc. AND (G10L15/$ OR G10L17/$ OR G10L21/$ OR G10L25/$ OR G10L13/$).ipcr,cipg,cicl,cips. AND (T01-C08A$ OR W04-V$).EMCD,CMCD. AND T01-J16$.EMCD,CMCD. AND US.pfpc. AND 704/2??.CCLS,COR.	Intersection of speech in CPC, IPC, Derwent™ and USPC. <u>CPC and/or IPC</u> G10L15: Speech Recognition G10L17: Speaker Identification G10L21: Processing of Speech or Voice Signal G10L25: Speech or voice analysis G10L13: Speech synthesis <u>Derwent™</u> T01-C08A: Speech recognition/synthesis W04-V: Analysis, synthesis, and processing of sound waves T01-J16: Artificial Intelligence <u>USPC</u> 704/200-278: Speech Signal Processing
Vision	(G06K9/$ OR G06T3/$ OR G06T5/$ OR G06T7/$).cpc. AND (G06K9/$ OR G06T3/$ OR G06T5/$ OR G06T7/$).ipcr,cipg,cicl,cips. AND (T01-J10B$ OR T04-D$).EMCD,CMCD. AND "382".clas. AND T01-J16$.EMCD,CMCD. AND US.pfpc.	Intersection of vision in CPC, IPC, Derwent™ and USPC. <u>CPC and IPC</u> G06K9: Recognition of characters or patterns G06T3: Image Transformation G06T5: Image enhancement/ restoration G06T7: Image Analysis <u>Derwent™</u> T01-J10B: Image Processing T04-D: Character and signal pattern recognition T01-J16: Artificial Intelligence <u>USPC</u> 382: Image Analysis

APPENDIX II: COMPARISON TO OTHER STUDIES

The queries to replicate the Cockburn et al. (2019) and WIPO (2019) studies are below. We make these queries using the USPTO EAST patent search tool and thus adapt the queries to use that tool. Additionally, we modify the queries for consistency with our analysis, i.e., time period, inclusion of U.S. PGPubs, and analysis limited to U.S. patent documents.

For the naive results we set all predictions equal to "not AI" for all patent documents and thus no query is needed.

Table A2: Queries to recreate Cockburn

Query type	EAST query
USPC classification	(@PY>="1990" AND @PY<="2014") AND (901/$.CIOR. OR 706/$.CIOR.)
Title keywords	(@PY>="1990" AND @PY<="2014") AND (OR "natural language processing" OR "image grammars" OR "pattern recognition" OR "image matching" OR "symbolic reasoning" OR "symbolic error analysis" OR "pattern analysis" OR "symbol processing" OR "physical symbol system" OR "natural languages" OR "pattern analysis" OR "image alignment" OR "optimal search" OR "symbolic reasoning" OR "symbolic error analysis" OR "machine learning" OR "neural networks" OR "reinforcement learning" OR "logic theorist" OR "bayesian belief networks" OR "unsupervised learning" OR "deep learning" OR "knowledge representation and reasoning" OR "crowdsourcing and human computation" OR "neuromorphic computing" OR "decision making" OR "machine intelligence" OR "neural network" OR "computer vision" OR "robot" OR "robots" OR "robot systems" OR "robotics" OR "robotic" OR "collaborative systems" OR "humanoid robotics" OR "sensor network" OR "sensor networks" OR "sensor data fusion" OR "systems and control theory" OR "layered control systems").ti.
Final query	Combined UPSC classification and title keywords without time limitations (i.e., remove (@PY>="1990" AND @PY<="2014")) using U.S. Patent and U.S. PGPub databases

Source: Cockburn, Henderson, and Stern (2019) as expanded by USPTO.

Table A3: Queries to replicate WIPO AI study

Query type	EAST query
Block 1	(Y10S706/$ OR G06N3/$ OR G06N5/003-027 OR G06N7/005-06 OR G06N99/005 OR G06T2207/20081 OR G06T2207/20084 OR G06T3/4046 OR G06T9/002 OR G06F17/16 OR G05B13/027 OR G05B13/0275 OR G05B13/028 OR G05B13/0285 OR G05B13/029 OR G05B13/0295 OR G05B2219/33002 OR G05D1/0088 OR G06K9/$ OR G10L15/$ OR G10L17/$ OR G06F17/27-2795 OR G06F17/28-289 OR G06F17/30029 OR G06F17/30035 OR G06F17/30247 OR G06F17/30262 OR G06F17/30401 OR G06F17/3043 OR G06F17/30522 OR G06F17/3053 OR G06F17/30654 OR G06F17/30663 OR G06F17/30666 OR G06F17/30669 OR G06F17/30672 OR G06F17/30684 OR G06F17/30687 OR G06F17/3069 OR G06F17/30702 OR G06F17/30705 OR G06F17/30713 OR G06F17/30731 OR G06F17/30737 OR G06F17/30743 OR G06F17/30746 OR G06F17/30784 OR G06F17/30814 OR G06F19/24 OR G06F19/707 OR G01R31/2846-2848 OR G01N2201/1296 OR G01N29/4481 OR G01N33/0034 OR G01R31/3651 OR G01S7/417 OR G06N3/004-008 OR G06F11/1476 OR G06F11/2257 OR G06F11/2263 OR G06F15/18 OR G06F2207/4824 OR G06K7/1482 OR G06N7/046 OR G11B20/10518 OR G10H2250/151 OR G10H2250/311 OR G10K2210/3024 OR H01J2237/30427 OR H01M8/04992 OR H02H1/0092 OR H02P21/0014 OR H02P23/0018 OR H03H2017/0208 OR H03H2222/04 OR H04L2012/5686 OR H04L2025/03464 OR H04L2025/03554 OR H04L25/0254 OR H04L25/03165 OR H04L41/16 OR H04L45/08 OR H04N21/4662-4666 OR H04Q2213/054 OR H04Q2213/13343 OR H04Q2213/343 OR H04R25/507 OR G08B29/186 OR B60G2600/1876 OR B60G2600/1878 OR B60G2600/1879 OR B64G2001/247 OR E21B2041/0028 OR B23K31/006 OR B29C2945/76979 OR B29C66/965 OR B25J9/161 OR A61B5/7264-7267 OR Y10S128/924 OR Y10S128/925 OR F02D41/1405 OR F03D7/046 OR F05B2270/707 OR F05B2270/709 OR F16H2061/0081 OR F16H2061/0084 OR B60W30/06 OR B60W30/10-12 OR B60W30/14-17 OR B62D15/0285 OR G06T2207/30248-30268 OR G06T2207/30236 OR G05D1/$ OR A61B5/7267 OR F05D2270/709 OR G06T2207/20084 OR G10K2210/3038 OR G10L25/30 OR H04N21/4666 OR A63F13/67 OR G06F17/2282).CPC.

Query type	**EAST query**
Block 2	(((ARTIFIC$ OR COMPUTATION$) ADJ INTELLIGEN$) OR (NEURAL ADJ NETWORK$4) OR ((NEURAL ADJ NETWORK$4) OR (NEURALNETWORK$)) OR (BAYES$ ADJ NETWORK$4) OR ((BAYESIAN ADJ NETWORK$4) OR (BAYESIANNETWORK$)) OR (CHATBOT$1) OR (DATA ADJ MINING$) OR (DECISION ADJ MODEL$1) OR (DEEP ADJ LEARNING$) OR ((DEEP ADJ LEARNING$) OR (DEEPLEARNING$)) OR (GENETIC ADJ ALGORITHM$1) OR ((INDUCTIVE ADJ LOGIC) NEAR1 PROGRAMM$) OR (MACHINE ADJ LEARNING$) OR ((MACHINE ADJ LEARNING$) OR (MACHINELEARNING$)) OR ((NATURAL NEAR1 LANGUAGE) ADJ (GENERATION OR PROCESSING)) OR (REINFORCEMENT ADJ LEARNING) OR (SUPERVISED ADJ (LEARNING$ OR TRAINING)) OR ((SUPERVISED ADJ LEARNING$) OR (SUPERVISEDLEARNING$)) OR (SWARM ADJ INTELLIGEN$) OR ((SWARM ADJ INTELLIGEN$) OR (SWARMINTELLIGEN$)) OR (UNSUPERVISED ADJ (LEARNING$ OR TRAINING)) OR ((UNSUPERVISED ADJ LEARNING$) OR (UNSUPERVISEDLEARNING$)) OR (SEMISUPERVISED ADJ (LEARNING$ OR TRAINING)) OR ((SEMI ADJ SUPERVISED ADJ LEARNING$) OR (SEMISUPERVISEDLEARNING$) OR (SEMISUPERVISED ADJ LEARNING$)) OR CONNECTIONIS? OR (EXPERT ADJ SYSTEM$1) OR (FUZZY ADJ LOGIC$1) OR ((TRANSFER ADJ LEARNING) OR (TRANSFERLEARNING)) OR (TRANSFER ADJ LEARNING) OR (LEARNING ADJ3 ALGORITHM$1) OR (LEARNING ADJ MODEL$1) OR (SUPPORT ADJ VECTOR ADJ MACHINE$1) OR (RANDOM ADJ FOREST$1) OR (DECISION ADJ TREE$1) OR (GRADIENT ADJ TREE ADJ BOOSTING) OR (XGBOOST) OR ADABOOST OR RANKBOOST OR (LOGISTIC ADJ REGRESSION) OR (STOCHASTIC ADJ GRADIENT DESCENT) OR (MULTILAYER ADJ PERCEPTRON$1) OR (LATENT ADJ SEMANTIC ADJ ANALYSIS) OR (LATENT ADJ DIRICHLET ADJ ALLOCATION) OR (MULTI-AGENT ADJ SYSTEM$1) OR (HIDDEN ADJ MARKOV ADJ MODEL$1)).ti,ab,clm.
Block 3 C1	(G06T7/$ OR G06T1/20 OR G10L13/$ OR G10L25/$ OR G10L99/$ OR G06F17/14-148 OR G06F17/153 OR G10H2250/005-021 OR G06F17/50 OR G06Q30/02-0284 OR G07C9/$ OR G06F21/$).CPC.

Query type	EAST query
Block 3 C2	(A61B5/$ OR A63F13/67 OR B23K31/$ OR B25J9/16 OR B25J9/18 OR B25J9/20 OR B29C65/$ OR B60W30/06 OR B60W30/10 OR B60W30/12 OR B60W30/14 OR B60W30/16 OR B60W30/165 OR B60W30/17 OR B62D15/02 OR B62D15/0295 OR B64G1/24 OR B64G1/26 OR B64G1/28 OR B64G1/32 OR B64G1/34 OR B64G1/36 OR B64G1/38 OR E21B41/$ OR F02D41/14 OR F02D41/16 OR F03D007/04 OR F03D7/048 OR F16H61/$ OR G01N29/44 OR G01N29/46 OR G01N29/48 OR G01N29/50 OR G01N29/52 OR G01N33/$ OR G01R31/28 OR G01R31/30$ OR G01R31/31$ OR G01R31/36$ OR G01S7/41$ OR G05B13/02 OR G05B13/04$ OR G05D1/$ OR G06F9/44$ OR G06F11/14$ OR G06F11/22$ OR G06F11/24$ OR G06F11/25$ OR G06F11/26$ OR G06F11/27$ OR G06F15/18 OR G06F17/14 OR G06F17/15 OR G06F17/16 OR G06F17/20 OR G06F17/27 OR G06F17/28 OR G06F19/24 OR G06K7/14$ OR G06K9/$ OR G06N3/$ OR G06N5/$ OR G06N7/$ OR G06N99/$ OR G06T1/20 OR G06T1/40$ OR G06T3/40$ OR G06T7/$ OR G06T9/$ OR G08B29/18$ OR G08B29/20$ OR G08B29/22$ OR G08B29/24$ OR G08B29/26$ OR G08B29/28$ OR G10L13/$ OR G10L15/$ OR G10L17/$ OR G10L25/$ OR G10L99/$ OR G11B20/10$ OR G11B20/12$ OR G11B20/14$ OR G11B20/16$ OR G11B20/18$ OR G16H50/20 OR H01M8/04992 OR H02H1/$ OR H02P21/$ OR H02P23/$ OR H03H17/02$ OR H03H17/04$ OR H03H17/06$ OR H04L12/24$ OR H04L12/70$ OR H04L12/751$ OR H04L25/02$ OR H04L25/03$ OR H04L25/04$ OR H04L25/05$ OR H04L25/06$ OR H04L25/08$ OR H04L25/10$ OR H04L25/12$ OR H04L25/14$ OR H04L25/17$ OR H04L25/18$ OR H04L25/20$ OR H04L25/22$ OR H04L25/24$ OR H04L25/26$ OR H04L25/03$ OR H04N21/466$ OR H04R25/$ OR G07C9/$ OR G06F21/$).IPC.
Block 3 C3	N/A for U.S. patent analysis, since the WIPO query pertains to Japanese applications
Block 3 C4	N/A for U.S. patent analysis, since the WIPO query appears to be a Questel classification associated with Japanese applications
Block 3 K2	(CLUSTERING OR (COMPUT$9 ADJ CREATIVITY) OR (DESCRIPTIVE ADJ MODEL$1) OR (INDUCTIVE ADJ REASONING) OR OVERFITTING OR (PREDICTIVE ADJ (ANALYTICS OR MODEL$1)) OR (TARGET ADJ FUNCTION$1) OR ((TEST OR TRAINING OR VALIDATION) NEAR1 DATA NEAR1 SET$1) OR BACKPROPAGATION$1 OR ((SELF ADJ LEARNING) OR (SELFLEARNING)) OR (OBJECTIVE ADJ FUNCTION$1) OR (FEATURE$1 ADJ SELECTION) OR (EMBEDDING$1) OR (ACTIVE ADJ LEARNING) OR (REGRESSION ADJ MODEL$1) OR ((STOCHASTIC OR PROBABILIST$) NEAR2 (APPROACH$ OR TECHNIQUE$1 OR METHOD$1 OR ALGORITHM$1)) OR (RECOMMEND$ ADJ SYSTEM$1) OR ((TEXT OR SPEECH OR ((HAND ADJ WRITING) OR (HANDWRITING)) OR FACIAL OR FACE$1 OR CHARACTER$1) ADJ (ANALYSIS OR ANALYTIC$1 OR RECOGNITION))).ti,ab,clm.
Final query	((Block 1) OR (Block 2) OR (((Block 3 C1) OR (Block 3 C2)) AND (Block 3 K2))) using U.S. Patent and U.S. PGPub databases

Source: WIPO (2019) as modified by USPTO.

REFERENCES

Abood, A. and D. Feltenberger. 2018. "Automated patent landscaping." Artificial Intelligence and Law, 26(2), pp.103-125.

Cockburn, I. M., R. Henderson, and S. Stern. 2019. "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis." In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 115–46. Chicago: University of Chicago Press.

Cox, N. 2004. "Adjacent values in graph_box" response on *The Stata listserver* (June 2). https://www.stata.com/statalist/archive/2004-06/msg00035.html.

Clarivate Derwent. 2020. "Derwent World Patents Index" webpage. https://clarivate.com/derwent/solutions/derwent-world-patent-index-dwpi/.

FCC (Federal Communications Commission). "API Documentation for Developers" webpage. https://geo.fcc.gov/api/census/.

Feltenberger, D. 2019. "Automated patent landscaping" github post (January 11). https://github.com/google/patents-public-data/tree/master/models/landscaping.

Krohn, J., G. Beyleveld, and A. Bassens. 2020. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Boston: Addison-Wesley Professional.

PatentsView. 2020. https://www.patentsview.org.

Persiyanov, D. 2018. "*2Vec File-based Training: API Tutorial." (last commit September 14). https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/Any2Vec_Filebased.ipynb.

Stata Corporation. "Graph box." Stata 13 online manual. https://www.stata.com/manuals13/g-2graphbox.pdf.

Stepner, M. "Maptile." https://michaelstepner.com/maptile.

Toole, A., N. Pairolero, A. Giczy, J. Forman, C. Pulliam, M. Such, K. Chaki, D. Orange, A. Thomas Homescu, J. Frumkin, Y.Y. Chen, V. Gonzales, C. Hannon, S. Melnick, E. Nilsson, and B. Rifkin. 2020. *Inventing AI: Tracing the diffusion of artificial intelligence with U.S. patents*. (October).

Trippe, A., 2015. *Guidelines for Preparing Patent Landscape Reports*. Geneva, Switz.: World Intellectual Property Organization.

USDA (United States Department of Agriculture). Natural Resources Conservation Service. "County FIPS Codes" webpage. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697.

USPTO (United States Patent and Trademark Office). 2020. "Block Data Storage System (BDSS)", version 1.1.0. https://bulkdata.uspto.gov/.

______. 2020. "Classification Resources, Cooperative Patent Classification Scheme", version 2020.08. https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html.

_______. 2020. *Manual of Patent Examining Procedure (MPEP)*, ninth edition, revision 10.2019 (last revised June). https://www.uspto.gov/web/offices/pac/mpep/index.html.

______. 2020. "Patent Classification" webpage (last modified September 2). https://www.uspto.gov/patents-application-process/patent-search/classification-standards-and-development.

______. 2020. "Public Search Facility" webpage (last modified June 5). https://www.uspto.gov/learning-and-resources/support-centers/public-search-facility/public-search-facility.

______. 2020. "PatentsView" webpage (last modified August 6). https://www.uspto.gov/ip-policy/economic-research/patentsview.

Wetherbee, I. 2017. "Google Patents Public Datasets: connecting public, paid, and private patent data." (October 31). https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data.

Wikipedia. 2020. "Lucent" webpage (last edited on August 18). https://en.wikipedia.org/wiki/Lucent.

_______. 2020. "Sun Microsystems" webpage (last edited on September 7). https://en.wikipedia.org/wiki/Sun_Microsystems.

_______. 2020. "Nokia" webpage (last edited on September 18). https://en.wikipedia.org/wiki/Nokia.

WIPO (World Intellectual Property Organization). 2019. *WIPO Technology Trends 2019: Artificial Intelligence*. Geneva, Switz.: World Intellectual Property Organization.

Yellowbrick.com. "Discrimination Threshold" webpage. https://www.scikit-yb.org/en/latest/api/classifier/threshold.html.