Evolution and scientific visualization of Machine learning field

Río-Belver, R., Garechana, G., Bildosola, I., and Zarrabeitia, E.
T.F.M. Technology Foresight and Management Research Group

Departamento de Organización de Empresas
University of the Basque Country UPV/EHU
author email: rosamaria.rio@ehu.eus

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

CAMPUS OF INTERNATIONAL EXCELLENCE

Index

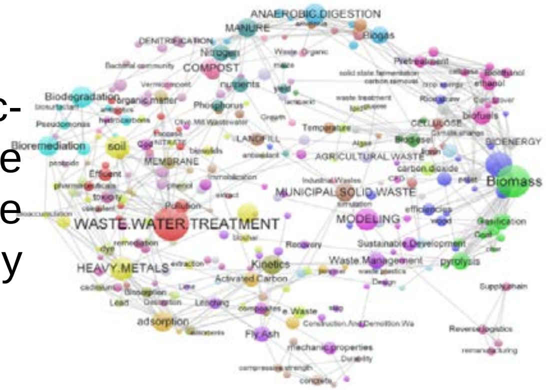
Evolution & Scientific visualization Machine learning

The competence of T.F.M. is the processing and analysis of technological data, applied to the detection, assessment and incorporation of new technologies in the industry.

Technology Management; Tech-mining; Technology maps; Foresight; Roadmaps; Knowledge management; Innovation; Competitive Intelligence

<u>Our research lines are the following:</u>

Technology. – It applies tech-mining to the analysis of the scientific-technological Information. Its objective is to know the state of the technology for which large amounts of technological information are identified, recovered and dealt with. This information is statistically analyzed and visualized using Technological Maps.



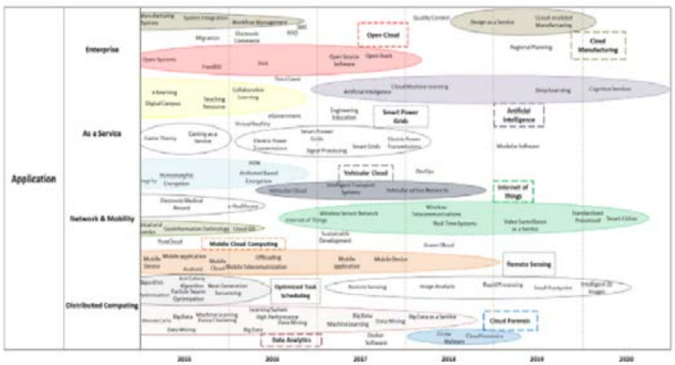
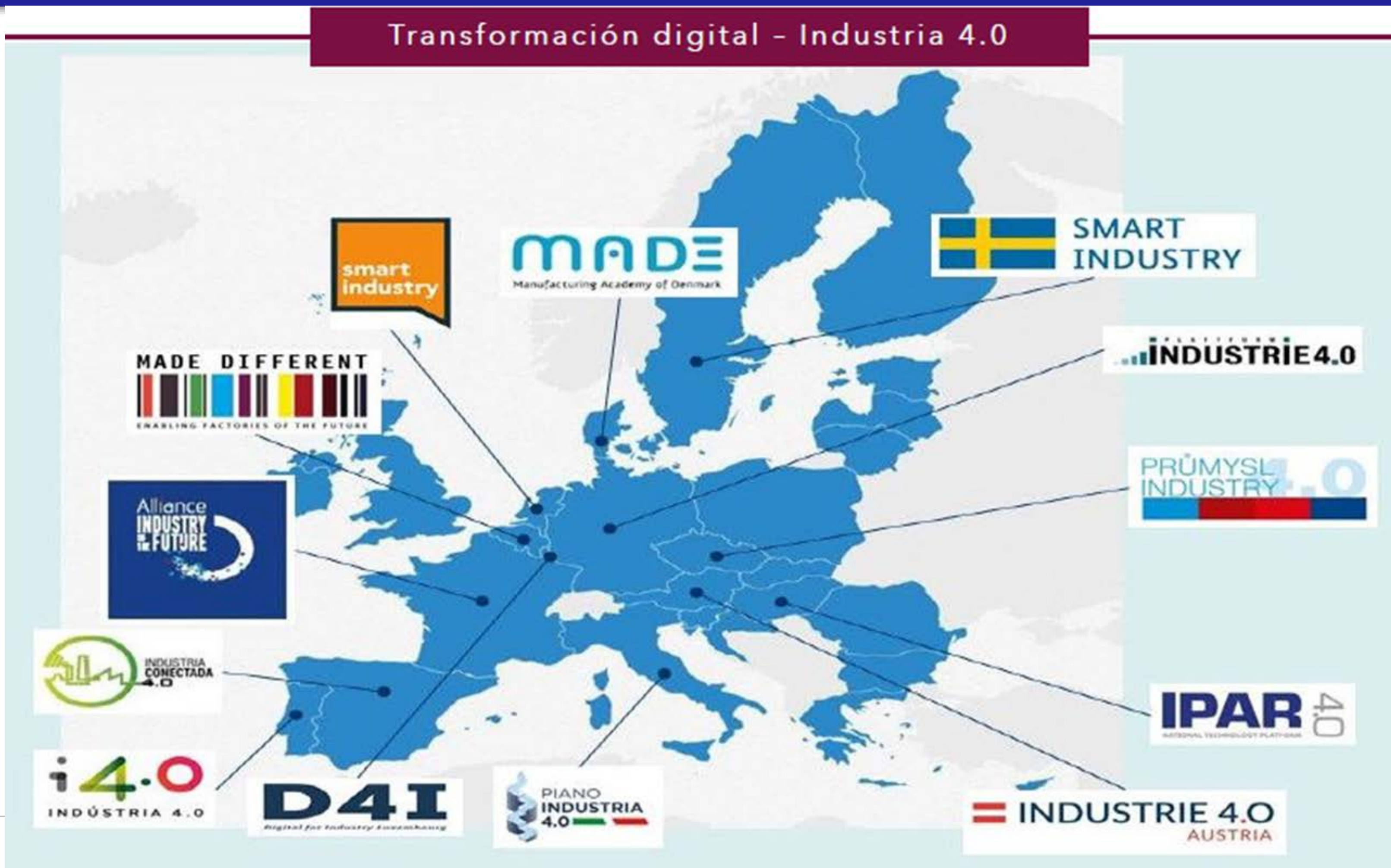


Fig 9. Zoom of the TRM which represents the period 2015–2020 for the application layer.

Foresight. – Its aim is to detect emerging technologies in the form of weak signals and to anticipate. It applies text mining with the purpose of extracting patterns and technological tendencies. It generates an output in the form of a Technology Roadmaps.

Technology Management. – Intelligence to support company decisions. Development of Advanced Competitive Intelligent Systems based in technological intelligence.

Industry 4.0
is generating an unprecedented revolution in the manufacturing sector



the new industrial revolution is understood as the **incorporation of fundamentally digital technologies** conductive to achieving the **smart factory**. (OECD)

The **application** of **automatic learning methods** defined by Alpaydin (2014) to **industrial production** will be **one of the pillars** of the new revolution

Understanding the keys to the development of the machine learning discipline makes it possible to understand the transfer process from algorithm development in the laboratory to machine programming in industry

This article provides a retrospective and an understanding of the development of automatic learning methods

How?

Using Bibliometrics text-mining methods, natural language processing and network theory

For?

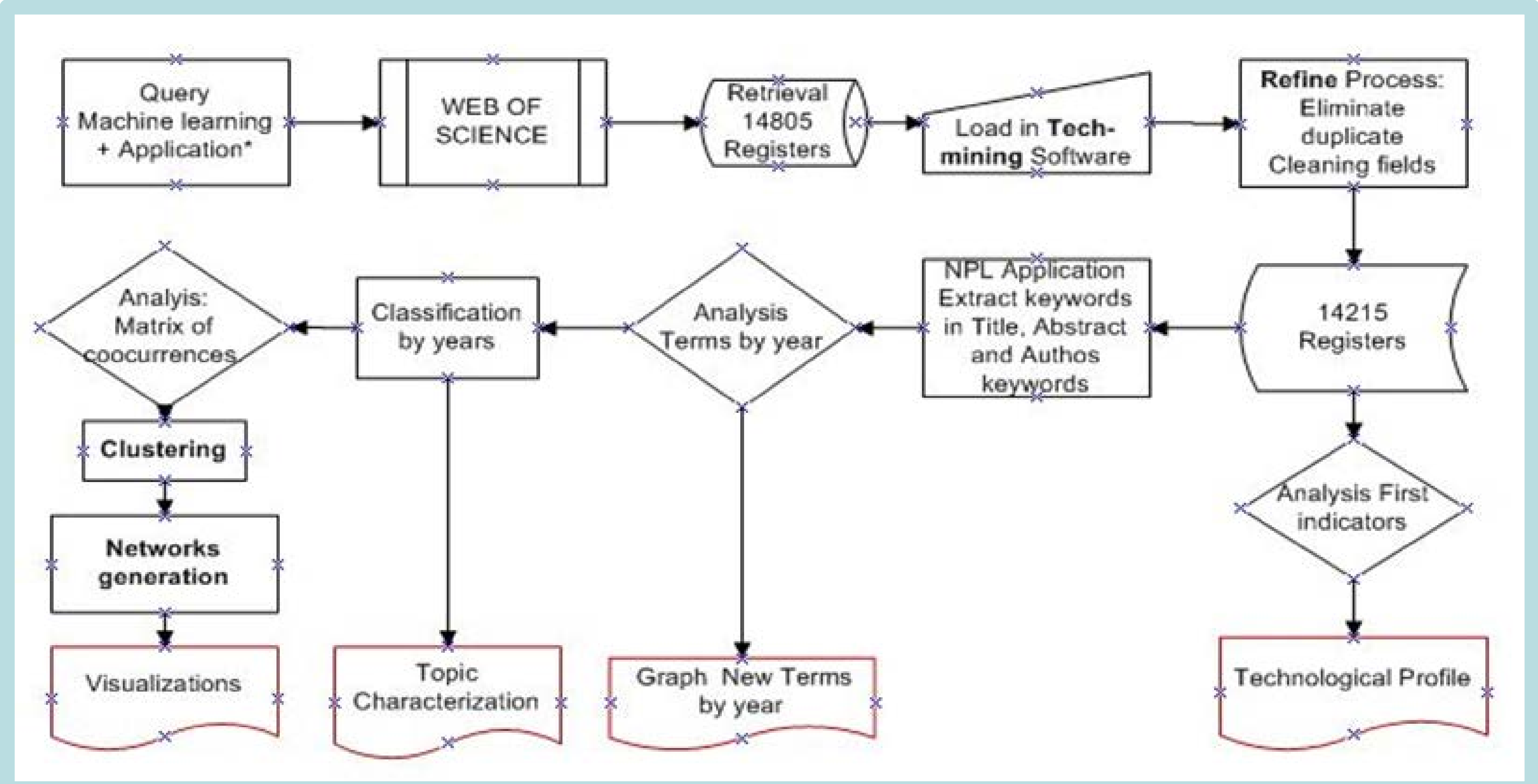
Define the stages of development, growth and maturity of the discipline Machine learning

The beginnings are visualized as a discipline within **Computer Sciences** in the **subcategory of Artificial Intelligence**, its development and the current **transfer of knowledge to other areas of Engineering** and its **industrial applications.**

Finally, we look for making technology transfer to industry visible

The Machine learning and applications (MLAA) data set has been obtained by retrieving the articles, conference proceedings and book chapters published from **1986 to 2017** from Web of Science, core collection.

The concepts ("Machine learning") AND (Application*) have been searched in the **fields Title, Abstract, Author Keywords and Keywords plus**, detecting **14805** records that were downloaded in their full record format.



The 14 805 records were imported in a software for tech-mining Vantage-point.

Cleaning Step

Eliminate duplicates
Group same authors, same Institutions… using fuzzy algoritms and create thesaurus

Tecnological Profile

Natural Lenguaje Processing

Extract Keywords from the fields

Quantitative data:

Main authors

More cited articles

Main Institutions

First Year of appearance
highlighting the evolution of the discipline

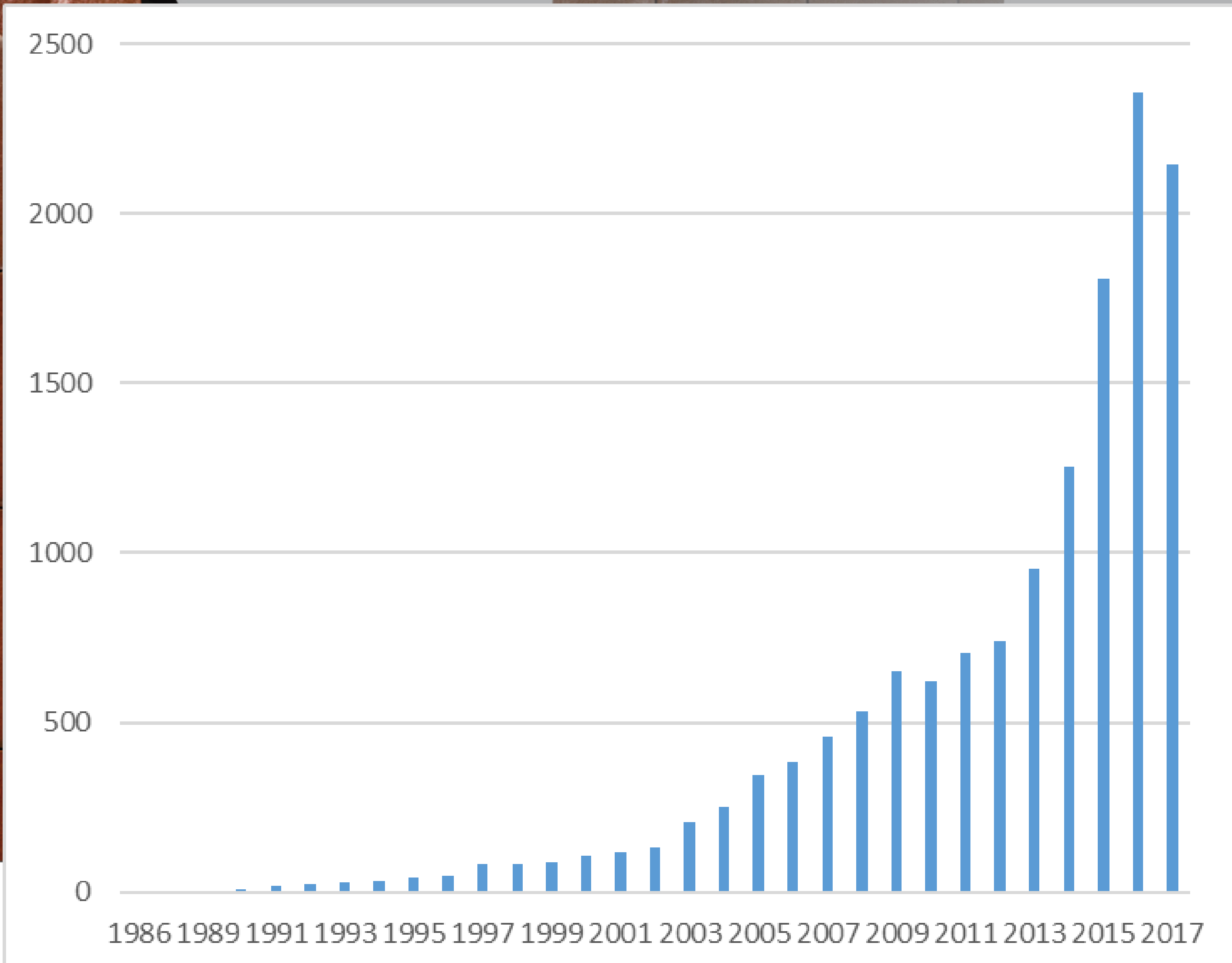
Create Sub-data sets

Network analysis

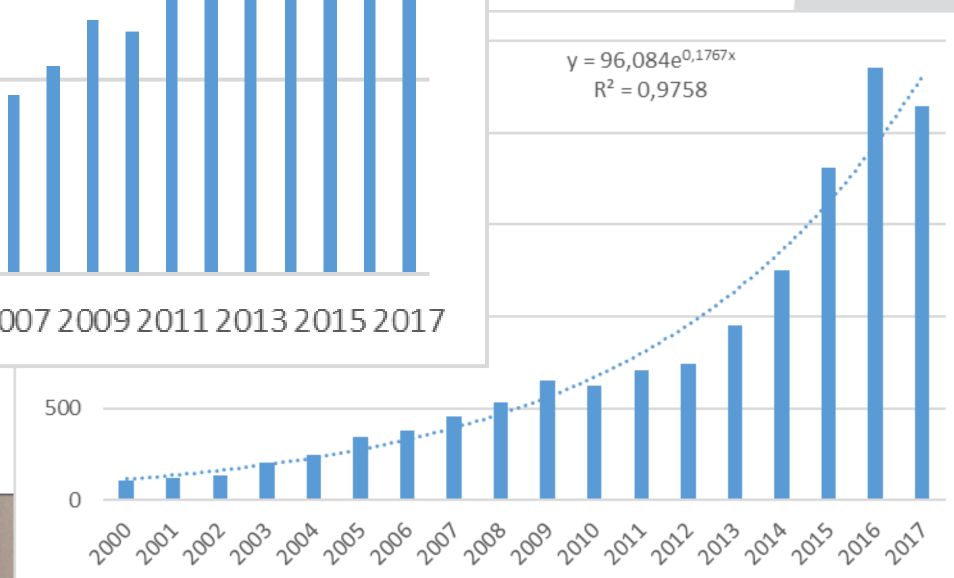
Autocorrelation maps of the Web of Science categories field

Visualizations

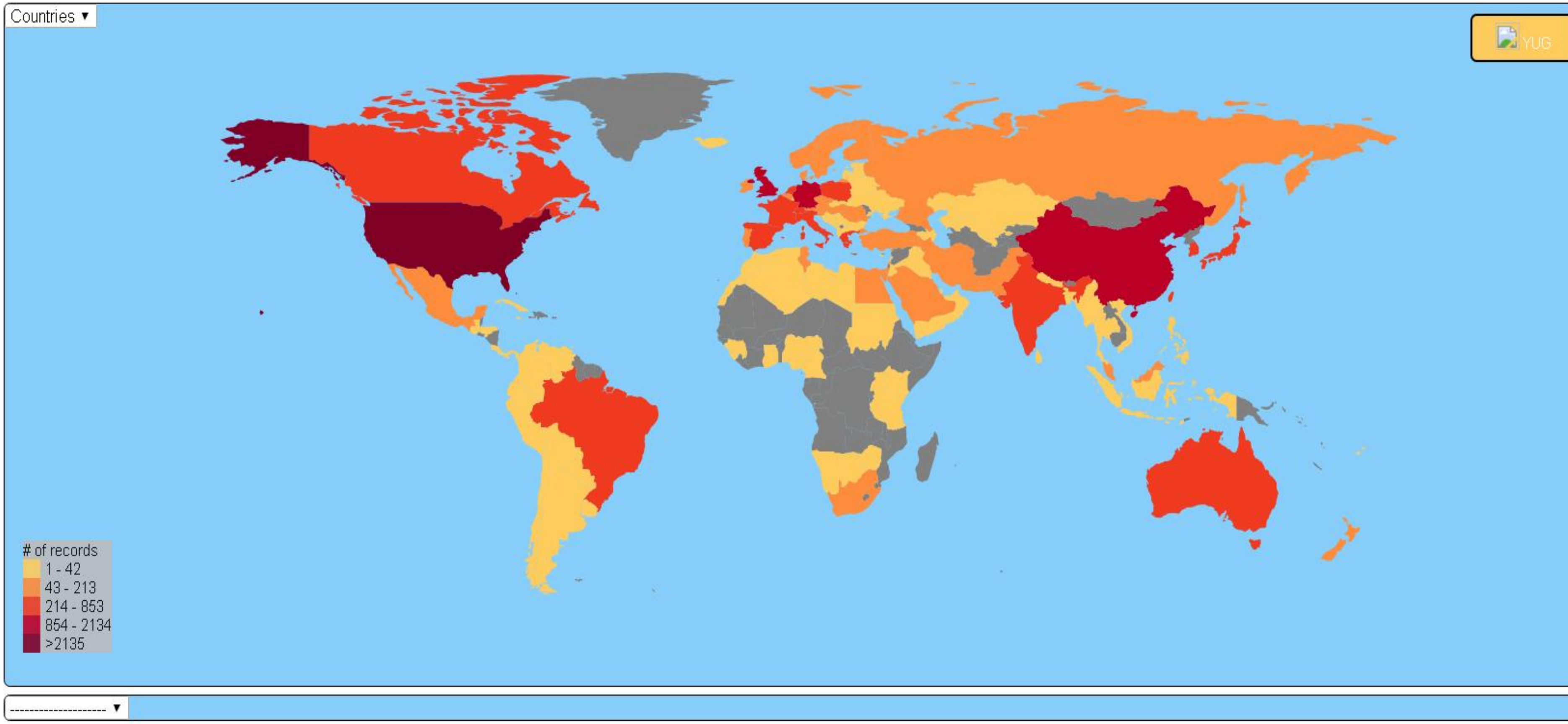
Vos Viewer



Number of Total Publications by year

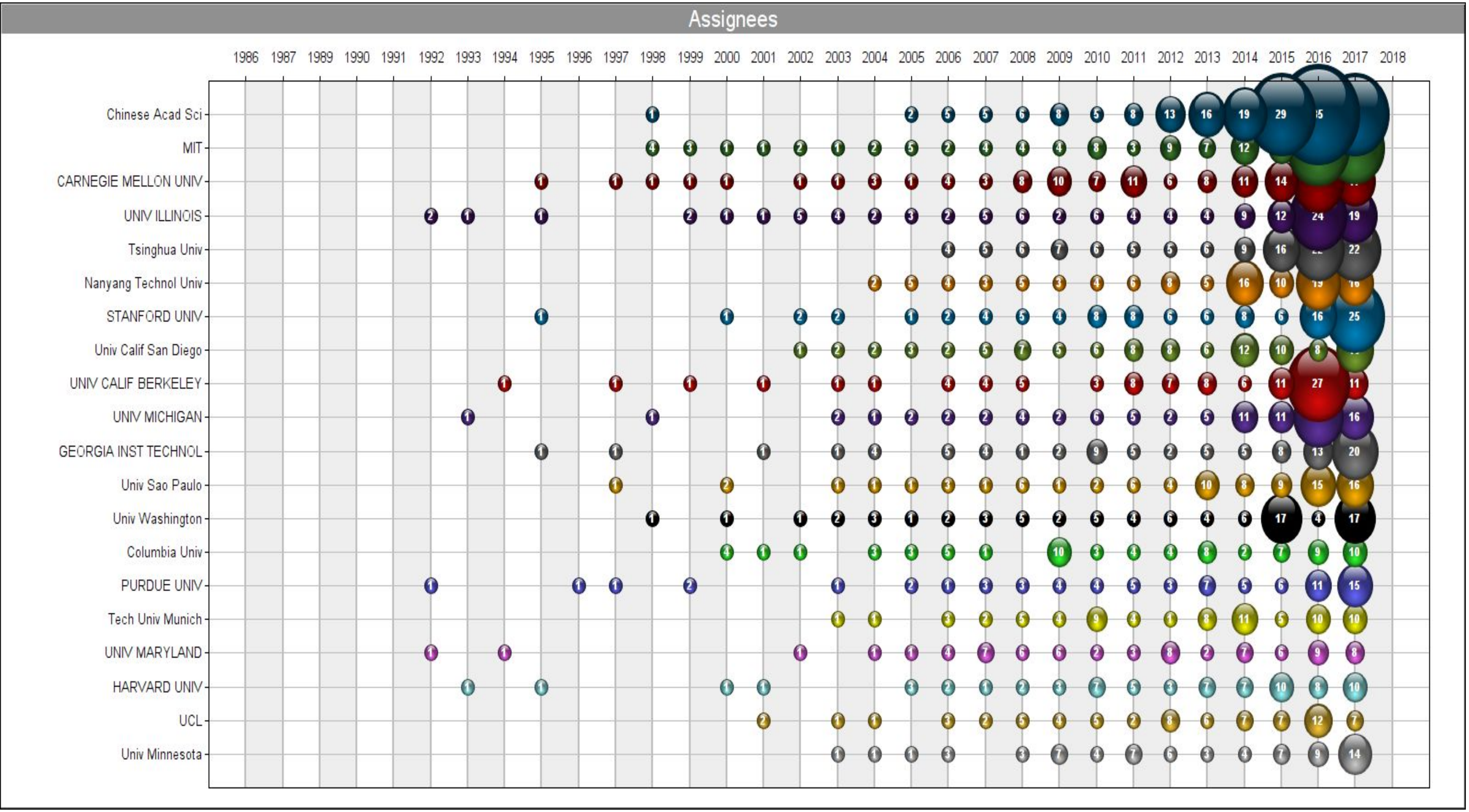


Map of Country Fields



20 Main Contributor Countries

Nº Records	Countries	Nº Records	Countries
4270	USA	446	Japan
2002	China	302	Brazil
1124	UK	266	South Korea
914	Germany	262	Switzerland
820	India	249	Poland
649	Spain	234	Taiwan
593	Canada	221	Greece
580	Italy	219	Netherlands
529	France	203	Turkey
512	Australia	196	Singapore



20 Main Institutions and its publication years

Times Cited	Authors	Keywords Plus	Author Affiliations	Countries	Publication
10390	Lin, Chih-Jen	ALGORITHMS	Natl Taiwan Univ	Taiwan	2011
	Chang, Chih-Chung	WORKING SET SELECTION			
4823	Anderson, RP	GENERALIZED ADDITIVE-MODELS	AT&T Labs Res	USA	2006
	Schapire, RE	OPERATING CHARACTERISTIC CURVES	CUNY City Coll		
	Phillips, SJ	NICHE	Amer Museum Nat Hist		
		SPATIAL PREDICTION	Princeton Univ		
		BIODIVERSITY INFORMATICS			
3476	Elith, J	HABITAT-SUITABILITY	Univ Melbourne	USA	2006
	Dudik, M	BIODIVERSITY	SUNY Stony Brook	Canada	
	Phillips, SJ	PLANT	Univ Sao Paulo	Switzerland	
	Guisan, A	SPATIAL PREDICTION	Princeton Univ	Mexico	
	Ferrier, S	DISTRIBUTION MODELS	AT&T Labs Res	Australia	
3207	Gabriel, Stacey B	GENOME	HARVARD UNIV	USA	2011
	DePristo, Mark A	ACCURACY	Brigham & Womens Hosp		
	del Angel, Guillermo	HUMAN EXOMES	Broad Inst Harvard & MIT		
	Altshuler, David	POPULATION-SCALE	Massachusetts Gen Hosp		
	Cibulskis, Kristian	QUALITY SCORES			
2738	Belkin, M	DATA REPRESENTATION	Univ Chicago	USA	2003
	Niyogi, P				
2591	Jones, M	VISUAL-ATTENTION	Mitsubishi Elect Res Labs	USA	2001
	Viola, P				
1941	Wunsch, D	NEURAL-NETWORKS	Univ Missouri	USA	2005
	Xu, R	COMPONENT ANALYSIS			
		K-MEANS ALGORITHM			
		PATTERN-RECOGNITION			
		HIDDEN MARKOV-MODELS			

A total of fifteen publications account for 49.54% of all citations.

Text mining techniques allow us to apply text classification to solve the categorization problems of a discipline

keywords defined by the authors themselves: the terms which better fit to their article

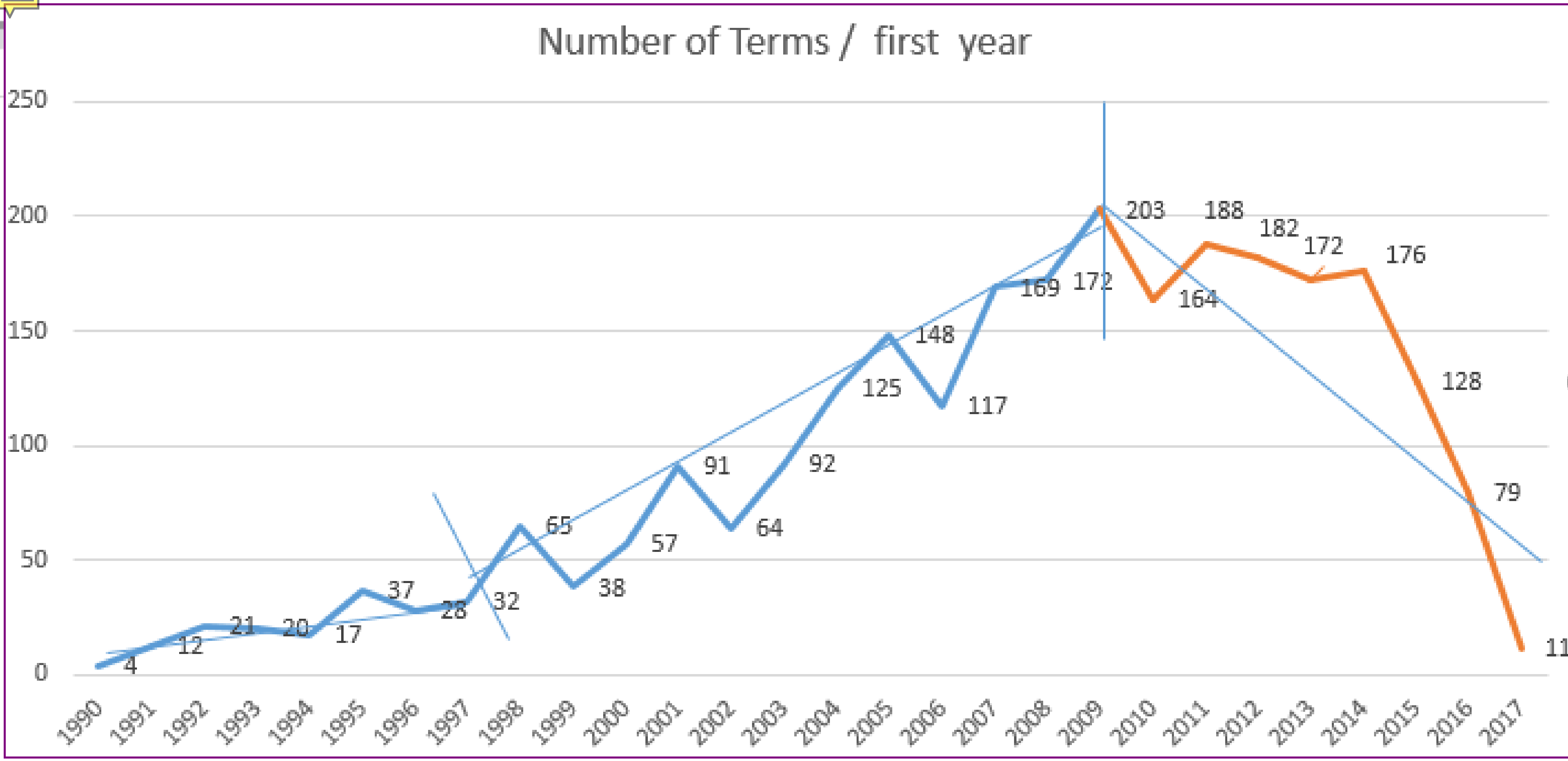
Keywords extracted from phrases identified **in the title and abstract fields**

Cleaned using fuzzy filters

22181 terms are available

discarding the terms which frequency of appearance is **less than 3**

2612 Field characterization terms



1990 (4)

machine learning. [1 of 4236]
expert system [1 of 46]
knowledge acquisition [1 of 38]
KNOWLEDGE BASE [1 of 14]

1997 (32)

data mining [2 of 515]
intelligent manufacturing systems [2 of 4]
regression [1 of 72]
graph theory [1 of 14]
fuzzy clustering [1 of 12]
Fuzzy system [1 of 10]
Neural Net [1 of 9]
applications of machine learning [1 of 7]
Semantics models [1 of 4]
Bayesian classification [1 of 3]

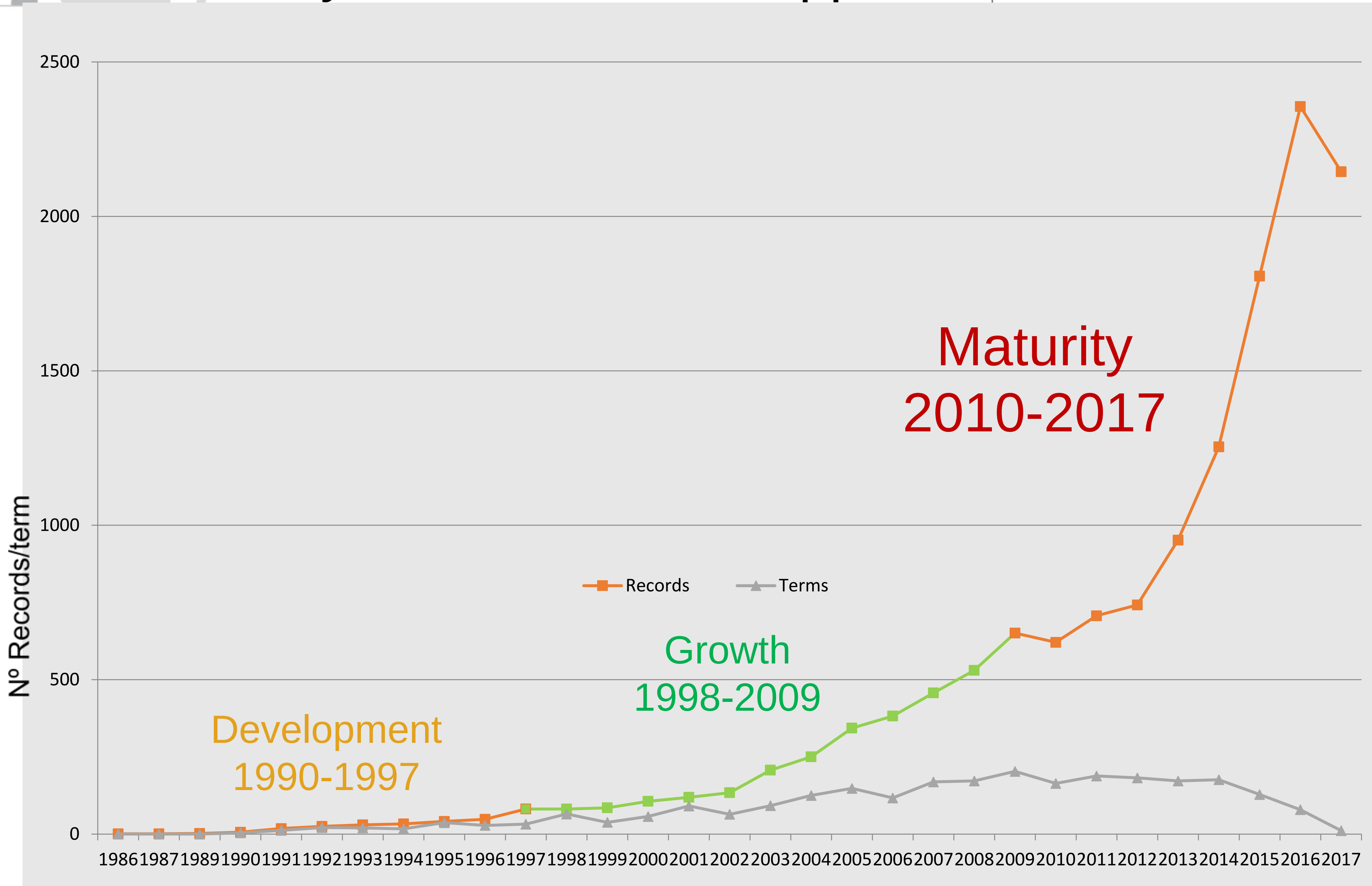
2009 (203)

Data extraction [3 of 6]
Instance selection [2 of 12]
k-means algorithm [2 of 6]
Social media [2 of 33]
fuzzy set theory [2 of 5]
CUDA [2 of 7]
Natural language processing (NLP) [2 of 5]
Facial expression [2 of 11]
Facial Recognition [2 of 6]

2017 (11)

Landslide [5 of 5]
Precision medicine [5 of 5]
Age prediction [3 of 3]
crowd-sourcing [3 of 3]
Dempster-Shafer theory [3 of 3]
DNN [3 of 3]
radiogenomics [3 of 3]
RNN [3 of 3]
self-optimizing [3 of 3]
Source Code Metrics [3 of 3]
Time series data [3 of 3]

First year that the term appeared

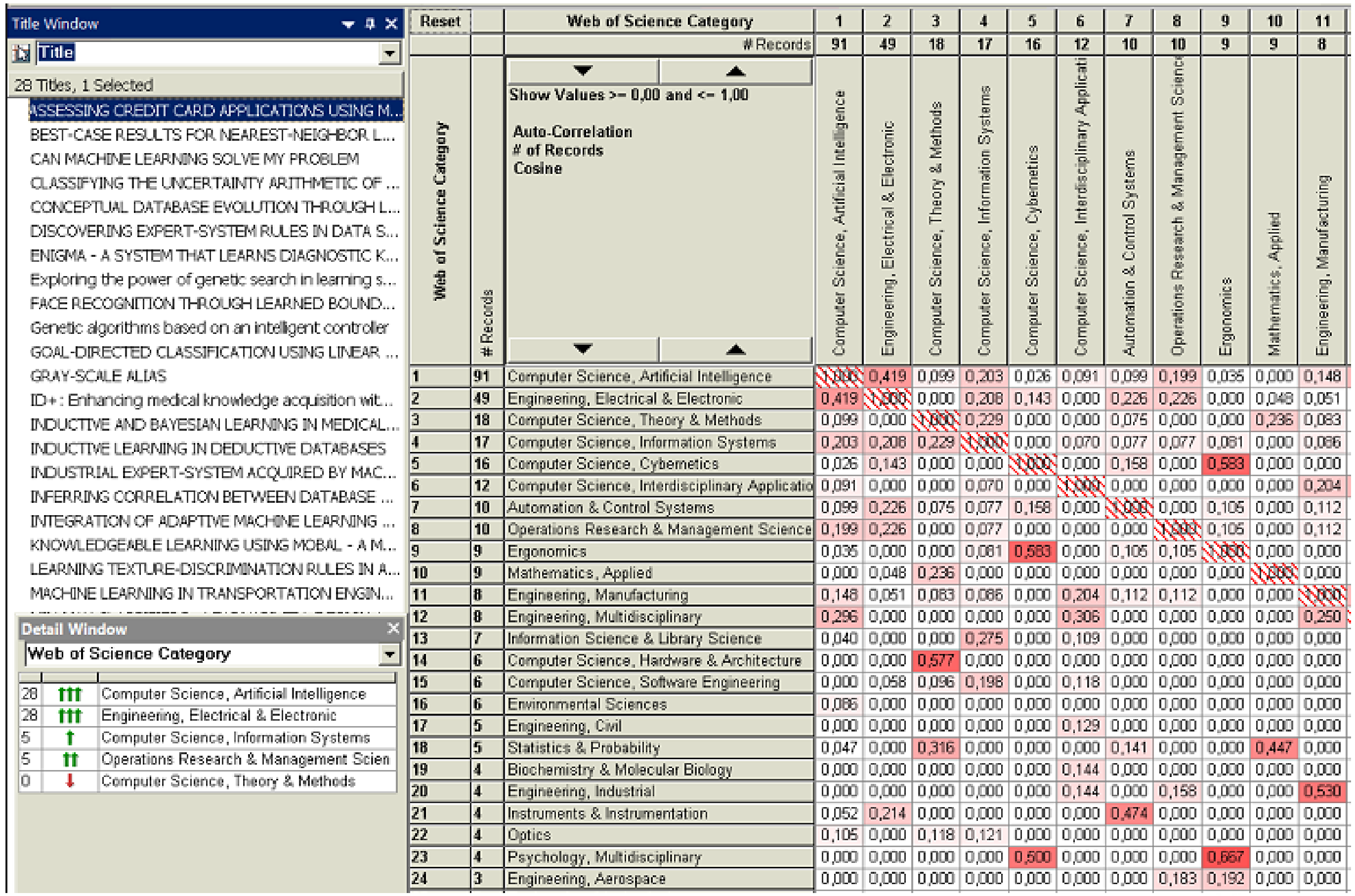


Number of new Author Keywords any year versus the number of records of that year.

Once the subdata are created, they are studied using network theory.

The aim is to understand how science has been developed. We analyse **the scientific fields present in data by exploiting the categories assigned by the Web of Science (WOS) to publications** (papers, proceedings or book chapters) of Leydersdoff (2013).

All books and journals included in the Web of Science Core Collection, the leading provider of scientific and technological publications, which includes references to leading scientific publications in any discipline of knowledge since 1945, **are assigned at least one of the 242 subject categories** predefined by Clarivate Analytics. This makes it possible to determine the scientific classification of the document



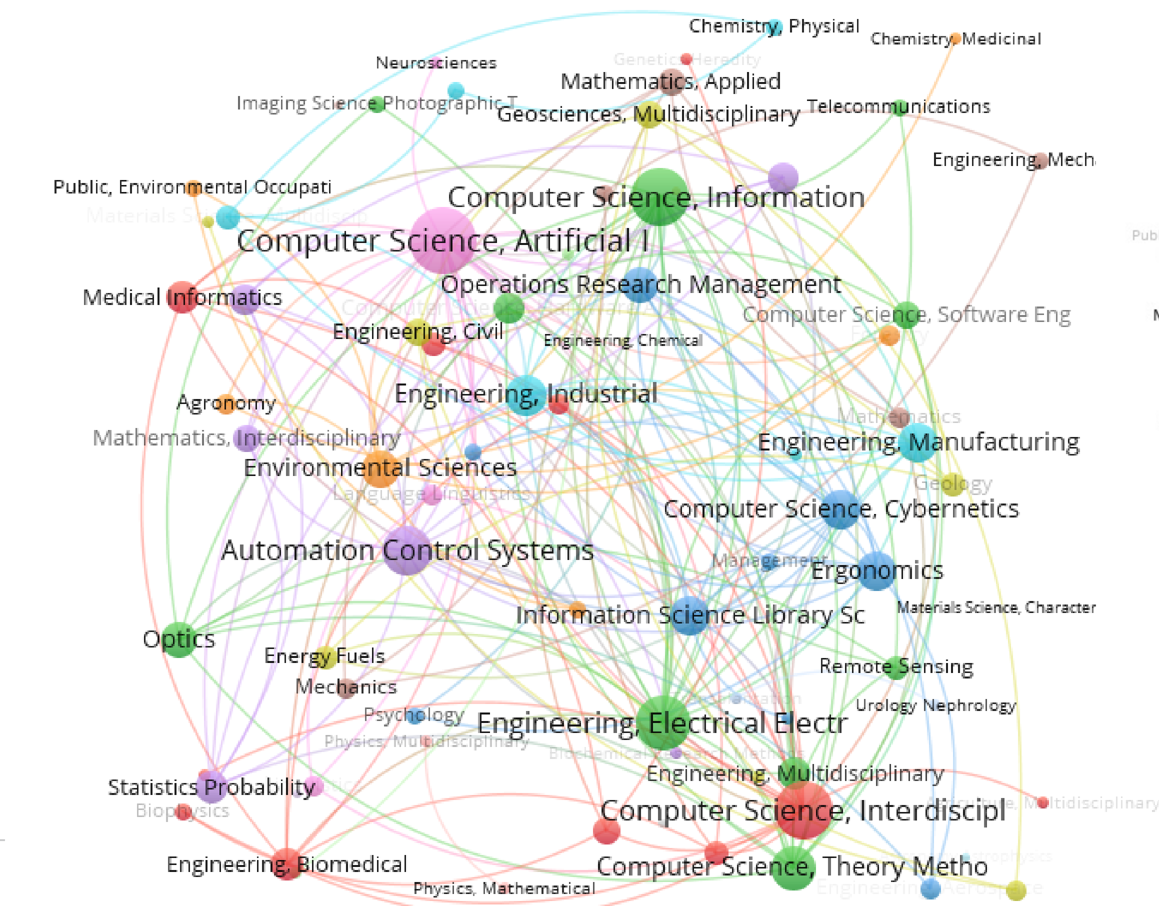
Web of Science Category	# Records		1 Computer Science, Artificial Intelligence (91)	2 Engineering, Electrical & Electronic (49)	3 Computer Science, Theory & Methods (18)	4 Computer Science, Information Systems (17)	5 Computer Science, Cybernetics (16)	6 Computer Science, Interdisciplinary Applicati (12)	7 Automation & Control Systems (10)	8 Operations Research & Management Scienc (10)	9 Ergonomics (9)	10 Mathematics, Applied (9)	11 Engineering, Manufacturing (8)
1	91	Computer Science, Artificial Intelligence	1,000	0,419	0,099	0,203	0,026	0,091	0,099	0,199	0,035	0,000	0,148
2	49	Engineering, Electrical & Electronic	0,419	1,000	0,000	0,208	0,143	0,000	0,226	0,226	0,000	0,048	0,051
3	18	Computer Science, Theory & Methods	0,099	0,000	1,000	0,229	0,000	0,000	0,075	0,000	0,000	0,236	0,083
4	17	Computer Science, Information Systems	0,203	0,208	0,229	1,000	0,000	0,070	0,077	0,077	0,081	0,000	0,086
5	16	Computer Science, Cybernetics	0,026	0,143	0,000	0,000	1,000	0,000	0,158	0,000	0,583	0,000	0,000
6	12	Computer Science, Interdisciplinary Applicatio	0,091	0,000	0,000	0,070	0,000	1,000	0,000	0,000	0,000	0,000	0,204
7	10	Automation & Control Systems	0,099	0,226	0,075	0,077	0,158	0,000	1,000	0,000	0,105	0,000	0,112
8	10	Operations Research & Management Science	0,199	0,226	0,000	0,077	0,000	0,000	0,000	1,000	0,105	0,000	0,112
9	9	Ergonomics	0,035	0,000	0,000	0,081	0,583	0,000	0,105	0,105	1,000	0,000	0,000
10	9	Mathematics, Applied	0,000	0,048	0,236	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000
11	8	Engineering, Manufacturing	0,148	0,051	0,083	0,086	0,000	0,204	0,112	0,112	0,000	0,000	1,000
12	8	Engineering, Multidisciplinary	0,296	0,000	0,000	0,000	0,000	0,306	0,000	0,000	0,000	0,000	0,250
13	7	Information Science & Library Science	0,040	0,000	0,000	0,275	0,000	0,109	0,000	0,000	0,000	0,000	0,000
14	6	Computer Science, Hardware & Architecture	0,000	0,000	0,577	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	6	Computer Science, Software Engineering	0,000	0,058	0,096	0,198	0,000	0,118	0,000	0,000	0,000	0,000	0,000
16	6	Environmental Sciences	0,086	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
17	5	Engineering, Civil	0,000	0,000	0,000	0,000	0,000	0,129	0,000	0,000	0,000	0,000	0,000
18	5	Statistics & Probability	0,047	0,000	0,316	0,000	0,000	0,000	0,141	0,000	0,000	0,447	0,000
19	4	Biochemistry & Molecular Biology	0,000	0,000	0,000	0,000	0,000	0,144	0,000	0,000	0,000	0,000	0,000
20	4	Engineering, Industrial	0,000	0,000	0,000	0,000	0,000	0,144	0,000	0,158	0,000	0,000	0,530
21	4	Instruments & Instrumentation	0,052	0,214	0,000	0,000	0,000	0,000	0,474	0,000	0,000	0,000	0,000
22	4	Optics	0,105	0,000	0,118	0,121	0,000	0,000	0,000	0,000	0,000	0,000	0,000
23	4	Psychology, Multidisciplinary	0,000	0,000	0,000	0,000	0,500	0,000	0,000	0,000	0,667	0,000	0,000
24	3	Engineering, Aerospace	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,183	0,192	0,000	0,000

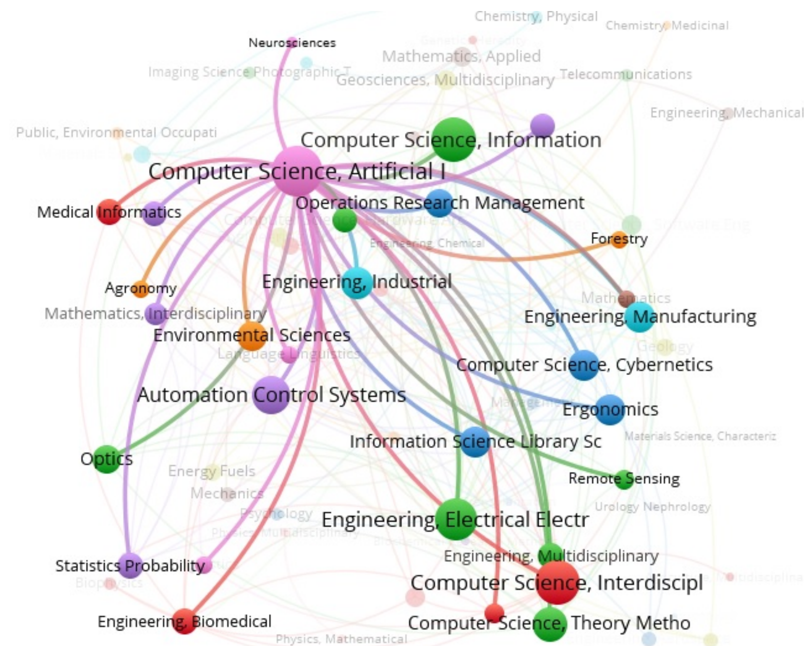
A Network composed of **nodes, WOS categories**, which are connected by edges. **The strength of the line represents the number of records in the line**

1990-1997,

This is a relatively low number and, the main collaborations are carried out between the same scientific field; Computer Sciences Artificial, C.S. information, C. S. interdisciplinary, CS Cybernetics, although there are tenuous connections with Information Sciences, Automation Control systems and Electrical Engineering

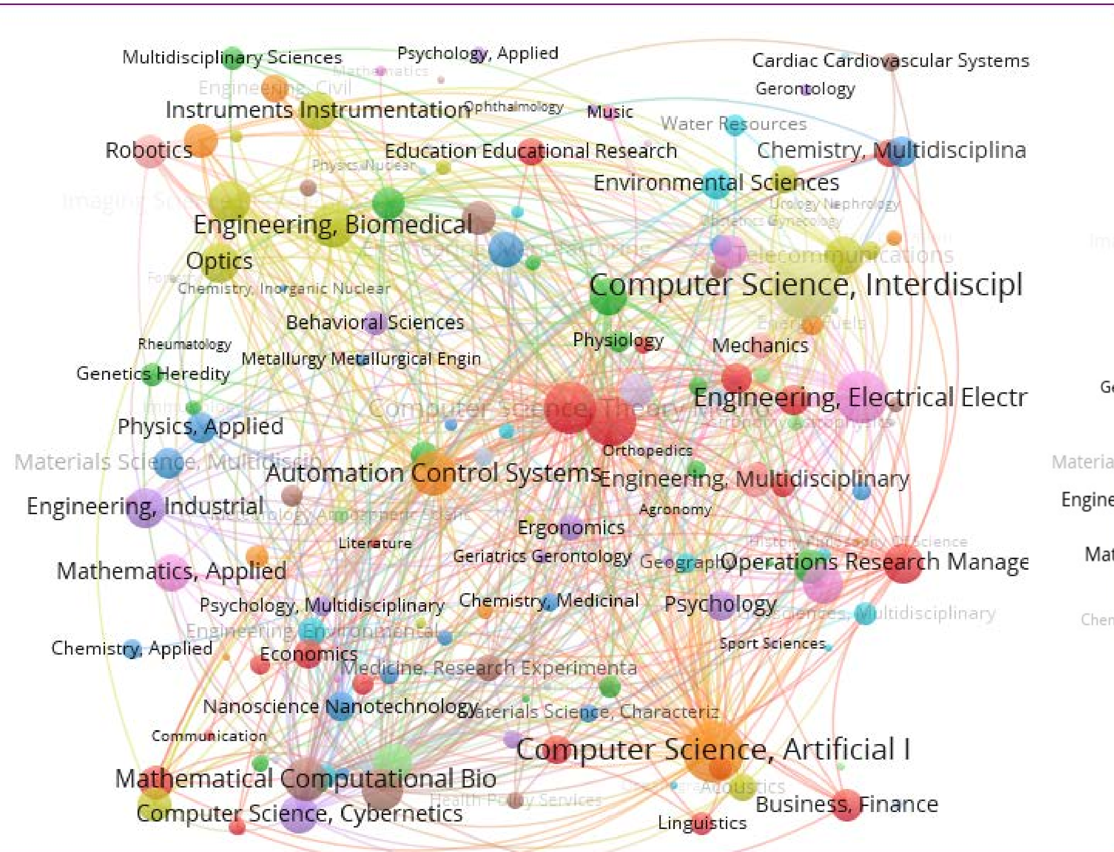
1990-1997	
Records	282
Research areas	47
Web of science categories	78
Nodes	78
Edges	181

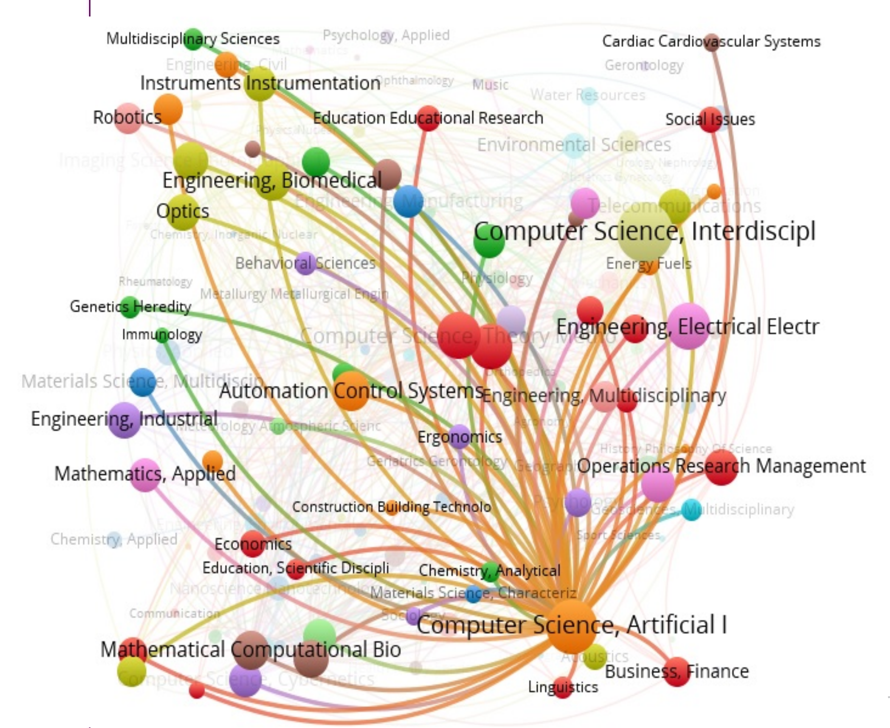




1998-2009, network grows and doubles the **number of nodes**, WOS Categories , **reaching 162**. The records are **3345**, so that more connections and **edges are generated (826)**. Computer Sciences Artificial Intelligence is connected to Medical, Bioinformatics, Biotechnology, Imaging Science photographic, Business Finance, Neuroscience, Biochemical Research methods,… We can define it as a GROWTH STAGE. Keywords at this time include terms such as: supervised learning; Bayesian decision theory; parametric, semi-parametric, and nonparametric methods; multivariate analysis; hidden Markov models; reinforcement learning; kernel machines; graphical models; Bayesian estimation; and statistical testing.

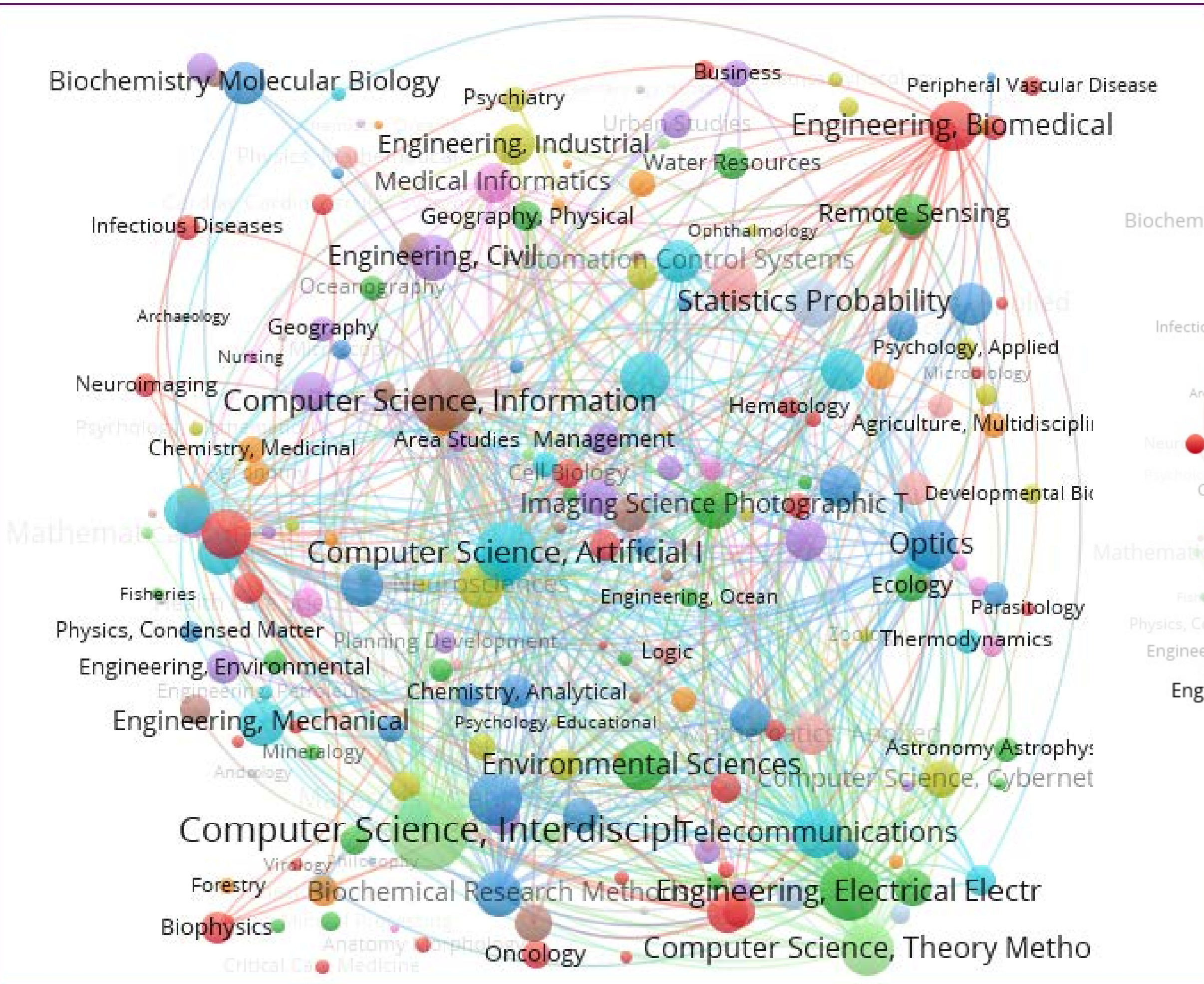
1998-2009	
Records	3345
Research areas	103
Web of science categories	162
Nodes	162
Edges	826

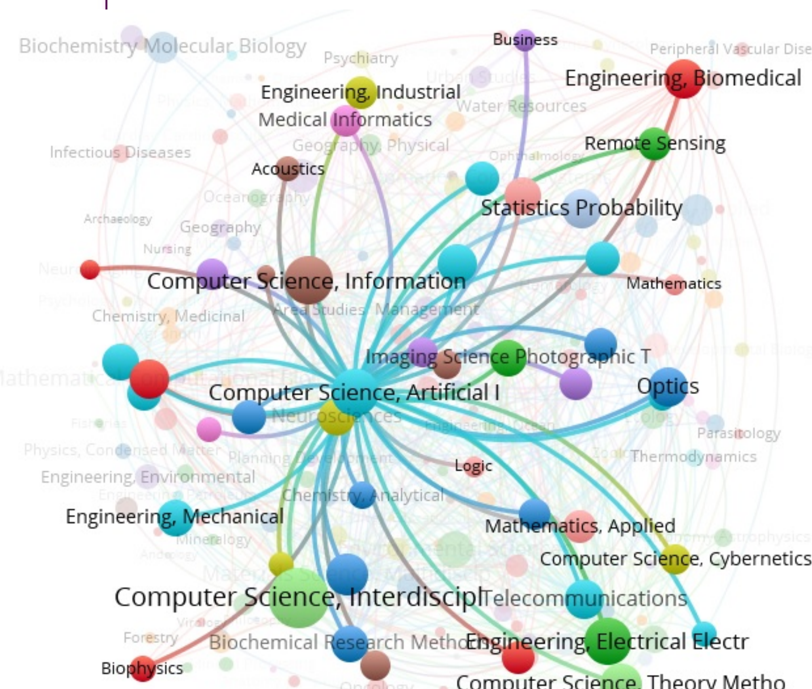




2010-2017, the network has become too extensive. **10584 records** are generated and **215 nodes and 1120 edges.** The area expands to almost all WOS categories (215/242). Areas of major applicability such as Industrial aplications: Engineering Industrial, Engineering Biomedical, Engineering electrical, Mathematics applied, Engineering Mechanical…

2010-2017	
Records	10584
Research areas	135
Web of science categories	215
Nodes	215
Edges	1120





Conclusions & Future work in Technology

The application of text mining techniques combined with **visualizations allows us to understand and interpret the evolution of a scientific discipline.**

Machine learning was born in the heart of Computer Sciences as a subdiscipline of Artificial Intelligence and has few links with other areas. From 1998 to 2009 it began to grow and branch out, connecting other areas of CS. From 2010 to 2017 we can see how the CS category branches out, goes beyond its own scope and expands into areas of applied techniques.

In the future, WOS data will be combined with patent databases and **the flows generated through the non-patent literature** collected in the industrial property registers will be analyzed.

Evolution and scientific visualization of Machine learning field

I.P. Rosa María Río-Belver
Rosamaria.rio@ehu.eus

TECHNOLOGY FORESIGHT MANAGEMENT
TFM research group
Industrial Organization and Management Engineering Department
University of the Basque Country (UPV/EHU)

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
CAMPUS OF INTERNATIONAL EXCELLENCE