Chapter 7

Technology Roadmapping of Emerging Technologies: Scientometrics and Time Series Approach

Iñaki Bildosola, Rosamaría Río-Bélver,
Gaizka Garechana and Enara Zarrabeitia

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76675

Abstract

The present work is framed within tech mining and technology forecasting fields. It proposes an approach which combines a set of quantitative methods to completely describe an emerging technology, based on science, technology & innovation data. These methods are scientometrics, with which a customized and clean database is generated; hierarchical clustering to generate the ontology of the technology; principal component analysis, which is used to identify the main sub-technologies; time series analysis to quantitatively analyze the evolution of the technology, as well as future development; and technology roadmapping to integrate all the generated information in a single visual element. The results can be regarded as inputs for competitive technical intelligence activities, as they provide information about the past evolution of the technology, as well as potential future fields of application. The practical application of the approach, to BD technology, yields outcomes that allow conclusions to be drawn, such as how competitive intelligence, query processing and internet of things sub-technologies have been dominating the basic technology during the initial evolution, and how competitive intelligence and data communications systems will do so in the short-term future.

Keywords: technology roadmapping, technology forecasting, time series analysis, emerging technologies, scientometrics, big data

1. Introduction

This work aims to contribute to the fields of tech mining and technology forecasting (TF), based on science, technology & innovation (ST&I) data, from a quantitative methodological point of view. Tech mining aims to generate Competitive Technical Intelligence (CTI) using bibliometric and text mining (TM) software for analyses of ST&I information resources [1].

Meanwhile, TF can be generically defined as a prediction of the future characteristics of useful machines, procedures, or techniques [2]. The interrelation of both fields is proved by the fact that TF studies in companies are often called CTI [3].

Both activities (CTI and TF) are crucial for current enterprises, since they address organizational and cultural barriers to adopt and harness the potential of strategic emerging technologies. In fact, literature suggests that this is even more important for SMEs, since they are slow adopters of technology, often purchasing long after release and regularly dealing with technology handed down from other companies [4]. If a company, especially medium or small, does not succeed in the early adoption of an emerging technology, it can be irremediably surpassed by those competitors who did know how to adopt it correctly. Additionally, the TF field also includes more social and diffuse measurements. For example, governments use national foresight studies to assess the course and impact of technological change for the purposes of effecting public policy [3], and some studies are also used as an awareness-raising tool, alerting industrialists to opportunities emerging in S&T or alerting researchers to the social or commercial significance and potential of their work [5].

Within this framework, the importance of correctly structuring the ST&I information for a consistent analysis of a given technology should be underscored, as it facilitates the elicitation of meaningful implications by reducing the dimensions of original data and eliminating noise that normally exists in multivariate data [6]. Accordingly, any attempt to understand the main characteristics of a technology and to discover its future evolution based on ST&I data should go through three phases: the application of scientometrics in order to structure and prepare the data related to it; the use of TM techniques, making it possible to go beyond processing the content of the data and transforming it into information; exploit the generated information to forecast the future evolution of the technology by means of TF techniques.

Based on the above, the present work proposes an approach which makes use of tech mining and TF techniques for describing an emerging technology in full. Its application to a specific field or technology brings out information that can be regarded as inputs for CTI activities. It provides the structure of the technology, the dominating subfields throughout its evolution and the potential dominating concepts of short-term future. Besides, all the information is condensed and structured in a technology roadmap (TRM), which allows a complete depiction of the technology in a single visual item.

The work is divided as follows. Section two introduces the background of the work, paying attention to similar efforts that can be found in literature. Section three describes the proposed approach, going into the detail of the techniques on which is structured and their combination. Section four is used to apply the approach to a specific technology: big data (BD). Finally, in section five the applicability and validity of the approach is discussed and the future lines of work are described.

2. Background

The interconnection among CTI, TF and TRM activities is identified by means of the abundance of reference literature. In the 90s, Porter et al. proposed a method, called technology

opportunities analysis (TOA), which used ST&I data and bibliometrics with the purpose of identifying and assessing the implications of emerging scientific areas and new research technologies [7]. Following this path, Lee and Jeong used bibliometric data, co-word analysis, to generate a strategic diagram to be used for the analysis of the development trends of a specific technology domain [8]. Similarly, Lee et al. proposed a new TRM methodology to increase roadmapping effectiveness to support effective decision-making in new product and technology planning processes. The data source was patents and the method was founded on keyword-based product–technology maps, from which objective and quantitative information can be derived [9].

Latest efforts in this field are focused on the integration of more complex statistical methods and (semi)automatization proposals. In this regard, works can be found such as that proposed by Zhang et al. [10], in which a TRM composing method is described where data inputs are raw science textual data sources. The method seeks to identify macro-trends for R&D decision makers and is primarily based on a clustering-based topic identification model, a multiple science data sources integration model, and a semi-automated fuzzy set-based TRM composing model with expert aid. With similar goals, Joung and Kim propose technical keyword-based analysis of patents to monitor emerging technologies [11]. The approach includes the automatic selection of keywords and the identification of the relatedness among them. This task is based on the analysis of a technical keyword-context matrix, which is obtained by means of text-mining tools and techniques.

However, when it comes to introduce a consistent forecasting method based on ST&I data, there is a lack of time series analysis (TSA) methods. In terms of statistical methods, the most common approach for forecasting the future evolution of a technology based on bibliometric data is growth curve analysis (see [12] for further discussion). When it comes to combine scientometrics and TF, the inclusion of specific time series models is hardly encountered within the reference literature (see for example [13, 14]). What is more, the time series commonly take the frequency of generic items, such as patents or articles, as indicators without going down to a lower level, such as keywords, which provide richer information about the technology or field that is being analyzed. This kind of strategy is roughly chosen by Park and Jun [15] within the patent analysis field. Here, time series regression and clustering techniques are combined to construct a technological trend model of identified clusters, and that furthermore, these clusters are described by means of top keywords.

The following section describes the proposed approach, which is based on the combination of methods and techniques discussed here, in an attempt to identify an optimal combination of the most representative ones.

3. Research approach

As previously stated, the present approach combines a set of methods which belong to tech mining and technology forecasting fields. Namely:

- Scientometrics: to retrieve scientific publications related to an emerging technology and structure a customized database of the corresponding records.

- Text mining: to structure and clean the text of the records and to generate time series based on the analysis of the content.
- Hierarchical clustering: to uncover the sub-technology-based structure of the technology.
- Principal component analysis (PCA): to identify the fields of greatest research activity within the technology.
- Time series modeling and forecasting: to specify appropriate models for obtained time series and to obtain forecasts of the short-term development of the research activity related to the technology.
- Technology roadmapping: to merge all the information in a single visual item.

All the methods are interrelated, in the sense that the results of the application for some represent the input for others. All the methods described below are repeated twice in the full application of the approach. The first round analyzes the research related to the basic technology of the field that is being studied; whereas the second round is focused on the applications of it. This fact impacts directly on the first task, the retrieval of research publications. The data sources for this task are multidisciplinary online databases, whose online search tools are used to perform the query and set the required Boolean conditions. Thus, making use of a scientometrics approach, when it comes to retrieve data related to basic technology, terms such as 'based on…', 'application of…', 'using…' etc., have to be avoided; and only those research areas that are directly related to the technology should be included in the query. Conversely, when it comes to the applications, those terms are not restricted in the query and the research fields should be those in which the technology is presented as an application to improve features such as performance or efficiency. The objective fields of those publications are the title, abstract, publication date and keywords.

The data set is then processed by means of TM in order to clean and structure it. Those records which lack title, abstract, publication date or keywords are removed. Natural language processing (NLP) is applied to titles and abstracts to obtain meaningful words and phrases, and these terms are combined with the keywords in order to obtain a single list of significant terms, sorted by frequency of appearance. This list is subsequently treated with fuzzy logic to group all those terms which have equivalent meanings but are not written in exactly the same way into a single term. This task falls within the text summarization field and is largely used when it comes to condense large text data (see [16] for more discussion).

The obtained terms are the base to identify the structure of the technology research. They represent the hot topics and, by means of clustering techniques, the relationships between them can be identified. Thus, the application of a hierarchical clustering method to this data will provide the vertical structure of the technology in which the main fields of research, as well as the most important subfields, can be identified.

Once a static picture of the technology is obtained, it is time to analyze the dynamics, i.e. the evolution. First of all, main sub-technologies have to be identified, as the evolution of the technology as a whole will be based on the evolution of its most important sub-technologies. To do so, PCA is applied to the list of terms generated in the previous step. PCA is a basic

method within factor analysis, which is a statistical approach that can be used to analyze interrelationships among a large number of variables, and to explain these variables in terms of their common underlying dimensions (factors or components) [17]. In the present case, it yields a number of components which are characterized by means of a vector of terms. These terms are grouped within the same component because they appear frequently together within the publications, and PCA identifies this fact. Thus, these components can be treated as sub-technologies, and the terms included in them as the main topics of within those sub-technologies (see [18] for PCA applications in text mining).

The evolution of the sub-technologies is subsequently obtained by means of time series. The generation of these series starts by splitting the previously obtained list of significant terms into months. This task is made possible because publication date of all the records is available and to which record each term belongs is also known. Thus, this split produces a set of sub-lists, each corresponding to each month of the analyzed time-range. Then a counting process is applied to generate the time series of each sub-technology. For example, if the vector of terms corresponding to sub-technology_1 is composed for three terms (term_1, term_2 and term_3), and these terms occur 2, 4 and 3 times respectively in the list of terms of a specific month, the value of the time series for that point in time is the sum of those frequencies: 9. This value is called the frequency of related terms (FRT), and represents the y-axis of the time series. If this counting process is repeated for all the months of the sample, a time series representing the evolution of each sub-technology is generated. This task is of utmost importance, as the time series is used as proxy for the intensity and trend of the activity related to a specific sub-technology.

In order to perform a consistent analysis of the evolution and forecasting, the time series has to be modeled. There is a range of models within the TSA field, and depending on the nature of the series, the simplest possible model that fits the data correctly and fulfills the objectives properly should be selected. In the case of the present work, as an initial approach, a linear time trend model (LTTM) [19] has been selected to model the last 3 years of the series, with which the trend of the series is consistently identified.

Finally, all the information previously generated is integrated into a TRM. The x axis is the temporal axis, defined by the time-range of the analysis. Whereas the y axis has two main layers: technology and application, each being completed with the information from each round of application of the approach, as described in the first task. These two vertical layers are in turn divided into sub-layers, which are directly the components of the first row of the vertical structure, obtained by means of hierarchical clustering. Once we have the TRM structured, it is filled year by year with those top terms contained in the list that comes from the text summarization task. In addition, these terms are grouped within each sub-technology, based on the corresponding vector of terms. Finally, there is room for short-term future, which will be completed with those terms that represent ascending sub-technologies. Logically, the ascending, maintained or decreasing nature is directly obtained from the time series modeling.

All these items are therefore integrated into a single visual element, full of information, the TRM. By means of this, the application of the approach aims to provide a mechanism to help experts forecast S&T developments within a specific area; or raise awareness among

practitioners concerning the characteristics and future potential applications and developments of emerging technologies.

4. Results and discussion

In order to test the applicability of the approach, and to analyze the outcomes obtained from its application, the whole approach was applied to a cutting edge technology, big data (BD). The definition of BD has evolved rapidly since the term was coined, which has caused some confusion. Gartner, Inc. gave a nice definition: "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" (Gartner IT Glossary (n.d.). The appearance of such a concept was driven by several facts. Among other things, the decrease in storage costs, which dropped from $14,000,000 (1980) to approximately $50 nowadays ($ per terabyte); the number of nodes a company might have, which have gone from 1(1969) to 1 billion hosts; and bandwidth costs, which was approximately $1200 in 1998 to the current $5 ($per Mbps) [20]. Thus, it is accepted that BD technology falls within the fields of computer science and mathematics, although it has been developed and applied in a myriad of fields, as we will see in the results of the approach.

All the tasks were applied interlaced, and partial and final outcomes were obtained. First of all, scientific publications were retrieved from the Web of Science (WOS) and Scopus databases. In order to establish the data time-range, the authors took into account what is considered as the "starting point" of BD technology research, a special issue of Nature on Big Data, in which it is distinguished from information and data science [21]. However, in order to considerate only those years in which the amount of publications was enough to analyze it from a time series point of view, the time-range was established in the range 2012–2016. The conditions imposed for the retrieving of the articles were based on similar works, in which was concluded that combining title and author keywords turned out to be the most relevant indicator in identifying related research on Big Data [22]. Thus, the term "Big Data" had to appear within the title and keywords. In the case of basic technology publications, only those within computer science and mathematics fields were allowed and those publications that contain the following terms were excluded: overview, review, based on big data, big data based, using big data, and big data application. A total of 6425 records were imported (WOS: 2740, SCOPUS: 3685). With regard to retrieving publications related to the applications of the technology, which is analyzed separately, the aforementioned excluded terms were permitted (save 'review' and 'overview'), and the allowed fields were all but computer sciences and mathematics. In this case, a total of 6864 records were imported (WOS: 3272, SCOPUS: 3592).

All the records were imported and merged in VantagePoint software (www.thevantagepoint.com). All the duplications and those records which lacked title, abstract, publication date or keywords were removed, finally obtaining a cleaned database of 5334 records for basic technology and 5991 for applications. NLP was then applied to titles and abstracts with which a set of terms was obtained. This allowed those concepts discussed within these fields to be identified. These terms were combined with those belonging to the keywords field in order to obtain

a complete set of descriptors. At the end of the task, a list of 20,5010 terms was obtained for basic technology and 29,573 terms for applications. These terms were processed by means of fuzzy matching/grouping equal terms in a single item; as a result the list was reduced to 18,434 and 26,905 respectively.

Once the lists were generated, hierarchical clustering was applied to obtain the structure of the technology. To carry out this task R software was used, as it offers various algorithms to perform this clustering process. For the present work, Agnes package [23] with Ward clustering method was selected, which has been used in a wide range of work related to term grouping. It should be noted that the clustering process needs a distance-matrix as an input, and to do so it is necessary to generate the co-occurrence matrix of the terms, which is available in VantagePoint. This matrix describes how often each term appears jointly with each of the rest of the terms, and this is the basis for the clustering task. That obtained is directly the ontology of BD technology, in which the vertical structure can be identified. This information can be found in **Figure 1** in the case of basic technology and **Figure 2** in the case of applications. Regarding the content of the ontologies, the main difference between the structures of both should be stressed. In the case of technology there are four clear main sub-fields, which represent the most important areas of research in BD: distributed systems, data mining, machine learning and privacy. Whereas in the case of application of BD, this first line is much more varied, and eight main subfields can be found: machine learning, business intelligence, cloud computing, distributed storage, internet of things, web-based big data and e-healthcare. This is justified by the fact that BD is applied in countless fields. The hierarchical clustering shows this feature by generating a first line of the ontology with multiple subfields. A further analysis provides a deeper insight of the structure, in which various levels and more specific fields of research can be identified.

The application of the approach follows with the identification of the main sub-technologies and their evolution, by means of PCA analysis. This task is carried out in VantagePoint, which contains PCA functionality. The list of terms was once again used as an input, however, in this

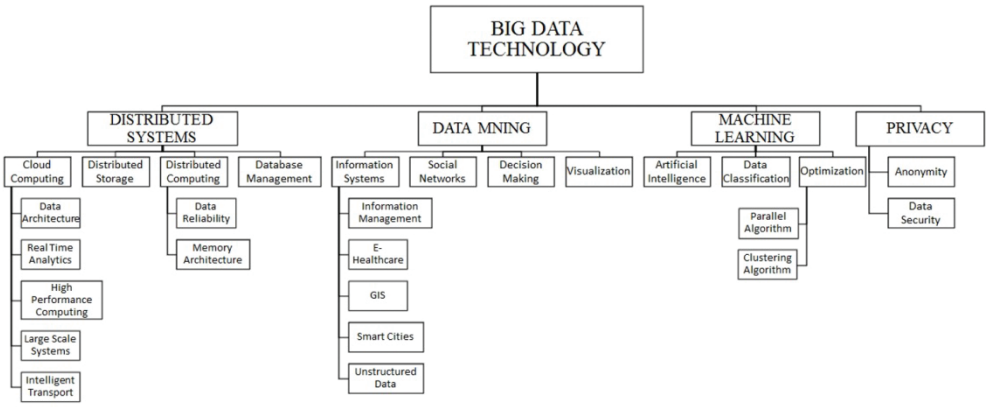

Figure 1. Big data technology ontology.

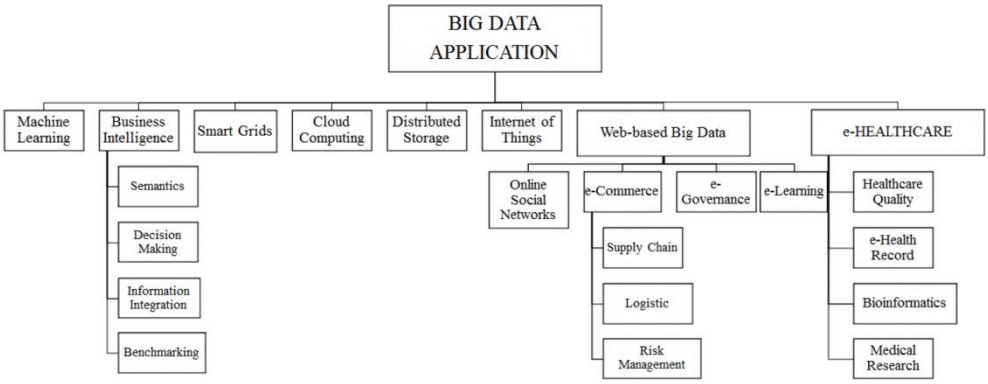

Figure 2. Big data application ontology.

Memory architecture	**Competitive intelligence**	**Learning Systems**	**Data privacy**	**Query processing**
Memory architecture Parallel architectures Program processors Parallel processing Data storage equipment Digital storage Computer hardware Network architecture Distributed storage Multiprocessing systems	Competitive intelligence Decision support system Business intelligent Decision support Decision making Management science Competition Information systems Competitive advantage Business Process	Learning systems Artificial intelligence Learning algorithms Machine learning Machine learning techniques Neural Network Deep learning Classification of information PCA Forecast	Data privacy Security of data Privacy Data security and privacy Privacy protection Privacy preserving Cryptography Privacy and security Mobile security Secure big data	Query processing Query language Query optimizer search engine Database System Computational linguistics Expert System Engines Information management Data integrity
Healthcare	**Data communication systems**	**Knowledge based systems**	**Internet of things**	**Data visualization**
Healthcare Medical computing Healthcare Hospitals Health Diagnosis Diseases Information science Medical images Data analytics	Data communication systems Data stream Stream big data Stream Computing Real time Data transfer Forestry Graphic methods Data handling	Knowledge based systems Knowledge base Semantic Web Ontology Semantic Natural language processing systems Information retrieval Extract information Knowledge extraction Knowledge management	Internet of things Internet Data reduction Data analysis Commerce Embedded systems Data acquisition Electronic commerce Cyber physical system Smart city	Data visualization Visualization Flow visualization Interactive visualization Big data visual Human computer interaction Visual analytics User interface Decision making Decision making process

Table 1. Big data basic technology top 10 components.

Internet of things	**Disaster prevention**	**Bioinformatics**	**Processing frameworks**	**Visual data**
Internet of things	Disaster	Bioinformatics	Processing	Visual data
Cyber physical systems	Disaster prevention	Biomedical	frameworks	Visuality
Embedded system	disaster management	engineering	Spark	Smart visual data
Industrial revolution	Emergency services	Biometrics	Map Reduce	Flow visualization
Network layers	Risk management	Alzheimer's disease	Computing	Three dimensional
Industry 4.0	Emergency management	Genetics	frameworks	computer graphics
Distributed computer	Online social network	Neuroimaging	Map Reduce	Information
systems	Risk perception	Genome	Hadoop	visualization
Ubiquitous computing	Social media	Biology	Open systems	Visual analytics
Manufacture	Data flow	Age	Information	Information system
Wireless		workflow	analysis	Big data visualization
telecommunication			Cluster	Data integrity
			computing	
			Open source	
			software	
Social big data	**Smart power grids**	**Machine learning**	**Energy efficiency**	**Traffic control**
Social network	Smart power grids	Machine learning	Energy efficient	Intelligent system
Natural language	Electric power	Artificial intelligence	Hardware	Traffic control
processing systems	distribution	Learning algorithms	Network	Intelligent transport
Online social network	Electric utilities	Natural language	architecture	system
Natural language	Electric power systems	processing	Energy	Traffic congestion
processing	Condition monitoring	Learning systems	conservation	Advanced technology
Machine learning	Electric power system	Online social	Computer	Motor transportation
Twitter	control	network	architecture	Vehicle
Sentiment analysis	Operation and	Classification of	Memory	Transportation
Recommender system	maintenance	information	architecture	Smart traffic control
Online learning	Data Processing Electric	Knowledge	System	Sustainable
Search engine	load forecasting	management	architecture	development
	Monitoring	Recommender	Energy	
	Electric power utilization	system	utilization	
		Forecast	Ecology	
			observatory	

Table 2. Big data application top 10 components.

case all the variables (terms) were grouped in components, and sorted by importance. Each component is represented by a vector of terms, which identifies the underlying topic. **Table 1** shows the main components of basic technology, interpreted as sub-technologies, and the top 10 terms for each. **Table 2** shows the same information in the case of applications. They are sorted by the explained variance, which means that the first contain more information about the complete original set of variables (terms). It should be noted that in order to keep as close as possible to the obtained quantitative results, the denomination of each component is always the corresponding first term, except in a few cases.

As shown, in the case of technology, even though the components were obtained from the content of publications directly related to basic technology research, topics which are actually applications of the technology can be identified. Once again, this is due to the characteristics of BD which, since the first research works, was already being applied to different fields. Thus, together with basic embryonic sub-technologies, such as memory architecture and data privacy, concepts like competitive intelligence or healthcare can be found, which are not strictly BD

foundational fields. As regards the components that belong to applications, logically these represent more specific fields, even though it might be another topic, the explained variance of each component is quite smaller than in the case of basic technology components. This means that the information is much more diversified, as expected when it comes to analyze the applications of a technology with the characteristics of BD. Lastly, it is worth mentioning the wealth of information contained in the vectors of each component. Consequently, by means of statistical techniques it is possible to identify such components, all of them with a high degree of homogeneity, and which show related and complementary concepts for different sub-technologies.

The utility of these components goes beyond their content, as a counting process to generate the corresponding time series - as previously described - can be applied. These series will provide complementary information, as they show both the intensity and the trend of each component, regarded as sub-technologies. As described in the approach's explanation, the y-axis values are measured in FRTs. Thus, those series with higher values represent those sub-technologies that have dominated the evolution of the technology in a given period of time. Additionally, the trends of the series provide meaningful information about how they have evolved throughout the analyzed period. Moreover, the trend for the last part of the series is valuable information allowing the future of the dominant and emerging sub-technologies to be forecast. However, whereas analysis of the FRT values can be done directly from the series, a consistent analysis of trends requires modeling, as this feature is not an observable component.

Figures 3 and **4** show the graphs of the top components (the complete set of values can be found in the Appendix). Note that the disparity in the range of values of the series prevents us from drawing all the graphs to the same scale. With regards to BD technology, the first analysis is centered on the levels of the series. In terms of absolute FRT values, attention should be paid to those components that have dominated the field throughout the years, which in this case are the sub-technologies of competitive intelligence, query processing and internet of things. The terms related to these have had a prominent presence, and therefore should be considered as key sub-technologies.

Additionally, which series started to present activity earlier in time can be analyzed. Thus, although all of them have a similar behavior, memory architecture and data visualization can be highlighted as those components that soon reached an important level of interest, within their range. These components can therefore be regarded as embryonic sub-technologies, since from the very beginning of the evolution of BD they started to have researchers and practitioners involved in their development. The same analysis for BD applications yields significant results. There is a clear dominant in terms of level values, social big data which, once activated, has values much higher than the rest. This indicates that it has attracted a lot of interest, directly related to its huge potential in a myriad of fields, ranging from marketing to customer relationship management (CRM). In terms of early starters, visual data is again one of those which started its activity earlier, together with processing frameworks. The latter, from the very beginning has been a field of interest, especially when it is approached from a benchmarking point of view, a fact confirmed by the data.

http://dx.doi.org/10.5772/intechopen.76675

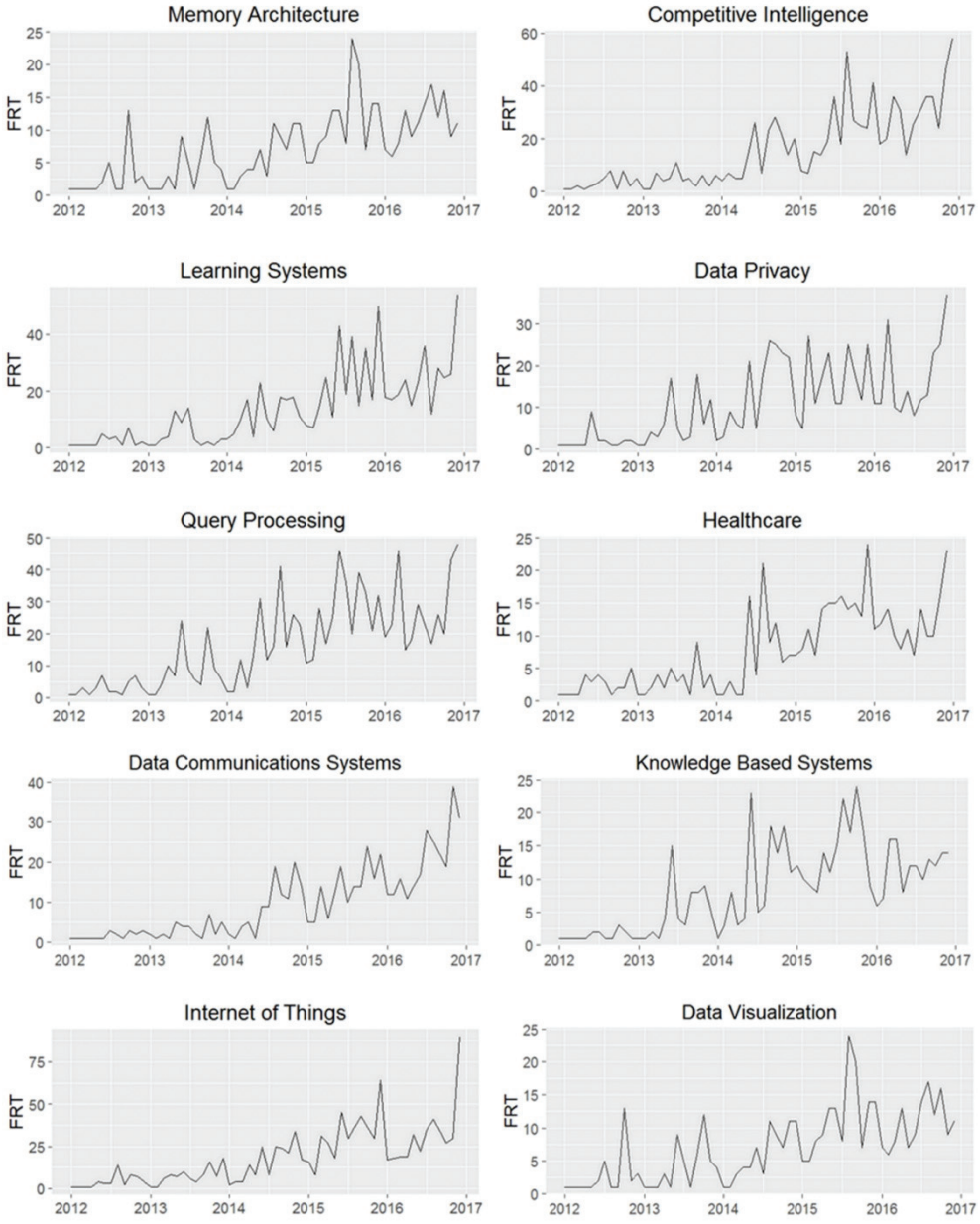

Figure 3. Time series graphs of big data basic technology top components.

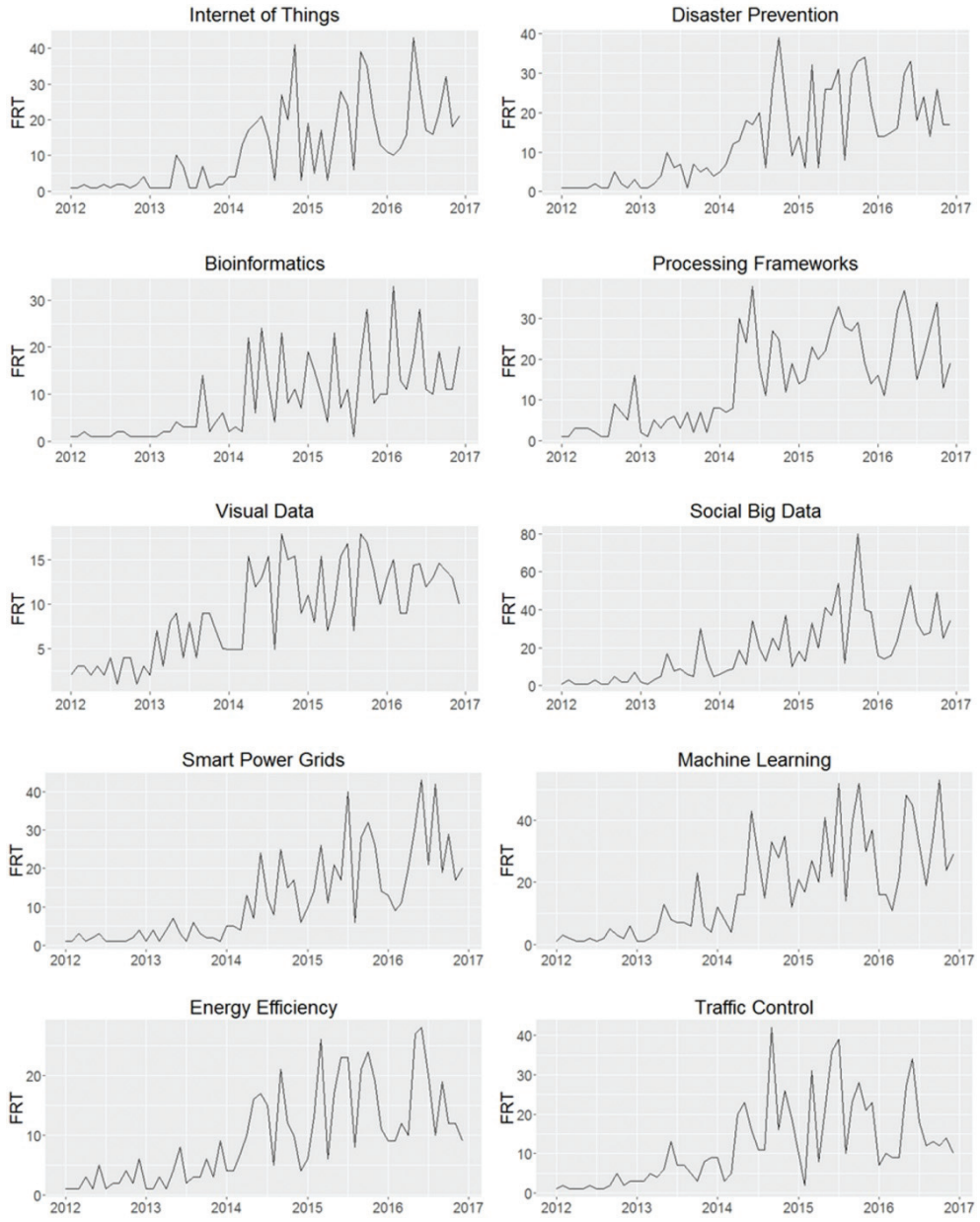

Figure 4. Time series graphs of big data applications top components.

The second part of the analysis is based on the modeling and trend identification of the series. As mentioned, the selected model was LTTM, and it was applied to the last 3 years of the series, since the goal was to identify the trend of the last phase of the evolution, in order to

Basic technology			**Applications**		
Sub-technology	R^2	Slope (p value)	Sub-technology	R^2	Slope (p value)
Memory architecture	0.35	0.032 (3.05e-04)	Internet of things	0.25	0.032 (1.09e-03)
Competitive intelligence	0.57	0.047 (1.08e-06)	Disaster prevention	0.10	0.019 (3.24e-02)
Learning Systems	0.40	0.042 (2.05e-05)	Bioinformatics	0.12	0.029 (2.23e-02)
Data privacy	0.16	0.028 (8.55e-03)	Processing frameworks	0.13	0.016 (1.94e-02)
Query processing	0.31	0.042 (2.26e-04)	Visual data	0.10	0.013 (3.71e-02)
Healthcare	0.37	0.052 (4.87e-05)	Social big data	0.27	0.031 (6.47e-04)
Data communication systems	0.52	0.059 (4.03e-07) 11	Smart power grids	0.31	0.034 (2.78e-04)
			Machine learning	0.17	0.024 (7.27e-03)
Knowledge based systems Internet of things Data Visualization	0.19	0.029 (4.14e-03)	Energy efficiency	0.10	0.019 (3.17e-02)
	0.42	0.049 (1.03e-05)	Traffic control	0.23	0.028 (3.10e-04)
	0.39	0.043 (3.07–05)			

Table 3. Parameter estimates and model validation of the main sub-technologies time series.

project it into the future. Thus, the model form is as follows: $\log(y_t) = a + bt + e_t$; where y_t represents the FRT value for a given month t = 1, 2, …, 36; a is the intercept of the model, which has no interpretation in the case of the present work; b represents the slope of the linear regression, which can be interpreted as the monthly percentage of growth of the series; and e_t represents the unexplained portion of the model, or term of error. The goodness of fit is given by the coefficient of determinations of the model (R^2), and the *p value* of the slope coefficient. If the series are observed it is clear that a linear model will not produce a good R^2 value, nevertheless, it is interesting that the *p value* of the slope coefficient is significant, since this is what is used as a proxy for the future projection. **Table 3** shows all the mentioned information for the complete set of time series.

As was expected, the R^2 values are not high enough to consider that the model is fitting the series tightly. The series present important variability and, logically, the linear model fails to follow it. However, trend identification by means of the slope value is statistically significant for all the cases at 5%. Based on these models, it is possible to analyze which sub-technologies are expected to raise more interest, and therefore develop further than others. Focusing on basic technology, the cases of data communication systems and healthcare should be noted, with a monthly percentage of increase of 5.9 and 5.2% respectively. The first is centered on issues arising from the management of communication of a huge quantity of data in the BD environment, and is apparently involving more people in its improvement. The second case, healthcare, has always been regarded as a promising field within BD technology, and the data show that it will gain importance in the short-term future. This is not the case for those that dominated the past years in terms of the series' absolute levels, memory architecture and data visualization, which with percentages of 3.5 and 3.9%, respectively have lost their dominance within the technology development.

In the case of applications, analysis of the values allows further conclusions to be drawn. Smart power grids (3.4%), internet of things (3.2%) and social big data (3.1%) are the ones with the highest trend values. All of them are growing faster than the rest of the sub-technologies and should be regarded as fields of great development. The case of social big data is even more remarkable, as it has also dominated the applications in terms of absolute

values, thus its great importance within BD applications is expected to increase. Once again, there are some sub-technologies that present lower increase values, such as energy efficiency, visual data and disaster prevention; all of them with a 10% value. Accordingly, these should be considered as fields that will gradually lose importance at the level of development and investment. In any case there is a general conclusion, which is the fact that the whole set of series present a positive trend value. This leads to a clear conclusion: BD as such is still increasing its importance among researchers and practitioners. It is still an emerging technology.

The final outcome of the approach is the TRM, in which all the previous partial results are integrated. What is more, the structuring and content of the TRM itself is conditioned by the partial results that have been obtained. The vertical structure is derived directly from the first level of the ontology in the case of the technology layer. This is not the case with the application layer, since the first line of its ontology had too many elements to sub-divide the layer based on them. Accordingly, the layer is presented without sub-divisions. The included terms are the most frequent terms, year by year, extracted from the list generated by means of the NLP task. It is required that terms exceed a certain level of frequency to be included in the TRM, and that is why more gaps appear during the initial years. In fact, it is from year 2014 when the TRM starts to be full of information, which coincides with the moment that the time series grew consistently. Furthermore, it is in the last years when the diversity of terms grows significantly, and consequently, the terms that describe more general concepts give way to others that represent more specific fields. The terms are grouped within the main sub-technologies identified above, and those terms that do not belong to any of these are placed loose. The vertical position of both the sub-technologies and loose terms, in the case of the technology layer, is based on the vertical structure of the TRM itself. Whereas for the application layer, as there is no such sub-division, placement is done by following the structure of the technology layer, as far as possible, to maintain a unified criterion throughout the TRM. Finally, the slope value of the models for each sub-technology is incorporated. The set of sub-technologies have been divided into five levels, from least to greatest slope, and have been painted accordingly with the following colors: gray; green; blue; orange; and red. Additionally, those with greater slopes have been extended further into the future, representing the probability of these being dominating fields in the short-term future. Thus, a third dimension has been added through the colors.

With regard to the content, the TRM provides a good summarization of the evolution of the technology characteristics. It can be seen how the first years show initial ideas that were developed within the different sub-technologies. For the technology layer, foundational terms such as distributed database systems in memory architecture and information management in competitive intelligence can be found. As time passes, more specific fields begin to appear, such as smart cities in internet of things and semantic web in knowledge based systems. Together with this, those topics within the fastest growing sub-technologies can be identified, which are candidates to have a strong presence in the short-term, such as business intelligence

in competitive intelligence, or diagnosis in healthcare. Similar behavior can be found in the application layer. Initially the TRM is filled with terms that refer to generalist fields, such as industry research in internet of things, MapReduce and Hadoop in processing frameworks or visual analytics in visual data. However, as you move forward in time, more specific ideas start dominating the roadmap, with examples such as industry 4.0 in internet of things and neuroimaging in bioinformatics. Finally, paying attention to emerging sub-technologies, attention should be paid to topics such as intelligent transport systems in traffic control, or sentiment analysis in social big data. All this information is presented in **Figures 5** and **6**, where the complete TRMs can be seen.

5. Conclusions and future work

The present work proposes an approach which makes use of tech mining and TF techniques for describing an emerging technology in full. The approach has been designed as a combination of quantitative methods through which various partial results are obtained, with which the technology analyzed is fully described. Within these methods, the main contribution is the idea of combining a more classical analysis based on scientometrics and common TM methods, such as clustering and text summarization; with less usual and more current methods such as PCA and especially TSA. Furthermore, technology roadmapping has been introduced to generate a final integrating element, in which all the information is aggregated. All this has permitted a fuller description of the technology, as well as a prospective exercise. To validate the applicability of the approach, it has been applied to BD technology, an emerging cutting edge technology. In that application, based on scientometrics analysis to generate a clean usable database, we have been able to apply the different methods with which the ontology of technology has been generated (hierarchical clustering method); and the main sub-technologies have been identified (PCA) (**Figures 5** and **6**).

Furthermore, a novel counting process has been presented to generate time series. These series have made it possible to understand the evolution of technology in detail. Additionally, they have been used to identify which sub-technologies have dominated the field throughout the years, and by means of a modeling process, which ones are expected to do so in the short-term future. It is at this point that it has been possible to identify that certain sub-technologies, such as memory architecture or energy efficiency, have shown limited growth in recent years, while others have accelerated their activity, with examples like competitive intelligence and smart power grids.

The results obtained come directly from the input data of the application: scientific publications. While more sophisticated results and deeper insights can be achieved on the analyzed technology, the aim has been to demonstrate that it is possible to generate such a powerful and information-filled element as the TRM by means of quantitative analysis of the data. In this sense, future lines of work should be directed towards the integration of more input data for the approach. In following with this, there are two elements that are being considered: patents

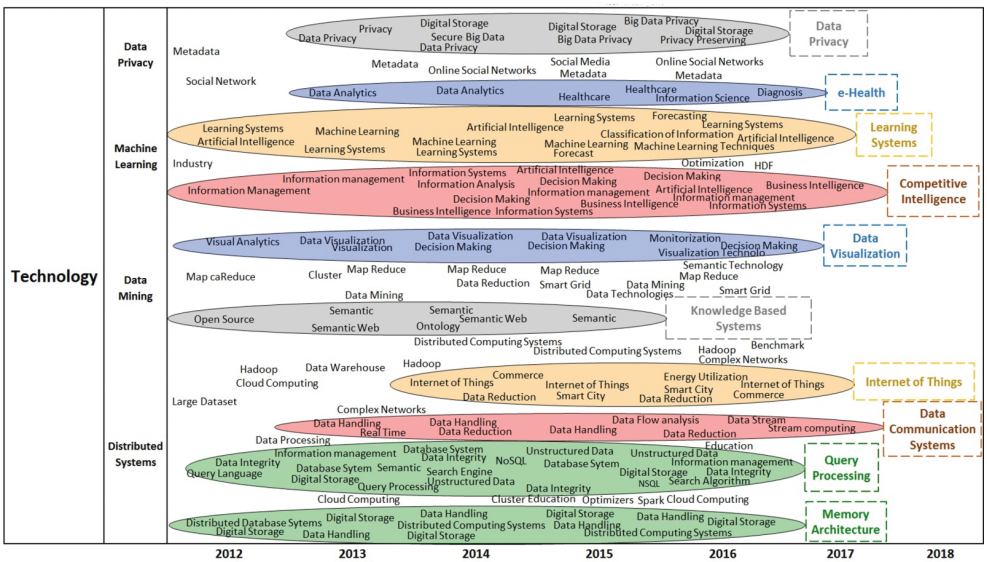

Figure 5. Technology roadmap of BD basic technology.

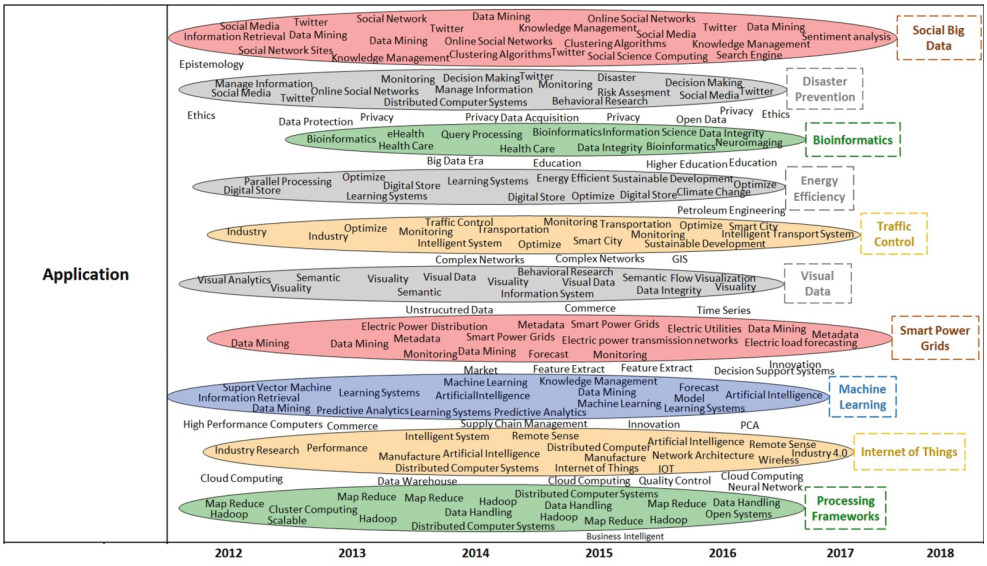

Figure 6. Technology roadmap of BD applications.

and web pages. The first will provide information about products or highly developed applications, while the webs will be used to analyze the technology at market level, based on web pages of enterprises that commercialize the technology. The same methods can be applied to these data and the results can be integrated by means of new layers in the TRM.

A. Appendix

	Memory arch.	Competitive intelligence	Learning Systems	Data privacy	Query processing	Health care	Data comm. syst.	Knowledge based syst.	Internet of Things	Data visual.
01/2012	1	1	1	1	1	1	1	1	1	1
02/2012	1	1	1	1	1	1	1	1	1	1
03/2012	1	2	1	1	3	1	1	1	2	1
04/2012	1	1	1	1	1	1	1	1	1	1
05/2012	1	2	1	1	3	4	1	1	2	1
06/2012	2	3	5	9	7	3	1	2	1	2
07/2012	5	5	3	2	2	4	3	2	6	5
08/2012	1	8	4	2	2	3	2	1	1	1
09/2012	1	1	1	1	1	1	1	1	1	1
10/2012	13	8	7	1	5	2	3	3	4	13
11/2012	2	2	1	2	7	2	2	2	3	2
12/2012	3	5	2	2	3	5	3	1	4	3
01/2013	1	1	1	1	1	1	2	1	1	1
02/2013	1	1	1	1	1	1	1	1	1	1
03/2013	1	7	3	4	4	2	2	2	2	1
04/2013	3	4	4	3	10	4	1	1	3	3
05/2013	1	5	13	6	7	2	5	4	9	1
06/2013	9	11	9	17	24	5	4	15	6	9
07/2013	5	4	14	5	9	3	4	4	6	5
08/2013	1	5	3	2	6	4	2	3	10	1
09/2013	6	2	1	3	4	1	1	8	2	6
10/2013	12	6	2	18	22	9	7	8	6	12
11/2013	5	2	1	6	9	2	2	9	3	5

	Memory arch.	Competitive intelligence	Learning Systems	Data privacy	Query processing	Health care	Data comm. syst.	Knowledge based syst.	Internet of Things	Data visual.
12/2013	4	6	3	12	6	4	5	5	12	4
01/2014	1	4	3	2	2	1	2	1	2	1
02/2014	1	7	5	3	2	1	1	3	1	1
03/2014	3	5	10	9	12	3	4	8	7	3
04/2014	4	5	17	6	3	1	5	3	7	4
05/2014	4	14	4	5	13	1	1	4	11	4
06/2014	7	26	23	21	31	16	9	23	13	7
07/2014	3	7	10	5	12	4	9	5	16	3
08/2014	11	23	6	18	16	21	19	6	14	11
09/2014	9	28	18	26	41	9	12	18	25	9
10/2014	7	22	17	25	16	12	11	14	15	7
11/2014	11	14	18	23	26	6	20	18	9	11
12/2014	11	20	11	22	23	7	14	5	14	11
01/2015	5	8	8	8	4	7	5	5	5	5
02/2015	1	7	7	5	5	3	5	2	4	1
03/2015	8	15	14	27	28	11	14	9	10	8
04/2015	9	14	25	11	17	7	6	3	13	9
05/2015	13	19	11	17	25	14	12	14	10	13
06/2015	13	36	43	23	46	15	19	11	26	13
07/2015	8	18	19	11	36	15	10	15	15	8
08/2015	24	53	39	11	20	16	14	22	39	24
09/2015	20	27	15	25	39	14	14	17	24	20
10/2015	7	25	35	18	33	15	24	24	30	7
11/2015	14	24	17	12	21	13	16	17	19	14
12/2015	14	41	50	25	32	24	22	9	29	14
01/2015	7	9	12	5	2	8	8	6	11	7
01/2016	6	12	11	5	7	12	8	7	10	6
02/2016	8	36	19	31	46	14	16	16	25	8
03/2016	13	31	24	10	9	10	11	16	24	13
04/2016	4	14	15	9	18	9	14	8	18	4

	Memory arch.	Competitive intelligence	Learning Systems	Data privacy	Query processing	Health care	Data comm. syst.	Knowledge based syst.	Internet of Things	Data visual.
05/2016	9	25	23	14	29	12	17	12	26	9
06/2016	14	30	36	8	23	13	28	12	18	14
07/2016	17	36	12	12	17	14	25	10	27	17
08/2016	12	36	28	13	26	10	22	13	31	12
09/2016	16	24	25	23	20	10	19	12	26	16
10/2016	9	46	26	25	43	16	39	14	35	9
11/2016	11	58	54	47	48	26	31	14	48	11

Table A1. Time series values of big data basic technology top components.

	Internet of Things	Disaster prevention	Bio-informatics	Processing framework	Visual data	Social big data	Smart Pow. grids	Machine learning	Energy efficiency	Traffic control
01/2012	1	1	1	1	1	1	1	1	1	1
02/2012	1	1	1	1	1	3	1	3	1	2
03/2012	2	1	2	3	2	1	3	2	1	1
04/2012	1	1	1	3	1	1	1	1	3	1
05/2012	1	1	1	3	1	1	2	1	1	1
06/2012	1	2	1	2	2	3	3	2	5	2
07/2012	1	1	1	1	1	1	1	1	1	1
08/2012	2	1	2	1	1	1	1	2	2	1
09/2012	3	5	2	9	2	5	1	5	2	2
10/2012	1	2	1	7	2	2	1	3	4	5
11/2012	2	1	1	5	1	2	2	2	2	2
12/2012	4	3	1	16	2	7	4	6	6	3
01/2013	1	1	1	2	1	2	1	1	1	1
02/2013	1	1	1	1	1	1	4	1	1	1
03/2013	1	2	2	5	2	3	1	2	3	3
04/2013	1	4	2	3	5	5	4	4	1	2
05/2013	9	10	4	5	2	17	7	13	4	4
06/2013	7	6	3	6	3	8	3	8	8	11

	Internet of Things	Disaster prevention	Bio-informatics	Processing framework	Visual data	Social big data	Smart Pow. grids	Machine learning	Energy efficiency	Traffic control
07/2013	1	7	3	3	4	9	1	7	2	5
08/2013	1	1	3	7	4	6	6	7	3	5
09/2013	8	7	14	2	1	5	3	6	3	3
10/2013	2	5	2	7	3	30	2	23	6	1
11/2013	2	6	4	2	10	14	2	6	3	6
12/2013	3	4	6	8	2	5	1	4	9	7
01/2014	6	5	2	8	1	6	5	12	4	9
02/2014	6	7	3	7	1	8	5	8	4	3
03/2014	10	12	2	8	3	9	4	4	7	5
04/2014	16	13	22	30	18	19	13	16	10	20
05/2014	19	18	6	24	11	11	7	16	16	23
06/2014	23	17	24	38	12	34	24	43	17	16
07/2014	19	20	12	18	18	20	12	28	15	11
08/2014	7	6	4	11	2	13	8	15	5	11
09/2014	27	27	23	27	17	25	42	33	21	42
10/2014	23	39	8	25	14	19	15	28	12	16
11/2014	41	23	11	12	18	37	17	35	10	26
12/2014	4	9	7	19	8	10	6	12	4	19
01/2015	24	14	19	8	10	18	10	21	6	10
02/2015	8	6	15	4	1	13	7	17	13	2
03/2015	24	32	10	23	16	33	26	27	33	31
04/2015	4	6	8	8	4	34	3	22	16	8
05/2015	17	26	23	22	9	41	21	41	17	21
06/2015	27	26	7	28	15	37	17	22	23	36
07/2015	24	36	11	33	14	54	40	52	23	39
08/2015	16	8	1	24	6	12	6	14	8	10
09/2015	39	30	18	27	17	47	28	39	21	23
10/2015	37	35	28	29	18	80	32	51	30	28
11/2015	24	34	8	19	14	40	26	30	19	21
12/2015	14	22	10	14	9	39	14	37	11	23

	Internet of Things	Disaster prevention	Bio-informatics	Processing framework	Visual data	Social big data	Smart Pow. grids	Machine learning	Energy efficiency	Traffic control
01/2016	15	14	10	10	3	16	7	16	10	7
02/2016	11	14	52	11	14	14	15	16	10	10
03/2016	11	15	13	21	8	16	19	11	12	9
04/2016	15	9	11	32	8	24	20	22	10	9
05/2016	42	30	18	37	14	39	31	48	27	27
06/2016	32	32	28	29	14	53	43	45	28	35
07/2016	19	18	11	15	11	33	21	32	20	18
08/2016	19	24	10	16	15	27	42	19	10	12
09/2016	23	14	19	27	15	28	19	35	19	13
10/2016	30	26	11	34	16	49	29	53	12	12
11/2016	18	12	12	13	12	25	17	24	12	14
12/2016	23	13	20	19	10	34	20	29	9	10

Table A2. Time series values of big data applications top components.

Author details

Iñaki Bildosola[1]*, Rosamaría Río-Bélver[2], Gaizka Garechana[1] and Enara Zarrabeitia[1]

*Address all correspondence to: inaki.bildosola@ehu.es

1 Industrial Management Department, University of the Basque Country (UPV/EHU), Faculty of Engineering in Bilbao, Bilbao, Spain

2 Industrial Management Department, University of the Basque Country (UPV/EHU), Engineering School of Vitoria-Gasteiz, Vitoria, Spain

References

[1] Zhang Y, Porter AL, Chiavetta D. Scientometrics for tech mining: An introduction. Scientometrics. 2017;**111**(3):1875-1878. DOI: 10.1007/s11192-017-2344-8

[2] Martino JP. Technological Forecasting for Decision Making. 3rd ed. McGraw-Hill, Inc.; 1993. p. 484

[3] Firat AK, Woon WL, Madnick S. Technological Forecasting–A Review. Composite Information Systems Laboratory (CISL). Cambridge, MA: Massachusetts Institute of Technology; 2008

[4] Beekhuyzen J, Hellens L, Siedle M. Cultural barriers in the adoption of emerging technologies. Las Vegas, Nevada USA: Proceedings of HCI International; 2005

[5] Coates V, Faroque M, Klavins R, Lapid K, Linstone HA, Pistorius C, Porter AL. On the future of technological forecasting. Technology Forecasting and Social Change. 2001;**67**(1): 1-17. DOI: 10.1016/S0040-1625(00)00122-0

[6] Engelsman EC, van Raan AF. A patent-based cartography of technology. Research Policy. 1994;**23**(1):1-26. DOI: 10.1016/0048-7333(94)90024-8

[7] Porter AL, Jin XY, Gilmour JE, Cunningham S. Technology opportunities analysis: Integrating technology monitoring, forecasting, and assessment with strategic planning. SRA journal. 1994;**26**(2):21-32

[8] Lee B, Jeong YI. Mapping Korea's national R&D domain of robot technology by using the co-word analysis. Scientometrics. 2008;**77**(1):3-19. DOI: 10.1007/s11192-007-1819-4

[9] Lee S, Lee S, Seol H, Park Y. Using patent information for designing new product and technology: Keyword based technology roadmapping. R&D Management. 2008;**38**(2): 169-188. DOI: 10.1111/j.1467-9310.2008.00509.x

[10] Zhang Y, Chen H, Zhang G, Zhu D, Lu J. Multiple Science Data-Oriented Technology Roadmapping Method. In: Management of Engineering and Technology (PICMET), Portland International Conference on. IEEE. 2015;2278-2287

[11] Joung J, Kim K. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technological Forecasting and Social Change. 2017;**114**:281-292. DOI: 10.1016/j.techfore.2016.08.020

[12] Martino JP. A review of selected recent advances in technological forecasting. Technological Forecasting and Social Change. 2003;**70**(8):719-733. DOI: 10.1016/S0040-1625(02) 00375-X

[13] Jun S, Uhm D. Technology forecasting using frequency time series model: Bio-technology patent analysis. Journal of Modern Mathematics and Statistics. 2010;**4**(3):101-104. DOI: 10.3923/jmmstat.2010.101.104

[14] Chen H, Zhang G, Lu J. A time-series-based technology intelligence framework by trend prediction functionality. In: Systems, man, and cybernetics (SMC), 2013 IEEE International Conference on. IEEE. 2013:3477-3482

[15] Park SS, Jun S. New technology management using time series regression and clustering. International Journal of Software Engineering and Its Applications. 2012;**6**(2):155-160

[16] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence. 2009;**1**(1):60-76. DOI: 10.4304/jetwi.1.1.60-76

[17] Hair JF, Black WC, Babin BJ, Anderson RE, Tatham R. Multivariate Data Analysis. Prentice hall: Upper Saddle River, NJ; 1998

[18] Kongthon A. A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management. Atlanta, Georgia USA: Doctoral dissertation, Georgia Institute of Technology. 2008

[19] Maddala GS, Lahiri K. Introduction to Econometrics. Vol. 2. New York: Macmillan; 1992

[20] Storey VC, Song IY. Big data technologies and management: What conceptual modeling can do. Data & Knowledge Engineering. 2017;**108**:50-67. DOI: 10.1016/j.datak.2017.01.001

[21] Huang Y, Schuehle J, Porter AL. Youtie. A systematic method to create search strategies for emerging technologies based on the web of science: Illustrated for 'big data'. Scientometrics. 2015;**105**(3):2005-2022. DOI: 10.1007/s11192-015-1638-y

[22] Hu J, Zhang Y. Discovering the interdisciplinary nature of big data research through social network analysis and visualization. Scientometrics. 2017;**112**(1):91-109. DOI: 10.1007/s11192-017-2383-1

[23] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Vol. 344. Hoboken, New Jersey USA: John Wiley & Sons; 2009. p. 342. DOI: 10.1002/9780470316801