Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study

Yi Zhang[1], Ying Huang[2, *], Alan L. Porter[3], Guangquan Zhang[1], Jie Lu[1]

[1]Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

[2]Department of Public Administration, Hunan University, Changsha, Hunan, China

[3]Technology Policy and Assessment Centre, Georgia Institute of Technology, Atlanta, USA

*Corresponding email: huangying_work@126.com;

Abstract: As one of the most impactful emerging technologies, big data analytics and its related applications are powering the development of information technologies and are significantly shaping thinking and behavior in today's interconnected world. Exploring the technological evolution of big data research is an effective way to enhance technology management and create value for research and development strategies for both government and industry. This paper uses a learning-enhanced bibliometric study to discover interactions in big data research by detecting and visualizing its evolutionary pathways. Concentrating on a set of 5840 articles derived from Web of Science covering the period between 2000 and 2015, text mining and bibliometric techniques are combined to profile the hotspots in big data research and its core constituents. A learning process is used to enhance the ability to identify the interactive relationships between topics in sequential time slices, revealing technological evolution and death. The outputs include a landscape of interactions within big data research from 2000 to 2015 with a detailed map of the evolutionary pathways of specific technologies. Empirical insights for related studies in science policy, innovation management, and entrepreneurship are also provided.

Keywords: Technological evolution; text mining; bibliometrics; big data.

Highlights:

- A systematic framework to integrate machine learning and bibliometric techniques;

- Empirical insights on discovering interactions in big data R&D from 2000 to 2015;
- A forecasting study combining quantitative and qualitative approaches;

1. Introduction

It has been several years since the big data boom led a revolution in both re-shaping thinking and behavior in all sectors of modern society (Mayer-Schönberger & Cukier 2013). The broad range of big data applications in business intelligence has been highlighted by both academic and business communities (Chen et al. 2012). Defined as "the means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed, and heterogeneous data sets[1]", big data analytics has become essential for the success of commerce (McAfee et al. 2012). Such increasing significance of big data, of course, attracts the attention of science, technology, and innovation policy (STIP) communities, and bridging big data with real-world STIP issues has been raised as an urgent task -- e.g., investigating the potential (including both positive and negative impacts) of big data and providing feasible reactions for specific industry sectors (Marx 2013; Kwon et al. 2015; Nobre & Tavares 2017). Unfortunately, despite the ambition of big data analytics on creating "big impact" (Chen et al. 2012; Wamba et al. 2015), the success of big data analytics in non-IT companies is still limited (Court 2015). Now is an opportune time to trace the evolution of big data research to discover the interactions between the techniques used in big data analytics and identify the crucial connections that have the potential to create and extend the sphere of "big impact".

Down this line, some pilot studies attempted to exploit information technologies (e.g., bibliometrics and network analysis) to uncover hidden insights behind big data analytics for supporting STIP (Zhang et al. 2016; Hu & Zhang 2017; Huang et al. 2017a). However, the involvement of artificial intelligence techniques in traditional bibliometric models is still rare, and fixed computational models might lack capability in effectively grasping insights from dynamic data streams, since a potential topic change has been observed in a collection of scientific articles with time stamps (Lu et al. 2014). Apparently, as a representative emerging technology, the instability of big data techniques further increases such difficulty for traditional approaches (Bughin et al. 2010). Aiming to address these concerns, this paper exploits a learning-based method – scientific evolutionary pathways (Zhang et al. 2017) – to discover interactions in big data research, e.g., how did big data techniques evolve and what are the relationships between leading and lagging topics. Together with a bibliometrics-based research and development (R&D) profile, we seek to identify future directions of big data research.

We draw on 5840 articles derived from the Web of Science (WoS) to conduct this study. Specifically, the R&D profile is to review the landscape of big data research by: 1) profiling the statistical dynamics and geographic distribution of related scientific articles; and 2) identifying the core constituents, i.e., the leading

[1] The definition is given in *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)* Program Solicitation NSF 12-499. More information can be found at: https://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

journals, organizations, and countries, and their competitive and collaborative relationships in this area. The scientific evolutionary pathways (SEP) create a solution to detect and visualize the technological changes in big data research from 2000 to 2015, in which a learning process is used to track the interactive relationships between topics in sequential time slices, revealing technological evolution and death by identifying predecessors and descendants of big data topics. At the end, we combine expert knowledge and our analytic results to foresee future directions of big data research in the near future and provide recommendations.

The rest of this paper is organized as follows: We review related work on bibliometrics, technological evolution, and big data-related empirical studies in Section 2, and Section 3 proposes the research framework of our study and describes the empirical data. Section 4 follows, profiling the R&D status and tracing the evolutionary pathways of big data research from 2000 to 2015. A forecasting study with recommendations on science policy and entrepreneurship is also provided. We further conclude our research, limitations, and avenues for future study in Section 5.

2. Related Work

This section reviews previous study from the following two aspects: bibliometric approaches for tracing technological evolution, and technological evolution studies on big data research.

2.1. Bibliometric approaches for tracing technological evolution

Technological evolution, defined as a theme concerning innovation and competition, has been raised by economists since the late 1970s (Abernathy & Townsend 1975; Clark 1985). Exploiting econometric approaches to investigate technological evolution in specific industry sectors and its relationships with certain phenomenon of innovation management has been a mainstream for decades (Sood & Tellis 2005; Eisenman 2013). Starting from the early 2000s or even earlier, with the rapid development of information technologies (ITs), some pioneers attempted to integrate bibliometrics and ITs (e.g., text mining) for tracing technological evolution (Zhu & Porter 2002), in which the engagement of technology roadmapping has been becoming attractive (Kostoff & Schaller 2001). When traditional technology roadmapping emphasizes the use of qualitative methodologies, e.g., Delphi/interview (Phaal et al. 2004) and TRIZ theory (Moehrle et al. 2013), Kostoff et al. (1999) headed a direction that exploits technology roadmapping as a bridge to connect bibliometric approaches with technological evolution, and the combination of qualitative and quantitative approaches becomes a trend (Zhang et al. 2013). Related approaches include keyword-based analysis (Lee et al. 2009; Zhou et al. 2014), citation analysis (Choi & Park 2009), subject-action-object-based semantic analysis (Zhang et al. 2014b), diffusion modeling (Cunningham & Kwakkel 2014), and the combination of certain above approaches (Li 2015; Guo et al. 2016). Further, such approaches have already been applied to a number of industry sectors, in particular emerging sectors, such as electric vehicles (Huang et al. 2014), dye-sensitized solar cells (Zhang et al. 2014c), energy industry (Daim & Oliver 2008; Kajikawa et al. 2008; Daim et al. 2017), and 3D printing (Huang et al. 2017b). In parallel, based on the citation linkages between scientific articles, main path analysis (Lucio-Arias & Leydesdorff 2008) and hybrid models with both citation and co-citation statistics (Small et al. 2014) are exploited to identify emerging topics and trace technological evolution as well.

2.2. Technological evolution studies on big data research

While being considered as a significant representative for emerging technologies, big data is identified as a series of technological developments in the area of data storage and data processing (Schermann et al. 2014), and the rising big data analytics, including all hardware and software techniques that can be used to analyze large-scale and complex data for real-world applications, are believed to a Pandora box by entrepreneurs (Kwon et al. 2014). A large number of surveys have been conducted by researchers from the computer science communities, addressing concerns about related techniques, issues, opportunities, and challenges for big data research (Kaisler et al. 2013; Chen & Zhang 2014; Mao et al. 2015). Such significance also holds interest from both technology management and bibliometrics communities, and, in terms of technological evolution, not too many but empirical studies on big data research could be traced as well. For example, social network analysis and science maps are exploited to profile big data research from diverse perspectives, e.g., international collaboration, semantic networks, and interdisciplinary natures (Park & Leydesdorff 2013; Singh et al. 2015; Hu & Zhang 2017); The use of technology roadmapping and technology delivery systems further emphasizes the exploration of the evolutionary pathways of big data techniques, with complementary values created by the combination of quantitative and qualitative methodologies (Zhang et al. 2016; Huang et al. 2017a).

2.3. Comparison with related work

We summarize the main features of related work that integrate bibliometrics with machine learning to investigate big data research or some other emerging technologies in Table 1. Compared with the literature, the contributions of our research are highlighted as follows:

- We follow the line 2 given in Table 1 and explore insights on understanding technological evolution through bibliometric analysis, but our methods (in particular the scientific evolutionary pathways) exploit machine learning techniques to trace the internal changes of technological evolution, while traditional bibliometric models (even with the engagement of technology roadmapping and technology delivery systems) cannot automatically and effectively handle streaming data, i.e., a fixed analytic model cannot always adapt to hidden changes occurring gradually within a sequential time line.

- Even though network analysis and science maps hold great ability in identifying the relationships between big data techniques, such relationships are relatively "blind", i.e., we can only know they relate to each other but cannot figure out what the exact relationship is. The use of scientific evolutionary pathways provides a solution to identify the predecessor and descendants of technological topics, which can help understand the evolutionary relationships among sub-technologies in a given area.

Table 1. Main features of related work

NO	*Main Feature*	*Related References*
1	Emphasizing *expert knowledge*, with the use of some quantitative approaches for data analytics: This is one of the mainstreams in the area of technology management for investigating technological evolution.	Phaal et al. (2004), Daim and Oliver (2008), and Daim et al. (2017)
2	Emphasizing the involvement of quantitative approaches (in particular *bibliometric analysis*) in gaining insights, with limited use of expert knowledge: This is an emergent area that investigates technological evolution through bibliometrics.	Choi and Park (2009), Zhou et al. (2014), Cunningham and Kwakkel (2014), and Huang et al. (2017b)

3. Methodology and Data

3.1. Methodology

We define *interactions* in big data research from two aspects: 1) the correlation between big data-related entities, e.g., the semantic relationships between technological terms, and the collaboration networks among countries and between affiliations; 2) the evolutionary relationships among big data-related techniques, e.g., for a specific topic, which forward topic is the predecessor and which afterward topics are the descendants. Under this circumstance, a research framework for our study is constructed (Figure 1), which includes a model for research and development (R&D) profile and a model for scientific evolutionary pathways (SEP).

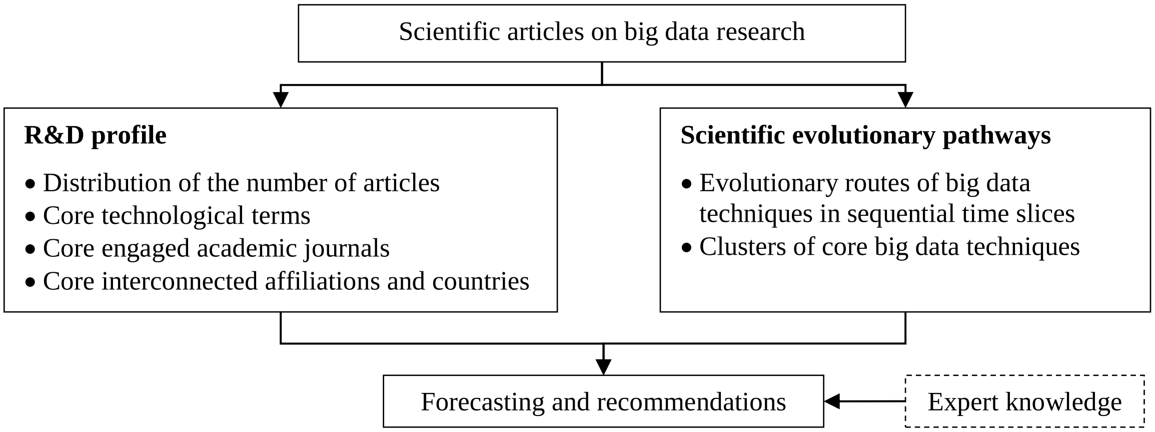
Scientific articles on big data research
R&D profile
• Distribution of the number of articles
• Core technological terms
• Core engaged academic journals
• Core interconnected affiliations and countries
Scientific evolutionary pathways
• Evolutionary routes of big data techniques in sequential time slices
• Clusters of core big data techniques
Forecasting and recommendations
Expert knowledge

Figure 1. Research framework of discovering interactions in big data research

The R&D profile model focuses on the basic statistical information of the collected scientific articles, including the number of articles, term frequency, and correlated journals, affiliations, and countries, and emphasizes the discovery of the correlations between big data-related entities by using social network analysis and certain traditional bibliometric approaches.

The SEP model was developed by Zhang et al. (2017); it specifically sheds light on discovering the detailed relationships among scientific and technological topics, in which a learning process is designed to detect potential topic changes from scientific articles in a sequential time line. Referring to the definition of emerging techniques, given by Rotolo et al. (2015), and certain empirical tests based on big data-related datasets, we observed that the radical novelty of big data techniques would result in relatively isolated technological groups, which share a low similarity with each other. Under this circumstance, we use the k-means algorithm refined by Zhang et al. (2016) to replace the hierarchical agglomerative clustering approach in the original SEP model, which would help generate a relative small number of topics, and also benefit the visual performance of the SEP model. Here are the main concepts of the SEP model.

Definition 1: a topic is a collection of articles that are mathematically represented by their centroid, which is identified as the article sharing the highest similarity with all other articles on the topic.

Definition 2: a topic is geometrically represented as a circle, and its boundary is the largest Euclidean distance between its centroid and all articles.

The stepwise algorithm in the SEP approach is described below:

Step 1 Simulate the dataset as a data stream that consists of a sequence of time slices. The k-means algorithm is used to group the articles in time slice 0 into several initial topics.

Step 2 Process the data stream in an iterative flow, i.e., a time slice is treated as one iteration, and the articles in a time slice are read one by one.

Step 3 Measure the similarity between a forthcoming article and the centroids of all existing topics, using Salton's cosine (Salton & McGill 1986), and assign it to the most similar topic.

Step 4 Calculate the Euclidean distance between the article and the centroid of its assigned topic. If the distance is within the boundary, we set the article as "normal". If it is near the boundary within a given interval, we set the article as "evolution". If it is much larger than the boundary, we set it as "novelty/noise", and its assignment within the existing topic is deleted.

Step 5 At the end of each iteration:

Step 5.1 The k-means algorithm groups the articles labeled with "evolution" and "novelty/noise" respectively. New topics containing articles labeled as "evolution" are set as the descendant of their assigned topic in Step 3. While new topics consisting of articles labeled with "novelty/noise" do not have a predecessor.

Step 5.2 The accumulated number of articles in each topic is detected, and a topic is set as "death" if the accumulation is 0 in a sequence of time slices.

Step 5.3 The similarity between all new and existing topics is measured, including dead topics. If a new topic shares the highest similarity with a topic that is not its predecessor, the new topic is combined with the old one, and the link to its predecessor is removed. If the old topic is dead, it will be resurged – this phenomena is identified as "sleeping beauties" by van Raan (2004), and our model follows the main concept of this idea.

Step 6 Recalculate the centroid and boundary of each existing topic and then return to Step 2 until the stream ends.

Generally, a traditional cluster approach applies a fixed algorithm to analyze the entire dataset, ignoring any difference resulting from the time and the content. For example, the term "data mining" closely related to database management and data warehouse decades ago, but now it is highly involved with machine learning and mostly replaced by data analytics and business analytics. Apparently, traditional models would fail to capture such change, and this ignorance will influence the accuracy of related similarity measures. The use of a learning process in this model simulates batched data in a stream, captures orderly information in an ordered queue, and modifies the algorithm in real time to adapt to possible underlying change, e.g., it modifies the centroid and boundary of a topic at the end of each iteration to improve the accuracy of classification in further iterations.

The evolutionary relationships between topics are vividly identified as predecessors and descendants, and linked topics would form a group to represent certain specific techniques or sub-areas in big data research. With the aid of science maps, we then visualize topics and their relationships as nodes and directed arcs, in which certain detailed evolutionary pathways of big data research can be traced.

Based on the analytic results (e.g., figures and tables) derived from the R&D profile and the SEP models, an expert panel, including experts whose expertise aligns with either big data research or relatively broad computer science disciplines (e.g., artificial intelligence and information systems), will be organized. Our questions concentrate on two parts: 1) whether the analytic results are reasonable? If not, how can we modify the results? 2) Based on the analytic results and their expertise, which evolutionary pathways generated by the SEP model can be an emergent direction in the near future? Is there any add-in direction? We then collect and summarize expert feedback by emails or face-to-face interviews, and forecast certain emergent directions of big data research, with expert knowledge-based recommendations for potential stakeholders, e.g., policy-makers, entrepreneurs, and academic researchers.

3.2. Data and Pre-processing

Scientific articles can be a good resource for exploring information on the research frontier of a given technological area (Zhang et al. 2015). Web of Science (WoS) is a quality-guaranteed database for this purpose. We collected 5840 articles from WoS using an updated version of the search strategy proposed by Huang et al. (2015), which is based on a combination of bibliometric techniques and expert knowledge. Details are provided in Table 2.

Table 2. Search strategy

NO	***Strategy***
#1	TS = ("Big Data" or Bigdata or "Map Reduce" or MapReduce or Hadoop or Hbase or Nosql or Newsql)
#2	TS = ((Big Near/1 Data or Huge Near/1 Data) or "Massive Data" or "Data Lake" or "Massive Information" or "Huge Information" or "Big Information" or "Large-scale Data" or Petabyte or Exabyte or Zettabyte or "Semi-Structured Data" or "Semistructured Data" or "Unstructured Data")
#3	TS = ("Cloud Comput*" or "Data Min*" or "Analytic*" or "Privacy" or "Data Manag*" or "Social Media*" or "Machine Learning" or "Social Network*" or "Security" or "Twitter*" or "Predict*" or "Stream*" or "Architect*" or "Distributed Comput*" or "Business Intelligence" or "GPU" or "Innovat*" or "GIS" or "Real-Time" or "Sensor Network*" or "Smart Grid*" or "Complex Network*" or

	"Genomics" or "Parallel Comput*" or "Support Vector Machine" or "SVM" or "Distributed" or "Scalab*" or "Time Serie*" or "Data Science" or "Informatics*" or "OLAP")
#4	#1 OR (#2 AND #3)

Note that the search strategy covered scientific articles published from January 1, 2000 to December 31, 2015.

We combined the title and abstract fields and used a function of natural language processing in VantagePoint[2] to retrieve the terms. A term clumping process (Zhang et al. 2014a) was used to remove noise and consolidate technological synonyms, and the stepwise results are given in Table 3. These 10,921 terms can be considered as the core technological terms in the big data research field.

Table 3. Stepwise results of the term clumping processing

Step	*Description*	*#Term*
0	Raw terms retrieved by the natural language processing technique	120,427
1	Removing terms starting with non-alphabetic characters, e.g., 1.5%	115,381
2	Removing meaningless and common terms, e.g., pronouns, prepositions, and conjunctions	110,362
3	Removing common terms in scientific articles, e.g., “method” and “introduction”	109,137
4	Consolidating technical synonyms in the field of computer science, e.g., “classification” and “classification analysis”	108,344
5	Consolidating terms with the same stem, e.g., the singular and plural of a noun	91,918
6	Removing single words [a], e.g., “internet” and “information”	84,949
7	Removing terms appearing in only one article	12,229
8	Consolidating technological synonyms with expert knowledge [b], e.g., “time series” was used to represent terms such as “time series forecasting,” “time series mining” and “time series economics”	10,921[c]

Note: (a) Considering the main concepts of most single words can be traced in multi-word terms and terms further enrich the meaning of single words from different perspectives, e.g., “internet” Vs. “internet of things”, and “information” Vs. “information systems”, we decided to remove single words. In addition, since we consolidated related terms into a single word (e.g., “classification” in Steps 4 and 5), these words were not removed. (b) Two authors of this paper manually reviewed the terms derived in Step 7 and, based on the list of big data techniques and technologies outlined by Manyika et al. (2011), related technological synonyms were consolidated. (c) Articles that did not contain any core technological terms were removed from the model of scientific evolutionary pathways; 5450 articles remained with coverage of 93.3%.

[2] VantagePoint is commercial software used in text mining and particularly in science, technology, and innovation text analysis. More detail can be found on their website: https://www.thevantagepoint.com/

4. Results: The Discovery of Interactions in Big Data Research and A Forecasting Study

The results of our study include the discovery of interactions in big data research (i.e., R&D profile and scientific evolutionary pathways) and a forecasting study to provide recommendations for stakeholders (e.g., policy-makers, entrepreneurs, and academic researchers).

4.1. Discovery of Interactions in Big Data Research

4.1.1. R&D Profile

The distribution of the number of articles in big data research per year is given in Figure 2. Despite the common perception that the big data boom started in the late 2000s when a number of world-leading IT companies developed architectures to handle large-scale data [e.g., MapReduce by Google in 2004 (Dean & Ghemawat 2008)], "big data" is still a new term to the public and to academia. This is validated by the few and relatively unchanged number of articles from 2000 to 2010 in Figure 2. The dramatic increase in the number of scientific articles after 2012 can be credited to the *Big Data Research and Development Initiative*[3] (Big Data Initiative) that was announced by the Obama administration. This formally raised the significance of big data research to the national strategy stage. Funding provided by the governments of the US, the EU, China, and many other countries effectively stimulated big data research.

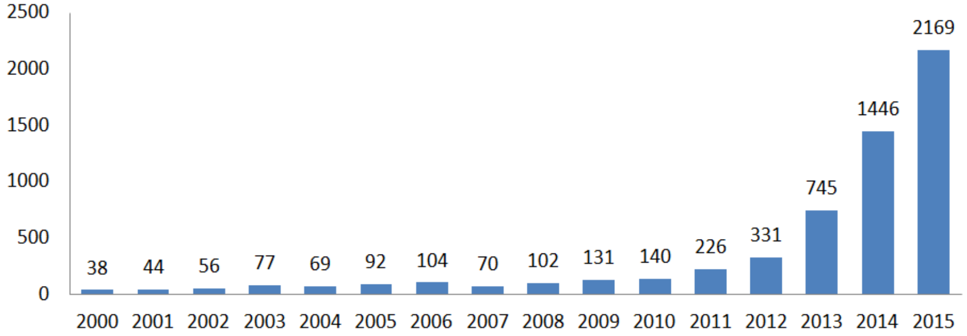

Figure 2. Distribution of the number of articles in big data research per year

The main components of big data (the Mckinsey List), summarized by the Mckinsey Global Institute in 2011 (Manyika et al. 2011), have been widely recognized by both industry and academia. Five years after the list's release it is interesting to explore questions such as: "What research has occurred over the past five years?" and "Which big data techniques are important now?" We selected 599 core technological terms with frequency no less than 10, which were identified in Step 8 of Table 3, and visualize them in Figure 3 using

[3] Information on the *Big Data Initiative* can be found at https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

VOSviewer (Waltman et al. 2010). The size of a node illustrates the weight (or, we say, importance) of the related term, and the color is painted based on VOSviewer's own algorithm for community detection.

Four clusters are clearly detected in Figure 3, i.e., cloud computing (with MapReduce, Hadoop, and related techniques), machine learning (with a broad range of artificial intelligence techniques), bioinformatics, and internet of things. Generally, there were no unexpected terms, and all terms in Figure 3 or their synonyms can be traced back to the Mckinsey List. However, it is clear that the importance and the internal content of some techniques have changed. Three examples are:

- The importance has weakened – terms relating to "A/B testing" cannot be found in our list. A/B testing is a basic approach to statistical hypothesis testing. Scientific articles might not have a strong interest in such a mature and basic technique when compared to novel and intelligent algorithms.
- The internal content has been extended – the term "artificial intelligence" has a relatively low frequency, which is surprising. However, it is clear that almost the entire cluster of machine learning and parts of the bioinformatics cluster relate to artificial intelligence, e.g., machine learning, neural network, prediction model, and natural language processing. In other words, many sub-domains of artificial intelligence have evolved into relatively mature research areas, which negatively influence the frequency of the term "artificial intelligence."
- The internal content has changed – the Mckinsey List did not consider internet of things as a big data technique but, undoubtedly, the interaction between big data and internet of things is broader and deeper than we imagined years ago. Technically, before 2010, internet of things was closely related to radio-frequency identification (RFID) techniques. Now it involves new techniques, such as sensor networks, Wi-Fi techniques, and a wide range of smart and mobile devices. From the perspective of a national R&D strategy, China identified internet of things as one of its top 5 emerging industries in 2009 (Wen 2009), and in 2014 the Obama administration highlighted internet of things in the report *Big Data: Seizing Opportunities Preserving Values*[4] (Big Data Report). [This may also have offered the first officially raised concern about data privacy in the big data age.]

[4] https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

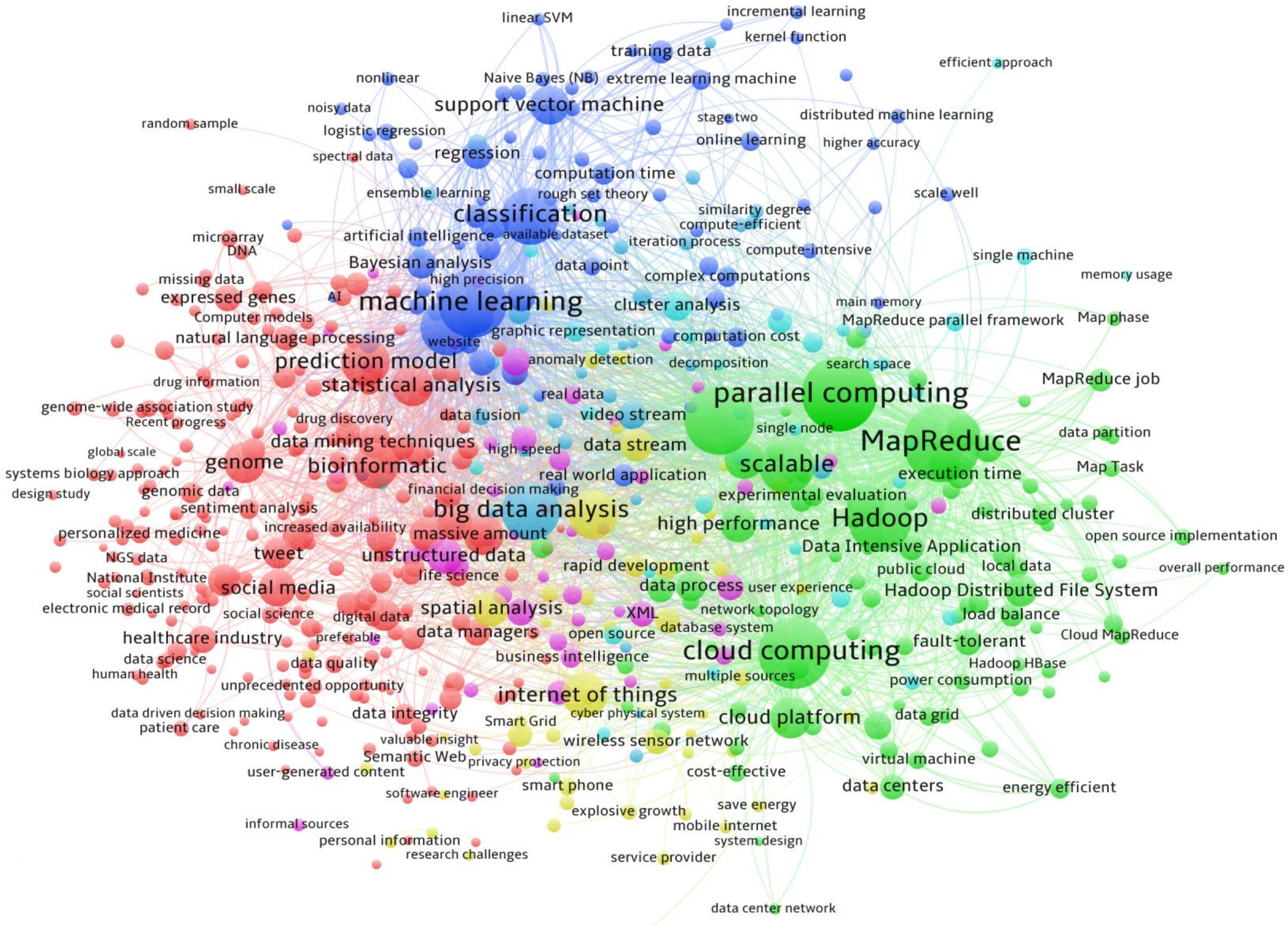

Figure 1. Term correlation map for big data research

Figure 3 provides an overview to answer the question: "What has happened in big data research?" Then, this section will focus on the core players in big data research, offering insights to questions such as:

- Which journals are holding interest in the frontier of big data research?
- Which countries are leading the global competition in big data?
- Which organizations are leading the world and how do they interact with each other?

Based on the 5840 articles we collected from WoS, we identified 1759 journals and list the 20 journals with the largest number of articles in Table 4. It is interesting that two journals related to bioinformatics are in the list, which might reflect the increasing interest in analyzing large-scale datasets in biological areas and, particularly, genomic data. Two multidisciplinary journals also attracted our attention. The special issue of

Nature entitled “Big Data” in 2008[5] and the special issue of *Science* entitled “Dealing with Data” in 2011[6] can be considered as milestones for big data research, indicating the start of the big data boom in academia. In this circumstance, the appearance of *Nature* in Table 4 demonstrates the continued interest in big data research by world-leading research communities. The fact that big data research is published in the domain of multidisciplinary sciences supports the argument that big data is an emerging technology, and it holds interest for researchers in both the natural and social sciences.

Table 4. Top 20 journals in big data research

# P [a]	***Journal***
84	Future Generation Computer Systems
79	PLoS One
69	Concurr. Comput. Pract. Exp.
62	Big Data
57	IEEE Transactions on Parallel and Distributed Systems
56	BMC Bioinformatics
56	IEEE Transactions on Knowledge and Data Engineering
51	The Journal of Supercomputing
49	International Journal of Distributed Sensor Networks
48	Cluster Computing
46	Computer
44	Bioinformatics
41	Information Sciences
36	Journal of Parallel and Distributed Computing
36	Nature
36	Neurocomputing
35	ACM SIGPLAN Notices
34	Expert Systems with Applications
34	IBM Journal of Research and Development
33	Communications of the ACM

Note: (a) The number of articles.

We also retrieved 824 journals, which were cited more than 20 times by the 5840 articles, and generated a journal citation map, as shown in Figure 4. It is interesting that the four major clusters in Figure 4 well match

[5] http://www.nature.com/news/specials/bigdata/index.html

[6] http://science.sciencemag.org/content/331/6018

with those in Figure 3. While considering the three multidisciplinary journals (i.e., Nature, Science, and PloS One) located at the center of journal citation map, some insights are identified as follows:

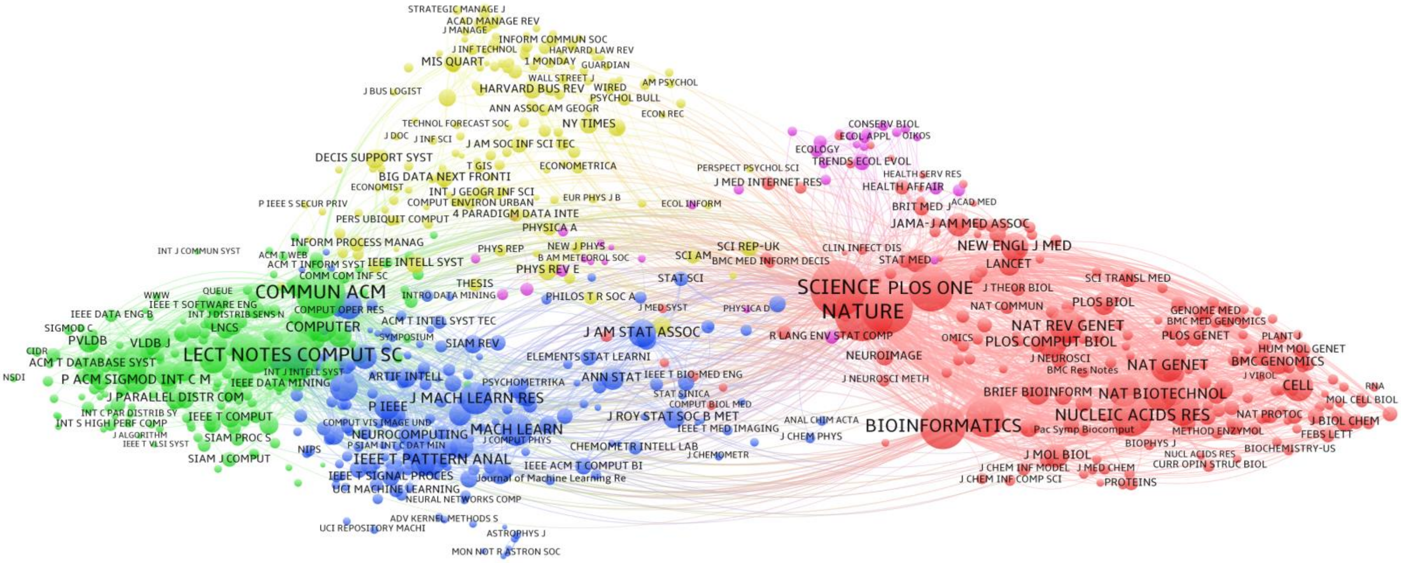

Figure 1. Journal citation map for big data research

- Bioinformatics journals (red nodes) are grouped as a large cluster, and illustrate close interactions with the three multidisciplinary journals, indicating the extensive applications of big data analytics in biological and medical fields.

- Undoubtedly, journals in the field of computer science (green and blue nodes) are the other major domain, and interestingly, *Communications of the ACM* is in the core of the cluster. One reason behind this might be this journal specifically addresses concerns to the actual applications of big data analytics, and such motivation could attract relatively broad audience from diverse research areas.

- Yellow and pink nodes act as a bridge to connect the two mentioned clusters and the three multidisciplinary journals with each other. For example, 1) journals aligning with the area of information systems (e.g., *MIS Quarterly*, *Decision Support Systems* and *Journal of the Association for Information Science and Technology*) extend the application of big data analytics to business and management domains. 2) Some physical journals (e.g., *Physical Review E* and *Physica A*) and mathematic journals (e.g., *Journal of the American Statistical Association*) seem like theoretical development for some new concepts announced in *Nature* and *Science*, and then link with further improvement or implementation in computer science journals. 3) Definitely, some top-level business and management journals and magazines are highly involved, e.g., *Harvard Business Review*, *Academy of Management Review*, and *New York Times*.

The authors of the 5840 articles are from 3807 organizations in 90 countries. Based on the number of articles in each country and the collaboration networks between the 90 countries, a country collaboration map is generated in Figure 5, in which a node indicates a country, and its size represents the number of articles in related country.

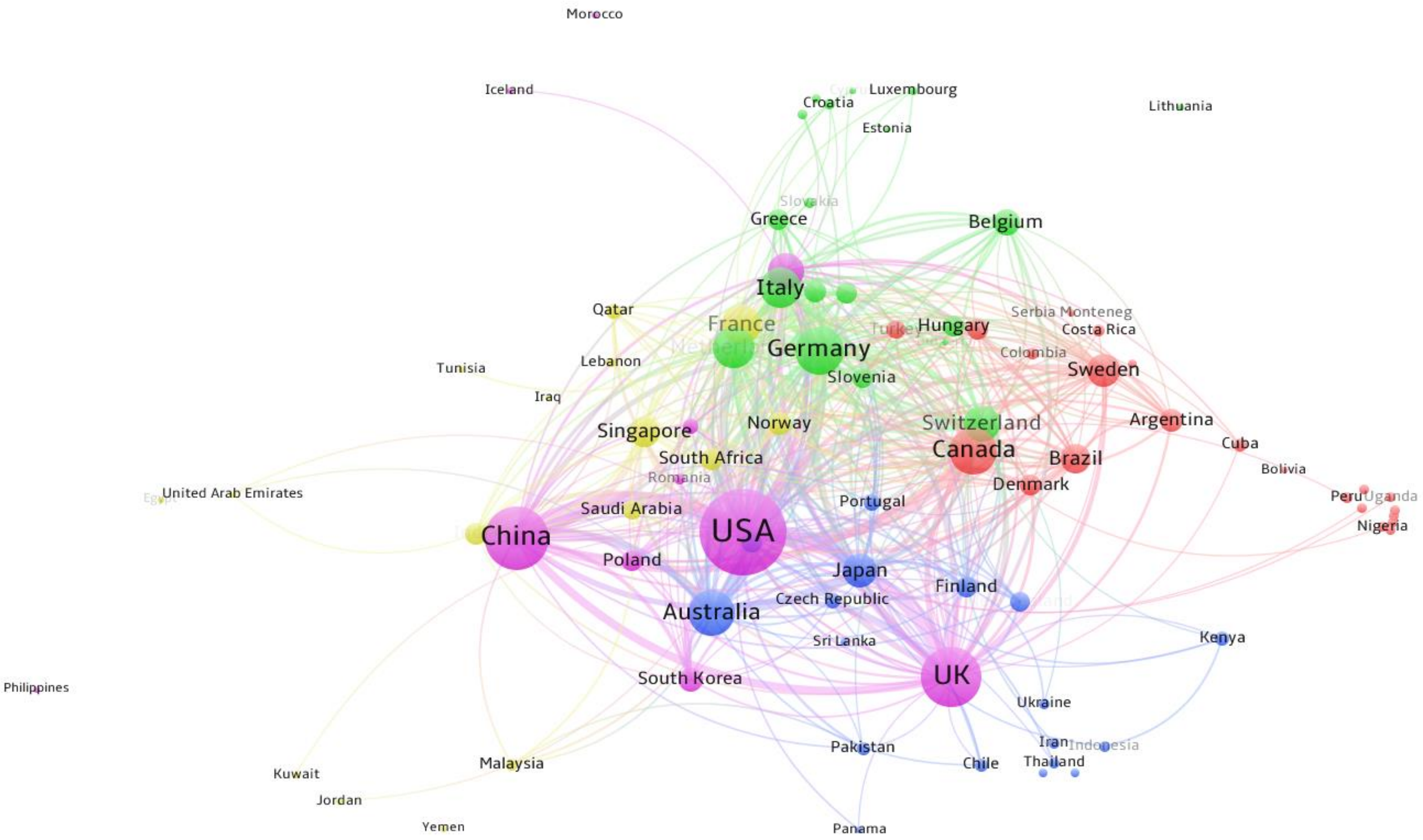

Figure 2. Country collaboration map for big data research

According to our records, the US leads the field with 42% of the articles, and China follows holding about 23%. The UK, Germany, and Canada are also relatively competitive. As shown in Figure 5, when the US, China, and the UK have established their own global collaborative networks, Germany pays more attention to European countries and Canada interestingly builds up connections with South American countries. In addition, Asian countries mostly collaborate with the US, China, or the UK, but Japan illustrates its accomplishment in global collaboration, as well as Australia.

The collaboration map of global organizations[7] in big data research (shown in Figure 6) further endorses our findings observed from Figure 5. The 3070 organizations (excluding those without any collaborative activities) cluster into four groups:

- The US group has a number of world-class universities such as Harvard, Stanford, MIT, and Carnegie Mellon. This group is located in the center of the graph and closely interacts with other groups.
- The Asian group presents its largest nodes from the Chinese Academy of Sciences (CAS) and Tsinghua University. Other important nodes are from Japan, South Korea, and Taiwan, China (e.g., Tokyo University, Tokyo Institute of Technology, Korea University, Kyung Hee University, and National Central University). Organizations in this group mostly have their own relatively isolated sub-groups, but some organizations from mainland China have constructed collaborations with other groups (e.g., CAS, Tsinghua Unversity, and Xi'an Jiaotong University).
- The European group with a large cluster on the right of the graph and a number of nodes scattered in other groups, e.g., Oxford, Leiden University, and Karolinska Institute, are located with the US group. Compared to the Asian group, European organizations have much stronger collaborations with organizations from the US group. This could be as a result of historical and cultural factors. Such collaboration leads to the European and the US groups being well-mixed in Figure 6, especially along the common boundary of the two groups. Interestingly, the University of College London is in that boundary area. It demonstrates a strong link with Harvard and Stanford, which might be a very representative example of an alliance among giants.
- The Australian group has relatively scattered nodes (e.g., University of Queensland, University of Melbourne, University of Adelaide) along the edge of the US and Asian groups and connects both. In one sense, this group belongs to the US group, but it is interesting to see them acting as a bridge prompting interactions between the two largest regions in the big data world.

[7] Most of the organizations we retrieved from the articles are, unsurprisingly, universities and academic institutions. Although the crucial role of companies, like Google, in the development of big data cannot be ignored, our study emphasizes academic research, and thus universities and academic institutions are promising candidates for study.

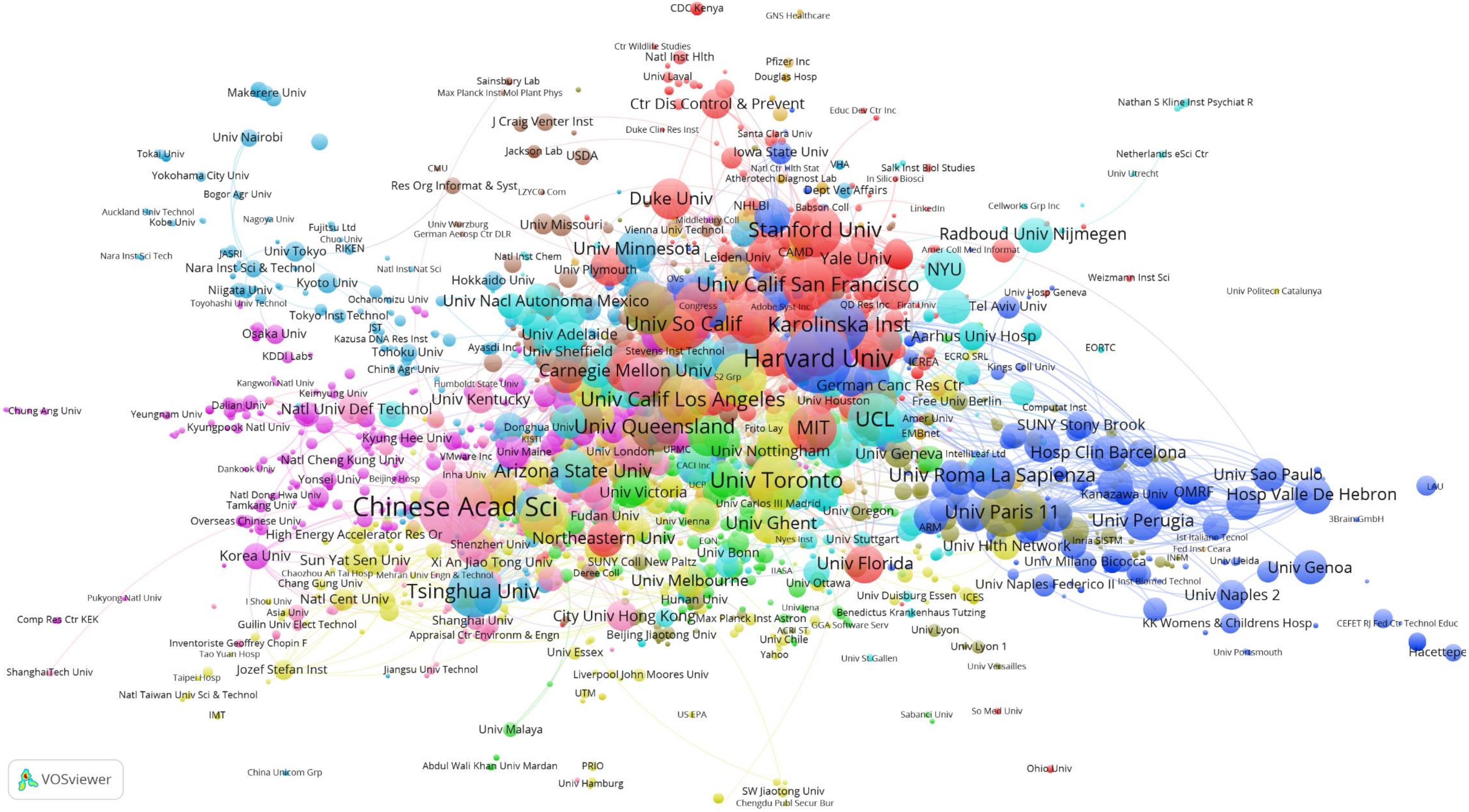

Figure 3. **Collaboration map of global organizations in big data research**

In summary, we outline certain key findings based on the results of the R&D profiling:

- Big data-related research has been conducted for several decades, but the boom started in 2010.
- The hotspots in big data analytics are concentrated on machine learning and cloud computing. MapReduce and Hadoop are still the two leading tools. The applications of big data analytics have been well involved with bioinformatics and internet of things.
- Beside *Nature* and *Science*, journals in the areas of computer science, business and management, and bioinformatics are playing active roles in publishing big data-related academic articles, while positive impacts of mathematic and physical journals can also be traced.
- With extensive collaborative networks with both Europe and Asia, the US leads big data research globally, and Europe illustrates strong competitiveness. China is demonstrating great potential in both research quantity and collaborations. However, collaboration within Asian academic institutions (especially Japan and South Korean) are somewhat limited.
- The role of world-leading universities and academic institutions in pushing big data research forward is significant, and examples of the alliance of giants occur here and there.

4.1.2. Scientific Evolutionary Pathways

The R&D profile provides an effective way to explore the insights of what happened in big data research and who the core players are. SEP concentrates on the changes to big data-related techniques, e.g., concepts, algorithms, software, services, and platforms, from 2000 to 2015.

As mentioned in Table 3, we used the 5450 articles containing the 10,921 core technological terms and set 16 time slices based on publication year. We grouped two initial topics via a k-means algorithm in time slice 0 – "statistical analysis" and "parallel processes." This decision was based on expert knowledge and the testing of a number of options for the value of k. We set the upper/lower ranges for the boundary to ±10% -- i.e., if the distance between an article and the centroid of its assigned topic was within the interval, we set the article as "evolution," and if the distance was larger than the upper range of the boundary, the article was set as "novelty/noise." In regards to the k-means algorithm grouping articles labeled with "evolution" and "novelty/noise," the strategy to decide the number of topics was: if the total number of articles waiting for the cluster analysis was less than 10, we grouped them as one topic; if the number was within the interval [10, 50] or [50, 100], we set k = 2 and k = 3, respectively; if the number was larger than 100, we set the value of k = 4. Since there were not too many articles labeled as such, k=1 or 2 were the most common occurrences.

Within the 5434 articles[8], 89 topics were detected, including the 2 initial topics. Descriptive statistics of these topics are given in Table 5, including the numbers of articles and terms, the value of term frequency inverse document frequency (TFIDF) analysis, and survival length.

Table 5. Descriptive statistics of topics

Indicator	*Min*	*Max*	*Avg.*	*S.D.* [a]
Number of articles	2	378	61.1	85.6
Number of terms	5	2143	361.8	499.5
TFIDF value [b]	0.0016	0.2390	0.0578	0.0596
Survival length [c]	1	16	3.5	3.3

Note: (a) Standard deviation; (b) The classic formula of TFIDF analysis proposed in (Salton & Buckley 1988) was used here; (c) The number of the time slices in which a topic is alive.

Based on the notion of "sleeping beauties," we identified 10 topics that died, were resurged later, and then were alive until 2015, while the remaining 79 topics stayed alive from their birth year to 2015. The resurgences can be attributed to the re-discovery of the potential of a topic. In these circumstances, "sleeping beauties" may represent emerging interests in big data research[9], especially those resurged not long ago. Detailed information about the 10 "sleeping beauties" is listed in Table 6, including the year when they were generated, the number of articles, the number of terms, the value of the TFIDF analysis, and their length of survival.

Table 6. Ten "Sleeping Beauty" Topics

Topic	*Birth*	*# P*	*# Term*	*TFIDF*	*S. L.* [a]
Statistical analysis	2000	161	922	0.136	9
Classification	2002	170	1146	0.166	9
Distributed system	2003	332	1959	0.229	7
Machine learning	2005	319	1859	0.220	4
Support vector machine	2006	54	340	0.066	10
Unstructured data	2009	26	189	0.042	8
Prediction model	2011	90	515	0.088	1
Metadata	2012	58	440	0.084	2
Video stream	2013	42	269	0.055	3
Clinical decision making	2014	42	245	0.049	4

Note: (a) Survival length.

[8] We removed two "novelty/noisy" topics comprising 16 articles since our expert panel deemed they were noisy rather than novelty topics.

[9] Considering "sleeping beauties" are topics that are identified at least twice, usually they cannot reflect noisiness. However, one exception is they can be a missing mention in a given year.

Analyses of these topics reveal a number of interesting analytic objects, crucial algorithms, and impressive applications in big data research. These insights are:

- Unstructured data is highlighted in big data research, and exploring insights from audio and video streaming is one of the basic requirements raised in the Big Data Initiative. At the same time, the definitions of metadata and unstructured data have overlapped and become broader in scope. They now cover rich information from almost all sectors of the real world, and, therefore, analyzing metadata has already become a hotspot in big data research.
- The majority of big data analytic techniques appeared long before the big data boom. Techniques and algorithms such as statistical analysis, classification, machine learning, and support vector machine are within this range. However, upgrading, optimizing, and recombining these techniques for big data have become emergent tasks. Prediction models are one such example, where big data techniques are now focused on solving the problems of both government and industry simultaneously.
- Distributed systems are not new and, as shown in Table 6, appeared as a topic in 2003 and its related research definitely started long before then. However, its importance was dramatically raised because of Hadoop, the programming framework used in distributed computing environments. A similar situation might also occur with parallel computing and MapReduce.
- Supporting medical diagnoses by analyzing large-scale medical records is currently highlighted. The project *Big Data to Knowledge* (Margolis et al. 2014), a response to the Big Data Initiative from the National Institutes of Health (NIH) in 2013, pushed this emergent task forward. As indicated in Figure 4, bioinformatics can be considered as one of the most successful applications of big data analytics.

We used the 89 topics and their relationships to create a network map to further visualize the dynamics of big data research. One node represents one topic, and the relationship between two topics (i.e., its predecessors or descendants) is represented as a directed arc. The SEP of big data research from 2000 to 2015 were generated by Gephi (Bastian et al. 2009), as shown in Figure 7. It is clear that four clusters of big data research can be identified:

- Large-scale data analytics – Starting from statistical analysis, three groups of analytic techniques are highlighted in this pathway: 1) machine learning and prediction models (including support vector machine, cluster analysis, and neural networks); 2) data warehouse and metadata (including XML and web mining); and 3) large-scale data mining and classification (including data stream processing, data visualization, time series analysis, and text mining). These analytic techniques existed long before the so-called big data age, but big data provided new opportunities, new objectives, and new problems. More importantly, attention to deep learning, an emergent area of machine learning, is rising these days, which further enhances the ability of analyzing complicated data by constructing multiple processing layers (LeCun et al. 2015). In addition, two problem-oriented research streams attracted new interest. Tweet-based analyses, including sentiment analysis and social network analysis, have become quite popular in text mining. Bioinformatics and medical diagnoses have also become oriented to real-world needs and are no longer out of reach with the help of big data analytics and related techniques.

- Cloud computing and distributed systems – Parallel processing and distributed systems appeared decades before the big data boom; however, the wide acceptance and popularization of MapReduce and Hadoop were clearly two forces pushing that boom. It is clear that these technologies introduced a disruptive revolution to traditional models of data storage, management, and processing. NoSQL databases were one of its outcomes. Yet, this pathway is relatively isolated and does not have close interactions with others.
- Internet of things – As discussed in the R&D Profile, internet of things may be independent of big data research. However, with the rapid growth of sensor-based applications, analyzing large-scale data generated from sensor networks connects internet of things and big data research. More importantly, the well-matched integration between mobile devices and social media has further prompted such interactions.
- Big data applications – besides bioinformatics, big data research can be, and has already been, applied to a broad range of real-world applications, e.g., forecasting climate change, constructing smart cities (usually bound with internet of things), assisting in decision making on social policy, and analyzing spatial problems. One highlight is the concern for data privacy. In fact, these debates existed before the big data boom, and the emphasis in the Big Data Report in 2014 escalated this issue to the national stage. Cyber trust has therefore become a hot topic for scholars in computer science. In addition, another interesting topic in this pathway is crowdsourcing, which does not have a strong relationship with big data research. However, as previously mentioned, using big data research to support decision making is one emergent need in almost all sectors, particularly decision support for investments.

When tracing the evolutionary pathways of big data research, it is interesting to note the techniques that started before the boom and those that started after. Most large-scale data analytic techniques and basic techniques in parallel computing and distributed systems emerged before 2009, while internet of things and big data applications emerged afterwards. Aiming to further track the dynamics in big data research from 2000 to 2015, we selected four time-nodes (i.e., 2004, 2008, 2011, and 2014) to demonstrate big data research in different time periods, as shown in Figure 8. Some observations are:

- Parallel processing and distributed systems appeared before 2005, and main analytic techniques at that time included statistical analysis, association rule mining, and classification. Comparably, data warehouse could be the main technique for data storage.
- Machine learning, neural networks, and support vector machine appeared as emergent techniques after 2004 but before 2009, which then became the most representative big data analytic techniques.
- The big data boom started in the time period between 2009 and 2011, and a large number of terms that closely relate to big data research were detected, e.g., MapReduce, Hadoop, and cloud computing. Data stream and unstructured data became analytic objects in big data research, and prediction models and data visualization also appeared at that time. In addition, bioinformatrics could be credited as one pioneer of big data applications.
- Big data has attracted more and more attention since 2012. Techniques related to cloud computing and distributed systems were further developed, and social network, metadata, and social media became hotspots in big data research. At the same time, the number of big data-related projects and applications increased rapidly, and the combination of big data and internet of things is one highlight at that time.

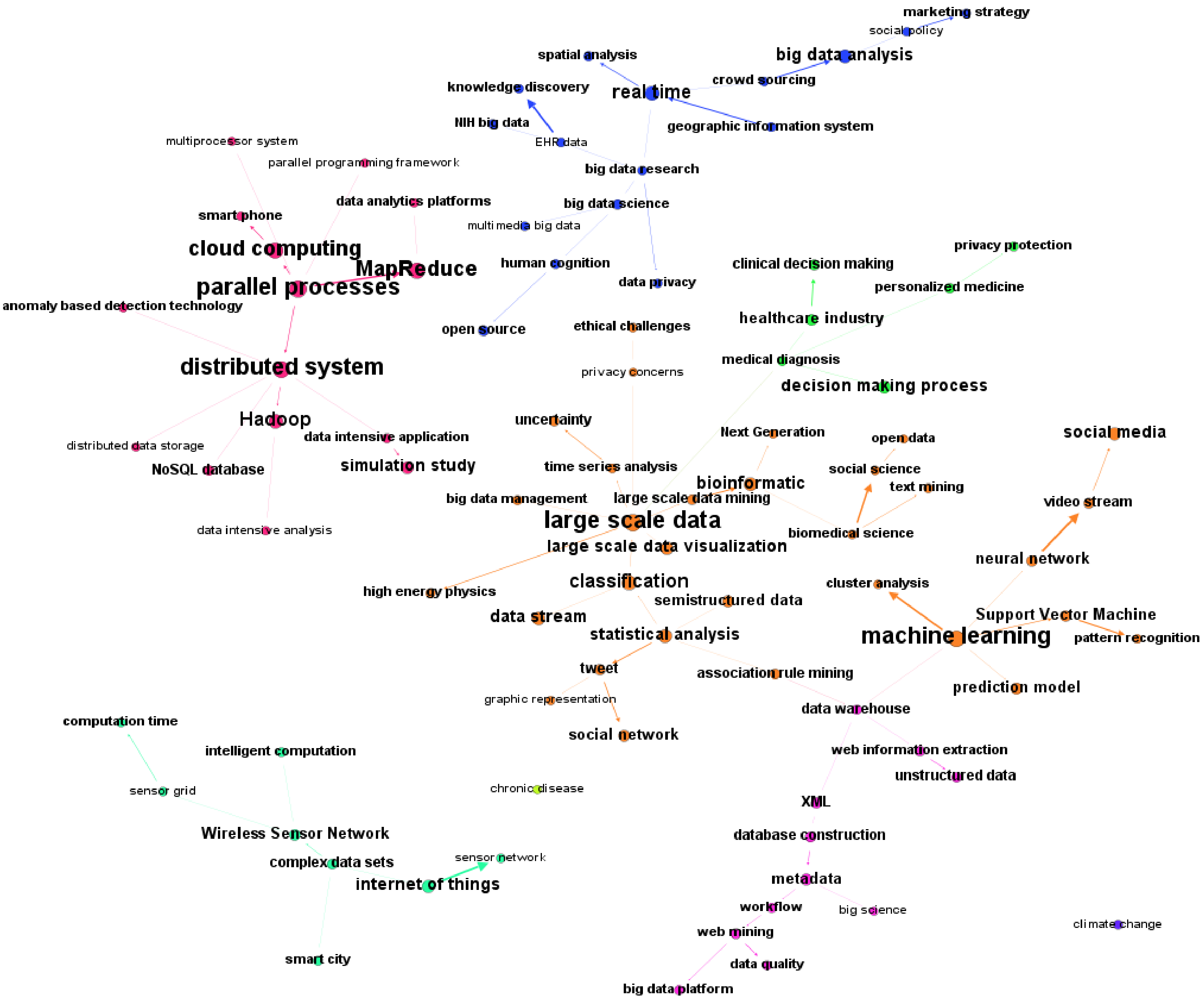

Figure 4. Scientific evolutionary pathways of big data research from 2000 to 2015

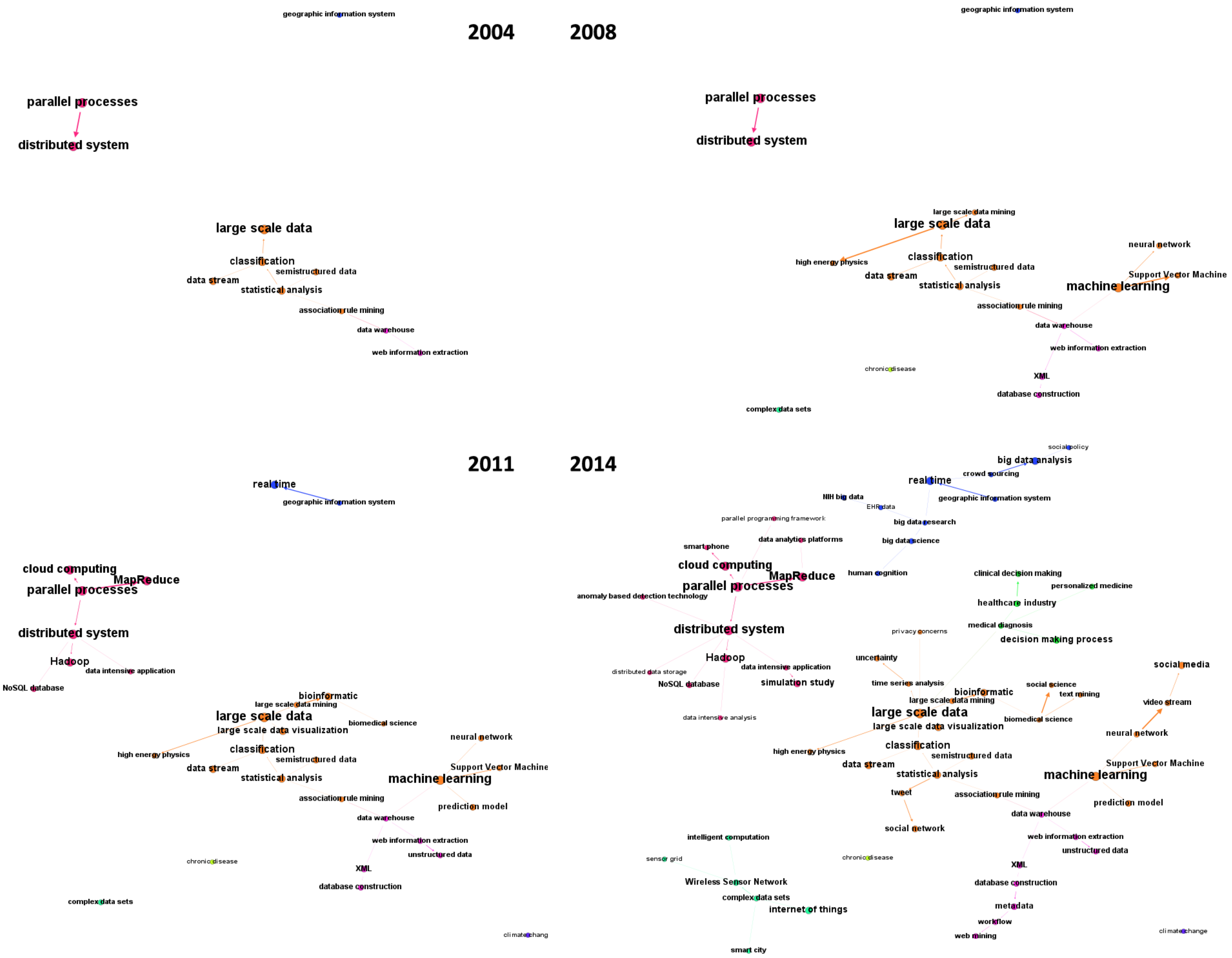

Figure 5. **Scientific evolutionary pathways of big data research in four time nodes**

In summary, based on the results of the SEP model of big data research, our key findings are:

1) Big data, especially its related techniques, reflects a recombination of largely existing techniques.
2) Analytic techniques to enhance artificial intelligence, including machine learning, data mining, pattern recognition, etc., have been widely developed, improved, and applied to big data analytics. These techniques will remain mainstream avenues of research in the near future.
3) The interaction between internet of things and big data research has deeply influences the daily lives of human beings, including the development of mobile devices, the popularization of e-communication, e-business, and e-government, and the construction of smart cities. However, these rapid developments also bring concerns for data privacy and ethical challenges, which in turn propels research on cyber trust.
4) Big data research has already been widely applied to a large number of real-world problems, and so far, its engagement with bioinformatics indicates future success.

4.2. Forecasting and Recommendations

The results of this bibliometric study provided insights on big data research from 2000 to 2015. The R&D profile addressed the questions, such as "What are the core techniques used in big data research?", "Which countries and organizations lead big data research globally?", and "How do they interact?" The SEP model explored the interactions between the techniques used within big data, as well as identifying the pathways for how these techniques evolved between 2000 and 2015.

An expert panel was arranged, and its members include academic researchers (from the School of Software and Centre for Artificial Intelligence at the University of Technology Sydney, and the School of Computer Science and School of Management and Economics at the Beijing Institute of Technology) and industry partners (from some Australian and Chinese IT companies).

We conducted interviews by either email or face-to-face consultancies, with the following steps:

1) The empirical insights (including tables, figures, and key findings) derived by the R&D profile and SEP models were delivered to experts. They first reviewed the results based on their knowledge: if disagree, give reasons or modifications (fortunately, such situation did not occur so far); if agree, extensively enrich our findings. Then, specifically based on Table 6 and Figures 7 and 8, they evaluated related topics and their emergence and identified certain possible highlights for future study.

2) We collected and manually reviewed the feedback, and then modified our results (in particular key findings). Since certain conflicts existed, which might be due to diverse knowledge background and working environments (academia and industry), we decided to emphasize the interest of academia (since our results were derived from scientific articles), and consulted several third-party academic researchers and finalized our key findings.

In addition, one co-author of this paper was selected as a consulted expert for a forum entitled Data and Analytics Innovation organized by the Government Accountability Office (GAO) of the US (Government Accountability Office 2016). Parts of the results were also presented in the workshop with GAO experts, and positive feedback was enriched.

We identified and briefly summarized the key findings of the forecasting study as follows:

- Artificial intelligence techniques will still lead big data analytics

Apparently, computer science is still the main battlefield for big data analytics. Specifically, the use of neural networks for processing imagines and videos is spearheading certain new directions in computer vision, and handling complicated and uncertain data by engaging various machine learning models can be considered another challenging direction.

- Real-world problem-driven applications will attract increasing attentions

Despite great accomplishments on developing novel big data analytic techniques and platforms, the most clearly visible application of big data analytics, beside IT sectors, is in the area of bioinformatics, since its relatively mature exploitation of modern computing techniques for biological issues. However, with the rapid development of big data analytic techniques, its application would be extended to a wide range of industry and government sectors. Considering the connections between bioinformatics and the healthcare industry, big data-enabled medical decision support or medical diagnosis could be an emergent direction.

Parallel, how to engage cloud computing to revolutionize daily life is becoming a multidisciplinary topic for both IT and business disciplines, and “cloud” would become an essential feature for future software and working platforms. At the same time, cloud computing would also act as a bridge to interactively link big data analytics with internet of things, and optimizing such linking strategy can be another interesting topic.

- Data privacy issues would be further raised and involve research communities from multiple disciplines

Together with the big data boom, open data[10] and data sharing also are becoming hot topics in public administration, which encourages building free-access platforms for anyone to access, use, and share data. However, both big data and open data would lead to data privacy issues. One example raised by one interviewed expert is some social media platforms would "steal" users' browsing data to feed their advertising functions, without conspicuously informing users. Under this circumstance, on the one hand, cyber security has already been identified as a crucial research direction for computer science (and sometimes also aligns with business disciplines by addressing certain real-world cases). On the other hand, ethical and legal issues have been widely and extensively discussed by not only academia but also government, industry, and the public.

We now provide recommendations from the perspective of both science policy and entrepreneurship.

Science policy: even though the US, the EU, China, and many other countries have already established national programs to bolster big data research, more must be done. Pursuing support for the development of big data analytics is one of the basic strengths of global competition, and prompting research institutes to extend and deepen interactions with both domestic and international collaborators will accelerate big data research – a way of standing on the shoulders of giants to achieve a win-win situation.

Entrepreneurship: collecting novel ideas in big data research from universities and academic institutions would be a smart way for companies to both explore possible technology transfers and profit at the same time. Combining internet of things, big data, and crowdsourcing, or applying big data research to non-IT sectors, such as healthcare and manufacturing, are two clear and immediate opportunities for partnership.

5. Conclusions

This paper constructs an empirical framework integrating machine learning and bibliometrics to investigate global big data research from 2000 to 2015. Specifically, an R&D profile was used to reveal insights into the statistical dynamics and geographic distribution of big data research, and, in particular, explore the global interactions between academic organizations worldwide. The SEP model introduced a machine learning process to identify the core technological clusters of big data research and detect their evolutionary pathways over the period. A forecasting study with the engagement of expert knowledge outlined certain future directions of big data research, and recommendations on science policy and entrepreneurship were provided.

Regarding limitations, future study can be conducted from the following aspects. 1) Extending the empirical data to include conference papers and web content could expose outcomes not addressed in journal articles,

[10] More information on open data can be found on the website of the European Union's Open Data Portal: https://www.europeandataportal.eu/elearning/en/module1/#/id/co-01

particularly since private companies are leading the big data boom. Understanding the role of companies in the development of big data research would also further enrich our study. 2) Given the possible negative influence of technological synonyms in term-based topic analyses, it may be interesting to further subdivide the technological areas, e.g., analytics techniques, parallel computing, distributed systems, and the internet of things. This could help improve the performance of the learning process in the SEP model. 3) Introducing a prediction model to foresee possible directions in big data research will create complementary value with the current expert knowledge-based forecasting study.

Acknowledgment

An early version of this paper has been presented in the 2017 Portland International Conference on Management of Engineering and Technology (PICMET).

This work is partially supported by the Australian Research Council under Discovery Grant DP150101645. We also acknowledge support from the US National Science Foundation (Award #1527370 – "Forecasting Innovation Pathways of Big Data & Analytics"). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the Australian Research Council or the US National Science Foundation.

References

Abernathy, W. J., & Townsend, P. L. (1975). Technology, productivity and process change. *Technological Forecasting and Social Change, 7*(4), 379-396.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of International AAAI Conference on Web and Social Media, 8*, 361-362.

Bughin, J., Chui, M., & Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly, 56*(1), 75-86.

Chen, C. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314-347.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly, 36*(4), 1165-1188.

Choi, C., & Park, Y. (2009). Monitoring the organic structure of technology based on the patent development paths. *Technological Forecasting and Social Change, 76*(6), 754-768.

Clark, K. B. (1985). The interaction of design hierarchies and market concepts in technological evolution. *Research Policy, 14*(5), 235-251.

Court, D. (2015). Getting big impact from big data. *McKinsey Quarterly, January*.

Cunningham, S., & Kwakkel, J. (2014). Tipping points in science: A catastrophe model of scientific change. *Journal of Engineering and Technology Management, 32*, 185-205.

Daim, T. U., & Oliver, T. (2008). Implementing technology roadmap process in the energy services sector: A case study of a government agency. *Technological Forecasting and Social Change, 75*(5), 687-720.

Daim, T. U., Yoon, B.-S., Lindenberg, J., Grizzi, R., Estep, J., & Oliver, T. (2017). Strategic roadmapping of robotics technologies for the power industry: A multicriteria technology assessment. *Technological Forecasting and Social Change*.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Eisenman, M. (2013). Understanding aesthetic innovation in the context of technological evolution. *Academy of Management review, 38*(3), 332-351.

Government Accountability Office. (2016). *Data and Analytics Innovation: Emerging Opportunities and Challenges Highlights of a Forum Convened by the Comptroller General of the U.S.* Washington DC: US Government Printing Office Retrieved from http://www.gao.gov/assets/690/680265.pdf.

Guo, J., Wang, X., Li, Q., & Zhu, D. (2016). Subject–action–object-based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change, 105*, 27-40.

Hu, J., & Zhang, Y. (2017). Discovering the interdisciplinary nature of Big Data research through social network analysis and visualization. *Scientometrics*, 1-19.

Huang, L., Zhang, Y., Guo, Y., Zhu, D., & Porter, A. L. (2014). Four dimensional science and technology planning: A new approach based on bibliometrics and technology roadmapping. *Technological Forecasting and Social Change, 81*, 39-48.

Huang, Y., Schuehle, J., Porter, A. L., & Youtie, J. (2015). A systematic method to create search strategies for emerging technologies based on the Web of Science: Illustrated for 'Big Data'. *Scientometrics, 105*(3), 2005-2022.

Huang, Y., Porter, A. L., Cunningham, S. W., Robinson, D. K., Liu, J., & Zhu, D. (2017a). A technology delivery system for characterizing the supply side of technology emergence: Illustrated for Big Data & Analytics. *Technological Forecasting and Social Change*.

Huang, Y., Zhu, D., Qian, Y., Zhang, Y., Porter, A. L., Liu, Y., & Guo, Y. (2017b). A hybrid method to trace technology evolution pathways: a case study of 3D printing. *Scientometrics, 111*(1), 185-204.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). *Big data: Issues and challenges moving forward.* System Sciences (HICSS), 2013 46th Hawaii International Conference on.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change, 75*(6), 771-782.

Kostoff, R. N., Eberhart, H. J., & Toothman, D. R. (1999). Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. *Journal of the Association for Information Science and Technology, 50*(5), 427.

Kostoff, R. N., & Schaller, R. R. (2001). Science and technology roadmaps. *IEEE Transactions on Engineering Management, 48*(2), 132-143.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management, 34*(3), 387-394.

Kwon, T. H., Kwak, J. H., & Kim, K. (2015). A study on the establishment of policies for the activation of a big data industry and prioritization of policies: Lessons from Korea. *Technological Forecasting and Social Change, 96*, 144-152.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6), 481-497.

Li, M. (2015). A novel three-dimension perspective to explore technology evolution. *Scientometrics, 105*(3), 1679-1697.

Lu, N., Zhang, G., & Lu, J. (2014). Concept drift detection via competence models. *Artificial Intelligence, 209*, 11-28.

Lucio-Arias, D., & Leydesdorff, L. (2008). Main‐path analysis and path‐dependent transitions in HistCite™‐based historiograms. *Journal of the American Society for Information Science and Technology, 59*(12), 1948-1962.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Mao, R., Xu, H., Wu, W., Li, J., Li, Y., & Lu, M. (2015). Overcoming the challenge of variety: big data abstraction, the next evolution of data management for AAL communication systems. *IEEE Communications Magazine, 53*(1), 42-47.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., . . . Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association, 21*(6), 957-958.

Marx, V. (2013). Biology: The big challenges of big data. *nature, 498*(7453), 255-260.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*: Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data. *The management revolution. Harvard Bus Rev, 90*(10), 61-67.

Moehrle, M. G., Isenmann, R., & Phaal, R. (2013). Technology Roadmapping for Strategy and Innovation. *Charting the Route to Success. Berlin et al.: Springer*.

Nobre, G. C., & Tavares, E. (2017). Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. *Scientometrics, 111*(1), 463-492.

Park, H. W., & Leydesdorff, L. (2013). Decomposing social and semantic networks in emerging "big data" research. *Journal of Informetrics, 7*(3), 756-765.

Phaal, R., Farrukh, C. J., & Probert, D. R. (2004). Technology roadmapping - a planning framework for evolution and revolution. *Technological Forecasting and Social Change, 71*(1), 5-26.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827-1843.

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. Auckland: McGraw-Hill.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Schermann, M., Hemsen, H., Buchmüller, C., Bitter, T., Krcmar, H., Markl, V., & Hoeren, T. (2014). Big Data. *Business & Information Systems Engineering, 6*(5), 261-266.

Singh, V. K., Banshal, S. K., Singhal, K., & Uddin, A. (2015). Scientometric mapping of research on 'Big Data'. *Scientometrics, 105*(2), 727-741.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450-1467.

Sood, A., & Tellis, G. J. (2005). Technological evolution and radical innovation. *Journal of marketing, 69*(3), 152-168.

van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467-472.

Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629-635.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234-246.

Wen, J. (2009). Let Science and Technology Lead China's Sustainable Development (Vol. 2014).

Zhang, Y., Guo, Y., Wang, X., Zhu, D., & Porter, A. L. (2013). A hybrid visualisation model for technology roadmapping: Bibliometrics, qualitative methodology and empirical study. *Technology Analysis & Strategic Management, 25*(6), 707-724.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014a). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.

Zhang, Y., Zhou, X., Porter, A. L., & Gomila, J. M. V. (2014b). How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: "Problem & Solution" pattern based semantic TRIZ tool and case study. *Scientometrics, 101*(2), 1375-1389.

Zhang, Y., Zhou, X., Porter, A. L., Gomila, J. M. V., & Yan, A. (2014c). Triple Helix innovation in China's dye-sensitized solar cell industry: Hybrid methods with semantic TRIZ and technology roadmapping. *Scientometrics, 99*(1), 55-75.

Zhang, Y., Chen, H., & Zhu, D. (2015). Semi-automatic technology roadmapping composing method for multiple science, technology, and innovation data incorporation. . In T. Daim, A. L. Porter & D. Chiavetta (Eds.), *Anticipating Future Innovation Pathways through Large Data Analytics*. New York: Springer.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179-191.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Science evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925-1939.

Zhou, X., Zhang, Y., Porter, A. L., Guo, Y., & Zhu, D. (2014). A patent analysis method to trace technology evolutionary pathways. *Scientometrics, 100*(3), 705-721.

Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change*, *69*(5), 495-506.