WIPO
Technology Trends 2019
Artificial Intelligence

Data collection method and clustering scheme

Background paper

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Data collection method and clustering scheme

Background paper for *WIPO Technology Trends 2019: Artificial Intelligence*

This background paper was commissioned in the context of *WIPO Technology Trends 2019: Artificial Intelligence*. The work was carried out by CNRS, science-miner and WIPO. The lead authors were Sophie Gojon, Adrien Migeon and Philippe Petit for CNRS; Patrice Lopez for science-miner; and Irene Kitsara for WIPO.

Acknowledgements

The paper draws on commissioned background research, based on search strategy and methodology developed by the core analytics team led by Irene Kitsara (WIPO), consisting of Sophie Gojon, Adrien Migeon and Philippe Petit (CNRS Innovation) and Patrice Lopez (science-miner).

The AI dimensions used for the report and the related glossary were developed by Patrice Lopez, who also provided expert advice on AI in patent literature, with inputs by the core team, the WIPO Advanced Technologies Application Center (ATAC) and team members of Mila (Simon Blackburn, Pierre Luc Carrier, Mathieu Germain, Margaux Luck, Gaétan Marceau Caron and Joao Felipe Santos).

Established in 1992, CNRS Innovation is a French public limited company, subsidiary of two major innovation players: CNRS and Bpifrance Financement. CNRS Innovation supports all technology transfer steps, from the evaluation and protection of technologies to the elaboration of the operating strategy.

Founded in June 2015, science-miner is an independent and self-financed company, focusing on scholar text mining, scientific knowledge extraction and knowledge engineering. The goal of science-miner is to accelerate and strengthen science with Machine Learning techniques. The company proposes open source tools and R&D services for helping the scientists to take advantage of the very rapidly growing amount of scientific and technical information, from traditional scholar publications to experimental and primary datasets.

The Montreal Institute for Learning Algorithms (Mila) is the machine learning laboratory of the University of Montreal, founded by Professor Yoshua Bengio. Mila includes researchers, students and scientific staff members working as software developer and specialist programmers in deep learning. It is one of the largest academic laboratories fully focusing on deep neural networks and their applications.

Table of contents

1. Patent collection strategy

a. Data source

Patent data is collected and analyzed using the **FAMPAT database**[1] provided by Questel. The FAMPAT collection indexes patent applications and granted patents from more than a hundred patenting authorities worldwide. FAMPAT organizes the patent applications into simple patent "families" build by grouping all the members sharing the exact same priority numbers, and thus corresponding to the same "invention".

At the time the patent search was conducted, **FAMPAT** contained approximately 59.3 million patent families.

In this report all the figures are given in number of patent families so that different applications associated to the same inventions are only counted once. Unless otherwise noted we use the date of the earliest priority filing of the family in the graphs showing temporal data.

b. Scope of the patent collection

The main objective of the search strategy is to capture all the patent families related to Artificial Intelligence. In order to analyze the evolution of this technology through time, we have chosen to apply **neither time nor geographical limits to the query.**

A first difficulty resides in the fact that nowadays AI is ubiquitous and finds applications in all major industrial sectors. For this reason, an efficient and complete search strategy cannot be limited to a specific industrial domain.

A second difficulty comes from the fact that the definition of AI is very fluid and evolves in time: what was yesterday considered as AI can now be considered as a regular technology, and new technologies are invented every day. It implies that the query may be very broad and has to take into account a very wide range of technologies.

Consequently, the strategy presented in this document has been designed to achieve an appropriate tradeoff between our will to collect **all the AI-related technologies in all their possible forms and uses** and the necessity to control the false positives that necessarily appears in this type of broad queries.

c. Search strategy

We have chosen to base most of our query strategy on the use of classification codes. It has the advantage of overcoming the limits of keywords-based queries when the technological field is intrinsically very large and the relation of the relevant inventions to Artificial Intelligence not always clearly explicit.

However, some crucial differences exist between the different classification systems in use. The scope, completeness, precision and structure of CPC, IPC and FI/F-terms are not the same so that different classification systems need to be treated differently in the final query.

Moreover, AI inventions may or may not always be classified in AI-specific subgroups. To reach completeness, less AI-specific classification groups need to be carefully taken into

[1] https://www.questel.com/wp-content/uploads/2016/04/FamPat-Rules.pdf

account, and we have chosen to control them with an extended list of keywords in order to limit the noise.

Nevertheless, a relatively large amount of AI-related patents are not classified in non-specifically AI-related classification codes and can only be captured using keywords, so that a part of the query has to be based on specific keywords only. In addition, the availability of a classification system is often limited to a subset of the patent publications, for instance FI/F-terms are limited to applications filed at the JPO.

Thus, the chosen query structure is the following:

> **Block 1**: List of CPC codes specific to AI technologies carefully selected and tested.
>
> **Block 2**: List of AI-specific keywords (K1) carefully selected and tested.
>
> **Block 3**: List of non-specific CPC or IPC or FI/F-terms classes controlled by AI-related keywords (K2).

Block 3 takes the form "CPC/IPC/FI/FTERMS classes" AND "list of AI-related keywords (K2)".

The final query is the union of Block 1, Block 2 and Block 3.

All the classification classes have been tested using Orbit database and samples of results have been manually checked in order to measure the relevance of the classes and their degree of specificity.

The keywords have been selected on the basis of a review of existing literature and web resources. All search strings, classification codes and keywords are provided in a section 7 along with a selected bibliography of sources used to select the keywords.

d. Control of the results

Results have been controlled using a built-in text-mining tool provided extraction of the main concepts contained in the tested datasets. The final query has been established using an iterative test/error process: each addition or modification of the query was examined via text-mining in order to identify sources of noise and/or additional keywords to use.

This query building process resulted in the establishment of a first list of results containing 369,211 patent families. This preliminary collection was explored using statistical tools and text-mining so that several source of noise have been identified. Most of the noise was due to old patent families (filed before 1975) related to automation and collected via non-AI specific classification codes. Other sources of noise due to non-AI computation approaches such as watermark embedding, database management and advertisement.

After this control process the final patent collection contains 339,828 patent families.

2. Scientific publication collection strategy

a. Data source

Scientific publication data is collected and analyzed using Elsevier's SCOPUS database. The SCOPUS database is a multidisciplinary database containing more than 70 million items (Articles, Conference Proceedings…) published in peer-reviewed journals[2].

b. Scope and comparability with the patent search strategy

The analysis of the AI scientific publication collection aims at providing statistical figures reflecting the worldwide academic activity in the AI field. These results are of course to be compared with the findings of the patent collection analysis, so that the queries must be as similar as possible.

To do so, we base our publication query strategy on the lists of keywords used in block 2 and block 3 in the patent search strategy. However, as there is no real equivalent to classification codes (IPC, CPC…) in the scientific publication databases, the patent query strategy presented in section 1 cannot be directly "translated" to create the publication query strategy. Two major problems occur:

- One cannot easily create an equivalent to the Block 1 of the patent query strategy
- The Block 3 that consists in a list of AI-related keywords controlled by IPC/CPC classes cannot be directly translated

The first issue was partially solved by extending the list of AI-specific or AI-related keywords to concepts contained in the definitions of the IPC/CPC classes used in Block 1. The use of the Subject Areas defined by Scopus helped solving the second one.

Another difference between patent and publication search strategies is that in general the level of noise induced by the use of certain keywords or keywords combinations might be different between patent and publication data. Then, the keywords lists and control strings or fields have also been adapted to the specificities of patent publication data.

Please note also that the basic search fields for keywords are different between publication and patent databases. Notably, it is generally impossible to search for keywords in the full text of scientific publications, and "Claims" do not exist. However, the Author Keywords are very helpful when searching scientific publications. Also, Scopus automatically classifies the scientific publication using "Indexterms" based on specialized thesauri. The field "Indexterms" is also used in query strategy.

c. Search strategy

The structure of the query strategy chosen to collect the worldwide AI-related scientific publications is the following :

Block A: List of AI-specific keywords or keywords combination (based on Block 2 of the Patent query strategy and augmented with additional keywords based on the definitions of the IPC/CPC codes of Block 1).

Block B: List of AI-related keywords or keywords combination controlled by the following Scopus' Subject Areas : "Mathematics", "Computer Science" or "Engineering".

---

[2] https://www.elsevier.com/solutions/scopus/how-scopus-works/content

The final query is the union of Block A and Block B.

All search strings and keywords are provided in section 7.

d. Control of the results

The final query has been established using an iterative test/error process: each addition or modification of the query was examined in order to identify sources of noise and/or additional keywords to use. The main feature we have examined in order to identify noise was the proportion of non-obviously AI-related Subject Areas or Journals associated to specific keywords or keywords combinations.

On June 15th the AI-related publications collection contains a total of 1,636,649 items.

3. Merger and Acquisition Data

Data related to merging & acquisitions and funding have been extracted from CrunchBase database in May 2018. Data refers to companies tagged in the Artificial Intelligence category group defined by CrunchBase. Although containing a lot of relevant data, this database should not be considered as an exhaustive one as it may lack information about companies originating from non-English speaking countries for instance.

A total 6,538 companies related to AI were listed in CrunchBase at May 2018.

4. Litigation and Opposition Data

Litigation and Opposition data was extracted from two sources. Litigation data was extracted from the Darts-ip database. Opposition data was extract from Darts-ip and Questal.

Darts-ip manages the largest database of intellectual property case law and is the only one of its kind to offer global coverage. Cases are gathered in four domains of intellectual property: patents, trademarks, design and models, and domain names.

Darts-ip Patent Cases Collection Coverage as of 20 August 2018 can be found here:

https://web.archive.org/web/20190123102155/https://www.darts-ip.com/ip-cases-database/case-law-database-patent/patent-coverage/

a. Litigations

Note:

- The litigation cases do not include patent oppositions and their associated follow-up.
- AI patent families involved in the litigation cases do not necessarily belong to the considered player. They only belong to the AI patent collection.
- Figures related to the cases in which companies are flagged as plaintiffs refer only to infringement cases.

b. Oppositions

Note:

- The difference between the number of patent families involved in oppositions and the number of cases is explained by the fact that in an opposition case filed against a single patent, other AI patents may be cited in the procedure.
- Important remark: it is possible to file oppositions anonymously. The figures related to opposing parties intrinsically cannot take into account the players involved in those cases.
- Some cases may not be closed, since appeals are possible during a certain length of time depending on the Patent Offices rules that apply.
- It is sometimes possible for the patent holder to modify the claims of its patent in order to avoid opposition. Depending on the case, it can be sometimes considered as a victory for the opposing party or for the patent holder. Since it would require to read each and every case to classify such situations, when a patent isn't fully revoked it shall be considered a win for the patent holder in the present study.
- The database used here (Darts IP) does not cover the opposition cases filed in the Korean patent office, so that the information about opposing parties is not available for this country.

Opposition data originates from two different sources :

- Darts-ip database that also covers the oppositions filed in the main worldwide offices (but not all of them, such as Korean office that is not covered as indicated in the word document). Here the data is rather complete (office of opposition, opposing parties…) for the covered offices.
- Questel also analyses the legal status of patent families and spots the opposition events in the legal status of patent families for all the offices covered by Orbit (e.g. including Korean office). It is possible to extract the offices of opposition for all the patent families involved in oppositions, however the opposing parties may still be missing however as they are not reported in the legal status of patent families.

5. Clustering scheme

a. Goal and limits of clustering scheme

For supporting visualization and interpretation of the main AI notions, a taxonomy is necessary to cluster the pool of documents,

However, developing a taxonomy is a difficult exercise. A consensus on the nature and the organization of the main concepts of a technical field is hard to reach by experts, in particular in a rapidly evolving field like AI where different well-founded points of view are possible. Over time, notions and relations evolve and the same representation cannot cover easily a field in a large time span. Moreover, science and innovation are by nature a place of debate, where models, new ideas and paradigms are continuously confronted. Having contradicting interpretation and representation is thus inevitable.

We try to limit these issues as follow:

- Focus on a taxonomy adapted to our practical task, with limited scope and complexity, we do not design an ontology or a generic taxonomy usable for multiples of tasks,
- Rely on an existing and well-accepted taxonomy for identifying the AI notions,
- Prioritize flat representations, limiting the number of hierarchical relations,
- Motivate the introduction of new categories by the actual data.

b. Choice of clustering scheme

The taxonomy used to categorize the documents must reflect aspects not only relevant to the different technical characteristics of AI but also the involved end-user applications and the business activities.

We assume from the beginning that each document can be associated to several classes of the taxonomy to capture the different characteristics in a multidimensional space. The multidimensional space should cover:

- The end-user application fields of AI (e.g. finance, marketing, health, etc.), which concretely can be any business activities given the impact of AI today.

- The functions realized by a particular AI work: these are typically the simulations of Human cognitive capacities such as Vision, Language Understanding, Decision making, etc. which are used to build end-user applications in various fields.
- The different AI techniques/algorithms such as machine learning, deep learning, fuzzy logic, etc. which are used to implement different AI functions. For instance, a particular type of neural network can be used to implement a variety of functions such as speech recognition, language understanding or object recognition in a visual

Important notice: The Scientific Publications collection was not clustered by end-users application fields. This is due to the fact that application fields are only marginally mentioned in scientific publications, at least in the document parts that we have access to (title-abstract-author keywords) so that presenting data about application fields would imply a strong bias and create potential misunderstandings about the work done in research labs.

Granularity: The goal of the clustering is to visualize and better understand the structure and the evolution of the AI technologies over time and more particularly in the last 10 years. For readability, the expected number of classes for a view should be typically around ten

and, consequently, we consider that the size of one hierarchical level should also be within this range when possible.

As the taxonomy will reflect the evolution of technologies over time, with some being successful and other declining, it is expected that the taxonomy will be not be balanced. It should be possible to use the taxonomy to focus on certain aspects by exploiting more detailed hierarchical descriptions.

For example Machine Learning (ML) was not very developed in the 80's in comparison to mainstream rule-based systems (expert systems) and fuzzy logics. However, today, ML has evolved into a large range of sub-areas. Understanding trends in ML in the last decade requires a detailed dedicated sub-hierarchy, while fuzzy logics do not require any further descriptions.

Complexity: Another practical aspect in the design of the taxonomy is to avoid introducing too many hierarchical relations for the following reasons:

- Maintain clear visualization: having 3 sub-levels for each AI root aspects for instance would mean to break down each view into many sub-views and make interpretation significantly more complicated without benefit.
- Avoid the lack of consensus about the relations between the categories: depending on the period or the point of view, we might group and sub-structure the categories differently, opening long debates. However, for our purpose, flat representations usually are enough.
- Avoid conflicts with CPC/IPC hierarchies: because we need to rely a lot on the CPC/IPC patent classification to cluster patents, we want to prevent introducing a hierarchical relations conflicting with the hierarchical relations of the CPC/IPC, which would raise issues regarding the inheritance of the keywords and CPC/IPC classes used in the clustering process.

Maturity: Developing a classification scheme is usually an iterative and slow process over many years. Its robustness requires its practical usage to many documents and to match a high-level top-down description with the actual bottom-up reality. We thus prioritize the reuse of an existing taxonomy, rather than developing one ourselves, and we will consider the adaptation of an existing taxonomy to our needs.

c. The ACM classification scheme

We have chosen to base our clustering hierarchies on the ACM classification scheme.

The ACM Computing Classification System (CCS)[3] is the most well-known and extended classification scheme in computing. It has been developed and used during more than 50 years to classify the complete content of the ACM Digital Library. It can thus be considered as a very robust description of the computing field. Relying on such an existing taxonomy is a way to avoid subjective choice from our side and to lead to a more consensual representation.

The primarily goals the ACM CCS are to provide the reader of a given article a quick content reference and to facilitate search for scientific literature. Given the academic nature of this taxonomy, the scientific and technical aspects are particularly deeply developed, and the business and application fields are more succinct.

The following sub-hierarchies are directly relevant to the technical matters of AI:

---

[3] https://dl.acm.org/ccs/ccs.cfm, with a "flat" hierarchical version of all the terms available here https://dl.acm.org/ccs/ccs_flat.cfm

- “Artificial intelligence” which covers AI *functions* and some AI techniques

- “Machine learning” which covers all the AI techniques related to ML

- “Applied computing”: the ACM CCS proposes a sub-hierarchy of application fields for computer science. As we can expect, this sub-hierarchy is extremely broad, covering almost all the business activities.

6. Detail of the clustering hierarchies

The main weakness of the ACM CCS is the date of its latest revision, 2012. Deep Learning is today a major AI technique, but was not considered five years ago in the scheme. Similarly, some new applications are now particularly scrutinized, like IoT, but absent from the ACM CCS. We thus adapted the ACM taxonomy to these recent changes and to findings made by analyzing the patent collection itself.

Note: For all the 3 hierarchies, a document can be associated to several classes from the same hierarchy. For example, a patent describing the embodiment of a machine learning technique would naturally cover at the same time "Supervised learning" (the way a ML model is trained) and a particular ML model such as a "Support Vector Machine" model.

We use Boolean queries with a mix of keywords and classification codes to realize the classification. Then a document appears in a cluster only if it **explicitly** mentions that it uses a technique, or performs a functional application, or is used in a specific application field, or if the patent family is tagged with a relevant classification code.

a. AI application fields

Derived from the "Applied computing" ACM CCS sub-hierarchy, we considered 20 different AI application fields. Some new fields were introduced to match new emerging business activities from the last 5 years, and some fields were merged to better reflect today impact of AI. The sub-fields have been introduced for the largest fields in term of retrieved documents. They are optional and could be used if relevant for understanding new trends at a more granular level.

Field	Sub-fields	Note
Physical sciences and engineering		Covering Archaeology, Astronomy, Chemistry, Earth and atmospheric sciences, Environmental sciences, Engineering, Computer-aided design, Physics, Mathematics and statistics, Electronics
Life and medical sciences	Physiological parameters monitoring Medical imaging Genetics/Genomics Public Health Medical informatics Bioinformatics (Computational biology) Preparation of (or involving)	Covering all biological and medical domains.

	biological compounds/materials Biomechanics Neuroscience/Neurorobotics Nutrition/Food science Drug discovery	
Law, social and behavioral sciences	Intellectual property	Intellectual property was added as sub-field to emphasize a particular area of interest for the WIPO community
Security	Cryptography Authentication Anomaly detection and surveillance Privacy/anonymity Cybersecurity	The introduced sub-fields cover the main aspects of security relevant to AI.
Arts and humanities		
Transportation	Autonomous vehicles Transportation/traffic engineering Driver/vehicle recognition Aerospace/Avionics	Added to ACM CCS to cover autonomous vehicles and other important application of AI.
Industry and manufacturing		
Education		
Document management and publishing		Merging of “Document processing” and “Publishing” from the original ACM to covers all the document management activities (e.g, document database, text editor, digitalization, etc.) including digital and web publishing.
Military		

Cartography		This field includes geo-localization, GPS-based applications and map services.
Agriculture		
Computing in government		
Personal devices, computing and HCI	Affective computing	This field includes Human-Machine Interface (HCI), Touch screens and displays, User interface. Affective computing has been added as sub-field to reflect recent trends.
Banking and finance	Insurance	Added to ACM CCS, covering financial trading
Telecommunications	Computer networks/Internet VoIP Telephony Radio and Television broadcasting Videoconferencing	Added to ACM CSS as a root field, because it was only considered as a subfield of engineering. "Telecommunications" covers all means of transmission (from radio/TV to internet) of data (from image, speech to sensor data). It includes specific applications like network monitoring and routing.
Networks	Social networks, IoT, Smart Cities	Added to ACM CCS to cover all emerging social network activities
Business	E-commerce, Enterprise computing, Customer services	Added as root category to cover commercial and enterprise business related activities, including customer services
Energy management		
Entertainment		Video games, special visual effects

b. AI functional applications

Derived from the "Artificial Intelligence" ACM CCS sub-hierarchy, we introduced 9 AI functional applications. The sub-fields are optional, i.e. they could be considered to focus on a particular field if more information appears relevant. They are particularly relevant to Computer Vision which is a very large cluster.

Field	Sub-fields	Note
Natural language processing	Information extraction, Machine translation, Dialogue, Natural language generation, Semantics, Morphology, Sentiment analysis	Covers all text processing in the sense of the Human capacity to apprehend text (read, summarize, classify, etc.). "Dialogue" covers Intelligent personal assistant, chatbot and Question answering systems. "Sentiment analysis" has been added to reflect the recent trend in NLP.
Speech processing	Speech recognition, Speech synthesis, Speech-to-speech, Speaker recognition, Phonology	Speech recognition includes dictation systems.
Knowledge representation and reasoning		
Planning and scheduling		
Control methods		
Distributed artificial intelligence		This field covers Multi-agent systems, Intelligent agents, Mobile agents, Cooperation and coordination and Swarm intelligence
Robotics		
Computer vision	Biometrics Scene understanding and Vision for robotics	This includes in particular image recognition, video analysis, object recognition and tracking, augmented and

	Image and video segmentation Object tracking Character recognition Augmented reality	virtual reality, visual biometrics
Predictive analytics	Recommender systems	Product recommendation falls under this class

c. AI techniques

Derived from the several ACM CCS sub-hierarchy and in particular “Machine Learning”, we introduced 6 main AI functional techniques completed by several sub-fields. The sub-fields are particularly relevant to the Machine Learning techniques/algorithms which is a very large cluster. Sub-clusters for Machine Learning are required to better understand the trends of this paradigm which, today, massively dominates the AI technology.

Field	Sub-fields	Note
Logic programming	Description logics, Expert systems	Semantic Web is the most well-known implementation of description logics
Fuzzy logic		
Probabilistic reasoning		
Ontology engineering		
Machine learning	Supervised learning, Unsupervised learning, Reinforcement learning, Multi-task learning, Classification and regression trees, Support vector machines, Neural networks, Deep Learning, Logical and relational learning,	In general we consider in a sub-field the mentioned technique and all its variants, e.g. Support Vector Machine (SVM) and its variants like Support Vector Clustering (SVC). Deep Learning was added as subfield. To avoid introducing too many hierarchical relations, Neural Network and Deep Learning are at

	Probabilistic graphical models, Rule learning, Instance-based learning, Latent representations, Bio-inspired approaches	the same level. Bio-inspired approaches include Genetic programing and artificial life.
Search methods		"Search methods" in AI refer to approaches reformulating an AI problem into the efficient retrieval of information stored within some data structure, often integrating heuristics."

All corresponding clustering queries are provided in a separate annex in Excel format.

7. Search strings

The **CPC, FI/F-terms and IPC classes** have been selected using:

- The background knowledge of the experts involved in the query building (Patrice Lopez, Irene Kitsara, Philippe Petit, Sophie Gojon).
- A deep exploration of the CPC, IPC and FI/F-terms classification trees using AI-related keywords.
- A comparison to the classification codes reported in already existing patent landscapes on artificial intelligence.

All the classes have been tested using Orbit database and samples of results have been manually checked in order to measure the relevance of the classes and their degree of specificity.

Please note that the vast majority of patents captured thanks to the use of FI/FTERMS are also captured either by Block 2 or by the IPC codes thanks to the facts that first : the Orbit database contains automatic translation of non-latin patent publications (namely here : Japanese publications) and second : 99% of these patents also possess IPC codes.

As explained in section 1, the noise that might come from the use of non-specific classifications codes is controlled by carefully selected keywords (**block 3**).

Two different lists of **keywords** have been selected and tested:

- A first list of keywords related to core AI concepts (K1)
- A second list of keywords related to general computing or mathematical concepts frequently used in AI technologies (but not specific to them). These keywords are used in combination with a list of non-specific CPC or IPC or FI/F-terms classes to control the noise while maintaining a high degree of completeness in the query.

The keywords list K1 has been established via a deep exploration of AI-related bibliographic records. We have selected these keywords based on their high degree of specificity.

The keywords list K2 has been selected in order to increase the completeness of the query and capture inventions related to AI but not mentioning specific AI concepts.

a. Search fields

The CPC, IPC and FI/F-terms are searched in the relevant Orbit indexes: /CPC for CPC classes, /IPC for IPC classes and /FI or /FTM for FI/F-terms. A specific effort has been made to ensure that all dependent subclasses are taken into account in the final query. For instance when the CPC group G06K9/00 is used in the query, all dependent groups are taken, from G06K9/00006 to G06K9/82.

The keywords used in block 2 and block 3 are searched in the following indexes:

- **/BI: English title**: all stages of publication: English language machine translations are included for the following publications, and are replaced with the official English translations when available: WO, EP, BR, CN, DE, DK, ES, FI, FR, JP, KR, RU, SE, TH and TW.
- **/BI: English abstract**: Original, or official translation from the EPO or machine translation by Questel. If there is no official English abstract available, the field will contain the English abstract of a family member if possible, or machine translated abstracts from the following authorities: WO, EP, BR, CN, DE, DK, ES, FI, FR, JP, KR, RU, SE, TH and TW, to be replaced by the official version when available.
- **/CLMS: English claims**: Original English publications or machine translations

- **/OBJ: Object of the invention**: Extracted from the full text of original English publications or machine translations using semantic technologies. Information contained in the full text (limited to the fields "description of the invention" and "advantages relative to the prior art" is analyzed via a contextual analysis tool.

Note that the search field /BI contains both English title and English abstract.

We chose not to search keywords in the full text to avoid collecting a large amount of false positives.

b. Search language

Based on the fact that the whole FAMPAT collection is either published in English or Machine-translated in English we chose to use keywords in english language only and thus not to use keywords in other latin languages such as French or German.

Furthermore, queries in non-latin languages (Chinese, Korean, Japanese) have been excluded after an in-depth test session: using literal translations to Chinese, Korean and Japanese languages of a list of english keywords related to core AI concepts, 98% of the patent families (33671 patent families) collected via non-latin language keywords lists were captured by the English keywords list and the remaining 2% (781 patent families) were in their vast majority consisting of irrelevant patent families.

c. Detailed search strings

As mentioned before the final query takes the form of the union of the three main search blocks: **Final query** = **Block 1** OR **Block 2** OR **Block 3**.

Details of the search strings used in the blocks are given below:

Block 1: (Y10S-706 OR G06N-003 OR G06N-005/003:G06N-005/027 OR G06N-007/005:G06N-007/06 OR G06N-099/005 OR G06T2207/20081 OR G06T2207/20084 OR G06T-003/4046 OR G06T-009/002 OR G06F-017/16 OR G05B-013/027 OR G05B-013/0275 OR G05B-013/028 OR G05B-013/0285 OR G05B-013/029 OR G05B-013/0295 OR G05B-2219/33002 OR G05D-001/0088 OR G06K-009 OR G10L-015 OR G10L-017 OR G06F-017/27:G06F-017/2795 OR G06F-017/28:G06F-017/289 OR G06F-017/30029:G06F-017/30035 OR G06F-017/30247:G06F-017/30262 OR G06F-017/30401 OR G06F-017/3043 OR G06F-017/30522:G06F-017/3053 OR G06F-017/30654 OR G06F-017/30663 OR G06F-017/30666 OR G06F-017/30669 OR G06F-017/30672 OR G06F-017/30684 OR G06F-017/30687 OR G06F-017/3069 OR G06F-017/30702 OR G06F-017/30705:G06F-017/30713 OR G06F-017/30731:G06F-017/30737 OR G06F-017/30743:G06F-017/30746 OR G06F-017/30784:G06F-017/30814 OR G06F-019/24 OR G06F-019/707 OR G01R-031/2846:G01R-031/2848 OR G01N-2201/1296 OR G01N-029/4481 OR G01N-033/0034 OR G01R-031/3651 OR G01S-007/417 OR G06N-003/004:G06N-003/008 OR G06F-011/1476 OR G06F-011/2257 OR G06F-011/2263 OR G06F-015/18 OR G06F-2207/4824 OR G06K-007/1482 OR G06N-007/046 OR G11B-020/10518 OR G10H-2250/151 OR G10H-2250/311 OR G10K-2210/3024 OR H01J-2237/30427 OR H01M-008/04992 OR H02H-001/0092 OR H02P-021/0014 OR H02P-023/0018 OR H03H-2017/0208 OR H03H-2222/04 OR H04L-2012/5686 OR H04L-2025/03464 OR H04L-2025/03554 OR H04L-025/0254 OR H04L-025/03165 OR H04L-041/16 OR H04L-045/08 OR H04N-021/4662:H04N-021/4666 OR H04Q-2213/054 OR H04Q-2213/13343 OR H04Q-2213/343 OR H04R-025/507 OR G08B-029/186 OR B60G-2600/1876 OR B60G-2600/1878 OR B60G-2600/1879 OR B64G-2001/247 OR E21B-2041/0028 OR B23K-031/006 OR B29C-2945/76979 OR B29C-066/965 OR B25J-009/161 OR A61B-005/7264:A61B-005/7267 OR Y10S-128/924 OR Y10S-128/925 OR F02D-041/1405 OR F03D-007/046 OR F05B-2270/707 OR F05B-2270/709 OR F16H-2061/0081 OR F16H-2061/0084 OR B60W-030/06

OR B60W-030/10:B60W-030/12 OR B60W-030/14:B60W-030/17 OR B62D-015/0285 OR G06T-2207/30248:G06T-2207/30268 OR G06T-2207/30236 OR G05D-001 OR A61B-005/7267 OR F05D-2270/709 OR G06T-2207/20084 OR G10K-2210/3038 OR G10L-025/30 OR H04N-021/4666 OR A63F-013/67 OR G06F-017/2282)/**CPC**

Block 2 = **K1** with:

K1 = (((ARTIFIC+ OR COMPUTATION+) 1W INTELLIGEN+) OR (NEURAL 1W NETWORK+) OR NEURAL_NETWORK+ OR NEURAL_NETWORK+ OR (BAYES+ 1W NETWORK+) OR BAYESIAN-NETWORK+ OR BAYESIAN_NETWORK+ OR (CHATBOT?) OR (DATA 1W MINING+) OR (DECISION 1W MODEL?) OR (DEEP 1W LEARNING+) OR DEEP-LEARNING+ OR DEEP_LEARNING+ OR (GENETIC 1W ALGORITHM?) OR ((INDUCTIVE 1W LOGIC) 1D PROGRAMM+) OR (MACHINE 1W LEARNING+) OR MACHINE_LEARNING+ OR MACHINE-LEARNING+ OR ((NATURAL 1D LANGUAGE) 1W (GENERATION OR PROCESSING)) OR (REINFORCEMENT 1W LEARNING) OR (SUPERVISED 1W (LEARNING+ OR TRAINING)) OR SUPERVISED-LEARNING+ OR SUPERVISED_LEARNING+ OR (SWARM 1W INTELLIGEN+) OR SWARM-INTELLIGEN+ OR SWARM_INTELLIGEN+ OR (UNSUPERVISED 1W (LEARNING+ OR TRAINING)) OR UNSUPERVISED-LEARNING+ OR UNSUPERVISED_LEARNING+ OR (SEMI-SUPERVISED 1W (LEARNING+ OR TRAINING)) OR SEMI-SUPERVISED-LEARNING OR SEMI_SUPERVISED_LEARNING+OR CONNECTIONIS# OR (EXPERT 1W SYSTEM?) OR (FUZZY 1W LOGIC?) OR TRANSFER-LEARNING OR TRANSFER_LEARNING OR (TRANSFER 1W LEARNING) OR (LEARNING 3W ALGORITHM?) OR (LEARNING 1W MODEL?) OR (SUPPORT VECTOR MACHINE?) OR (RANDOM FOREST?) OR (DECISION TREE?) OR (GRADIENT TREE BOOSTING) OR (XGBOOST) OR ADABOOST OR RANKBOOST OR (LOGISTIC REGRESSION) OR (STOCHASTIC GRADIENT DESCENT) OR (MULTILAYER PERCEPTRON?) OR (LATENT SEMANTIC ANALYSIS) OR (LATENT DIRICHLET ALLOCATION) OR (MULTI-AGENT SYSTEM?) OR (HIDDEN MARKOV MODEL?))/**BI/OBJ/CLMS**

Block 3 = (**C1** OR **C2** OR **C3** OR **C4**) AND (**K2**) with:

C1 = (G06T-007 OR G06T-001/20 OR G10L-013 OR G10L-025 OR G10L-099 OR G06F-017/14:G06F-017/148 OR G06F-017/153 OR G10H-2250/005:G10H-2250/021 OR G06F-017/50 OR G06Q-030/02:G06Q-030/0284 OR G07C-009 OR G06F-021)/**CPC**

C2 = (A61B-005 OR A63F-013/67 OR B23K-031 OR B25J-009/16:B25J-009/20 OR B29C-065 OR B60W-030/06 OR B60W-030/10:B60W-030/12 OR B60W-030/14:B60W-030/17 OR B62D-015/02:B62D-015/0295 OR B64G-001/24:B64G-001/38 OR E21B-041 OR F02D-041/14:F02D-041/16 OR F03D-007/04:F03D-007/048 OR F16H-061 OR G01N-029/44:G01N-029/52 OR G01N-033 OR G01R-031/28:G01R-031/31937 OR G01R-031/36:G01R-031/3696 OR G01S-007/41:G01S-007/418 OR G05B-013/02:G05B-013/048 OR G05D-001 OR G06F-009/44+ OR G06F-011/14:G06F-011/1497 OR G06F-011/22:G06F-011/277 OR G06F-015/18 OR G06F-017/14 OR G06F-017/15 OR G06F-017/16 OR G06F-017/20 OR G06F-017/27 OR G06F-017/28 OR G06F-019/24 OR G06K-007/14:G06K-007/1495 OR G06K-009 OR G06N-003 OR G06N-005 OR G06N-007 OR G06N-099 OR G06T-001/20 OR G06T-001/40+ OR G06T-003/40:G06T-003/4092 OR G06T-007 OR G06T-009 OR G08B-029/18:G08B-029/28 OR G10L-013 OR G10L-015 OR G10L-017 OR G10L-025 OR G10L-099 OR G11B-020/10:G11B-020/18 OR G16H-050/20 OR H01M-008/04992 OR H02H-001 OR H02P-021 OR H02P-023 OR H03H-017/02:H03H-017/06 OR H04L-012/24+ OR H04L-012/70+ OR H04L-012/751+ OR H04L-025/02:H04L-025/26 OR H04L-025/03:H04L-025/03993 OR H04N-021/466:H04N-021/4668 OR H04R-025 OR G07C-009 OR G06F-021)/**IPC**

C3 = (G06N3/02:G06N3/10 OR G06N3/08 OR G06N99 OR G06N7/04 OR G06K9 OR G06K9/00 OR G10L13 OR G10L25 OR G10L15 OR G10L17 OR G10L99 OR G06F17/27

OR G06F17/28 OR G06F17/30180A:G06F17/30180C OR G06F17/30210A OR G06F17/30210D OR G06F17/30220A OR G06F17/30310C OR G06F17/30330 C OR G06K9 OR G06F19/00130 OR G06N3/00140 OR G06F11/14676 OR G06F11/22657 OR G06F11/22663 OR G06K7/14082 OR H01M8/04992 OR H04N21/466 OR B60W30/06 OR B60W30/10:B60W30/12 OR B60W30/14:B60W30/17 OR F02D41/14310H)/**FI**

C4 = (5B078+ OR 5B178+ OR 5B064+ OR 5L096FA+ OR 5L096GA+ OR 5L096HA+ OR 5L096JA+ OR 5L096KA+ OR 5L096MA07 OR 5B043+ OR 5B064+ OR 5B057CH+ OR 5B057DA+ OR 5B057DC+ OR 5H004KD23 OR 5H004KD31 OR 5H004 KD32 OR 5H004KD33 OR 5H004KD35 OR 5H004KD63 OR 5H301DD02 OR 5H301JJ+ OR 5H301LL+ OR 5D045+ OR 5D015+ OR 5B056BB+ OR 5B056HH03 OR 5B056HH05 OR 5B109QA+ OR 5B109RD02 OR 5B109RD03 OR 5B091+ OR 5B075NK3+ OR 5B075PP04 OR 5B075PP24 OR 5B075PP25 OR 5B075QP+ OR 5B075QT04 OR 5B075QT05 OR 5B064+ OR 5L049DD04 OR 5J070BF16 OR 5B078+ OR 5B048DD12 OR 5K030KA07 OR 5K030KA18 OR 5K030KA20 OR 5C164PA43 OR 5C164YA12 OR 5C087GG02 OR 3D241AF05 OR 3D241AF07 OR 3D241BA+ OR 3D241CE05 OR 3D241CE06 OR 3D241CE08 OR 3D241CE10 OR 3C707KT11 OR 3C707 LW1+ OR 4C117XJ31 OR 4C117XK11 OR 3G301ND2+ OR 3G301ND3+ OR 3G301ND43 OR 3J552TA11 OR 3J552TA12 OR 3J552TA18 OR 3J552TA19 OR 3J552TA20)/**FTM**

K2 = (CLUSTERING OR (COMPUT+ CREATIVITY) OR (DESCRIPTIVE MODEL?) OR (INDUCTIVE REASONING) OR OVERFITTING OR (PREDICTIVE 1W (ANALYTICS OR MODEL?)) OR (TARGET 1W FUNCTION?) OR ((TEST OR TRAINING OR VALIDATION) 1D DATA 1D SET?) OR BACKPROPAGATION? OR SELF-LEARNING OR SELF_LEARNING OR (OBJECTIVE FUNCTION?) OR (FEATURE? SELECTION) OR (EMBEDDING?) OR (ACTIVE LEARNING) OR (REGRESSION MODEL?) OR ((STOCHASTIC OR PROBABILIST+) 2D (APPROACH+ OR TECHNIQUE? OR METHOD? OR ALGORITHM?)) OR (RECOMMEND+ SYSTEM?) OR ((TEXT OR SPEECH OR HAND_WRITING OR FACIAL OR FACE? OR CHARACTER?) 1W (ANALYSIS OR ANALYTIC? OR RECOGNITION)))/**BI/OBJ/CLMS**

8. Selected Bibliography

Web resources:

https://ai100.stanford.edu/2016-report/section-i-what-artificial-intelligence/defining-ai

https://ai100.stanford.edu/2016-report/appendix-i-short-history-ai

AI Index 2017 Report https://aiindex.org/2017-report.pdf

https://www.forbes.com/sites/bernardmarr/2016/12/08/what-is-the-difference-between-deep-learning-machine-learning-and-ai/

https://en.wikipedia.org/wiki/Artificial_intelligence

https://phrasee.co/ultimate-glossary-artificial-intelligence-terms/

https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/part-1-artificial-intelligence-defined.html#

https://www2.deloitte.com/content/dam/Deloitte/at/Documents/human-capital/artificial-intelligence-innovation-report.pdf

https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance include a number of definitions of encyclopedias

https://ai100.stanford.edu/2016-report/section-i-what-artificial-intelligence/ai-research-trends and trends by domain https://ai100.stanford.edu/2016-report/section-ii-ai-domain

https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf

McKinseys State of ML and AI 2017 https://www.forbes.com/sites/louiscolumbus/2017/07/09/mckinseys-state-of-machine-learning-and-ai-2017/#3f82e1c75b64 and https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx.

https://www.pwc.com/gx/en/industries/communications/assets/pwc-ai-and-iot.pdf

https://aisociety.fi/our-definition-artificial-intelligence-and-scope-research

Gartner IT Glossary: https://www.gartner.com/it-glossary/artificial-intelligence/

Free access books:

Poole and Mackworth (2017): Artificial Intelligence: Foundations of Computational Agents, 2nd Edition: http://artint.info/2e/html/ArtInt2e.html