The Spectrum of Big Data Analytics

Zhaohao Sun [1, 2], Yanxia Huo [3]

[1] Research Centre of Big Data Analytics and Intelligent Systems

Department of Business Studies

PNG University of Technology

Lae 411, Morobe, PNG

zhaohao.sun@pnguot.ac.pg or zhaohao.sun@gmail.com

&

[2] School of Science, Engineering and Information Technology

Federation University Australia

[3] Teaching Learning Methods Unit (TLMU)

PNG University of Technology

Lae 411, Morobe, PNG

monica.huo@pnguot.ac.pg

Abstract: Big data analytics is playing a pivotal role in big data, artificial intelligence, management, governance and society with the dramatic development of big data, analytics, artificial intelligence. However, what is the spectrum of big data analytics and how to develop the spectrum are still a fundamental issue in the academic community. This paper addresses these issues by presenting a big data derived small data approach. It then uses the proposed approach to analyse the top 150 profiles of Google Scholar including big data analytics as one research field and proposes a spectrum of big data analytics. The spectrum of big data analytics mainly includes data mining, machine learning, data science and systems, artificial intelligence, distributed computing and systems, and cloud computing, taking into account degree of importance. The proposed approach and findings will generalize to other researchers and practitioners of big data analytics, machine learning, artificial intelligence and data science.

Keywords: big data, big data analytics, machine learning, artificial intelligence, data science.

This paper has been published. To cite this paper, please use

Sun Z, & Huo Y (2021) The spectrum of big data analytics. Journal of Computer Information Systems 61(2): 154-162. DOI. 10.1080/08874417.2019.1571456.

1 INTRODUCTION

Big data are generated from various instruments, billions of phones, payment systems, cameras, sensors, Internet transactions, emails, videos, click streams, social networking services and other sources (Henke & Bughin, 2016). The characteristics of big data include at least 10 bigs: big volume, big velocity, big variety, big veracity, big intelligence, big analytics, big infrastructure, big service, big value, and big market (Sun, Strang, & Li, 2018) (Sun, Sun, & Strang, 2016) (Minelli, Chambers, & Dhiraj, 2013). Big data has become a strategic resource for industry, business, governance and national security. In addition, big data nowadays has also become a strategic enabler of exploring business insights and economy of services and economy of intelligence (Chen, Chiang, & Storey, 2012)(Sun, Strang, & Firmin, 2017) (Liang & Liu, 2018). In this regard big data has created significant new opportunities for an organization to derive big value and create competitive advantage (EMC, 2015).

Big data analytics or big analytics (BA) has been drawing increasing attention in academia of computer science, information technology, mathematics, operations research, decision science, business, management and industry of healthcare, medical science (Sun, Zou, & Strang, 2015) (Liang & Liu, 2018) (Laney & Jain, 2017). Big data analytics has become a mainstream market adopted broadly across industries, organizations, and geographic regions and among individuals to facilitate big data-driven decision making for organizations and individuals to achieve desired business outcomes (Sun, Strang, & Firmin, 2017) (Laney & Jain, 2017). Big data analytics is playing a pivotal role in big data, artificial intelligence, management, governance and society with the dramatic development of big data, analytics, artificial intelligence (Straetgy Analytics, 2018). However, little literature concerns the following research questions:

- What is a spectrum of big data analytics?
- How to develop the spectrum of big data analytics?
- What is the distribution of top scholars of big data analytics across the world?

This article addresses these three research questions. More specifically, it presents a big data derived small data approach as its theoretical and methodological foundation for addressing the second and third research questions. It then uses the proposed approach to analyse the top 150 profiles of Google Scholar including big data analytics as one research field and proposes a spectrum of big data analytics. The research demonstrates that the spectrum of big data analytics mainly includes data mining, machine learning, data science and systems, artificial intelligence, distributed computing and systems, and cloud computing, taking into account degree of importance. The research also identifies the top 10 countries where big data analytics scholars work. This article finally examines the theoretical, technical and social implications of this research. The proposed approach and findings could generalize to other researchers and practitioners of big data analytics, machine learning, artificial intelligence, and data science.

The remainder of this article is organized as follows. First of all, it presents a background for the proposed research. It proposes a big data derived small data approach. Then the spectrum of intelligent big data analytics is presented. The top 10 countries where big data analytics scholars work are identified and then the theoretical, technical, and social implications of this research are examined in this article. The final section ends with some concluding remarks and future research directions.

2 BACKGROUND

This section provides a background on big data, big data analytics, machine learning, data mining, artificial intelligence and data science for research of the spectrum of big data analytics.

Spectrum. In mathematics, a spectrum is a set of elements that meet certain conditions or properties (Wiktionary, 2018). Based on this mathematical definition, a spectrum of big data analytics is a set of research disciplines that have a close relationship with big data analytics.

Big data. Big data can be refined as "the datasets whose volume, velocity, variety and veracity are so big that is beyond the ability of typical ICT tools to capture, store, manage, and analyze" (Manyika, Chui, & Bughin, 2011). For example, big variety means that big diversity or big different types of data sources with different structures from which it arrived, and the types of data available to everyone (Sun, Strang, & Li, 2018). Big data can be classified into three types: structured, semi-structured, and unstructured at a higher level. The data stored in relational database systems like Oracle are structured. The data available on the Web are unstructured. 80% of the world's data is unstructured (Sathi, 2013). The big variety exists in the data on the Web. Blogs and tweets on social media are not structured data, because they contain a large amount of slang words, with a mix of languages in a multiethnic, multi-language environment (Sathi, 2013). Big data has become a new ubiquitous term. Big data is transforming science, engineering, technology, medicine, healthcare, finance, business and management, education, and ultimately our society itself using big data analytics (Minelli, Chambers, & Dhiraj, 2013)(Sun, Strang, & Li, 2018).

Big data analytics. Big data analytics is a science and technology about organizing big data, analyzing and discovering knowledge, patterns and intelligence from big data, visualizing and reporting the discovered knowledge for assisting decision making (Sun, Sun, & Strang, 2016). The main components of big analytics include big data descriptive analytics, predictive analytics and prescriptive analytics (Sun, Sun, & Strang, 2018), which correspondingly address the three questions of big data: when and what occurred? what will occur? and what is the best answer or choice under uncertainty? All these questions are often encountered in almost every part of science, technology, business, management, organization and industry.

Machine learning. Machine learning is concerned about how computer can adapt to new circumstances and to detect and extrapolate patterns (Russell & Norvig, 2010, p. 2). The essence of machine learning (ML) is an automatic process of pattern recognition by a learning machine (Wu, Buyya, & Ramamohana, 2016). Machine learning mainly aims to build systems that can perform at or exceed human level competence in handling many complex tasks or problems.

Data mining. Data mining is a process of discovering various models, summaries, and derived values, knowledge from a given collection of data (Kantardzic, 2011). Data mining includes descriptive data mining and predictive data mining. The former produces new non-trivial information and knowledge, while the latter produces models and roles of the systems. The primary tasks of descriptive data mining include clustering, summarization, dependency modelling. The primary tasks of predictive data mining include classification, regression, change and deviation detection.

Data mining has its origins mainly in statistics and machine learning. Statistics has its roots in mathematics (Conover, 1999); machine learning has its roots in artificial intelligence (Wu, Buyya, & Ramamohana, 2016). Data mining and artificial intelligence share the common: knowledge discovery from big data and learning from data (Kantardzic, 2011).

Artificial intelligence (AI) is concerned with imitating, extending, augmenting /amplifying, automating intelligent behaviors of human being (Russell & Norvig, 2010). AI attempts not only to understand how human think, understand, write, learn, act rationally and smartly, but also to build intelligent entities that can think, write, perceive, understand, predict and manipulate a world.

The relations among deep learning, machine learning, and artificial intelligence are mathematically represented as follows: deep learning $\sqsubset$ machine learning $\subset$ artificial intelligence. In other words, deep learning is a subset of machine learning, and machine learning is a subset of artificial intelligence (Russell & Norvig, 2010) (Wu, Buyya, & Ramamohana, 2016).

Data science can be defined as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data." (NIH, 2018). In other words, data science has become a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, including statistics, informatics, computing, communication, management, and sociology to translate data in general and big data in specific into information, knowledge, insight and intelligence for decision making (Cao, 2017).

The relations among big data, data mining and big data analytics are mathematically represented as follows: data mining $\sqsubset$ big data analytics $\subset$ big data $\subset$ data science (EMC, 2015) (Sun, Sun, & Strang, 2016). Data scientists aim to invent data and intelligence-driven technologies and machines to represent, learn, simulate, reinforce, and transfer human-like intuition, imagination, curiosity, and creative thinking through human-data interaction and cooperation (Cao, 2017).

Both artificial intelligence and data science are a type of "intelligence science" that aims to transform data into knowledge, intelligence, and wisdom (Cao, 2017). Therefore, mathematically, the symmetric difference, artificial intelligence$\oplus$data science, will become smaller and smaller rather than bigger and bigger. In other worlds, the relationship between artificial intelligence and data science become closer and closer.

3 BIG DATA DERIVED SMALL DATA APPROACH

This section presents a big data derived small data approach. As a process, a big data derived small data approach consists of 1. Big data reduction, 2. Big data derived small data collection, and 3. Big data derived small data analysis.

3.1 Big data Reduction

Big data reduction is the first step for the big data derived small data approach. Reducing big data is, in essence, a kind of selection. The proper selection of data is usually in the name of data collection.

For example, in order to review big data analytics and classify big data analytics into categories based on research focuses, Chong and Shi search the three databases (Compendex, GEOBASE, INSPEC) using the term of "big data analytics" and find out 2960 articles (Chong & Shi, 2015) . This is the first step of big data reduction, which uses special databases to collect data, that is, big data derived small data collection. After necessary exclusion of invalid papers, Chong and Shi review the abstracts, titles through focusing on development, implementation and discussion of big data analytics and reduce the papers from 2960 to 266. It can be considered as the second step of big data reduction. Then they analyse the 266 publications and classify big data analytics into categories based on research focuses.

3.2 Big Data Derived Small Data Collection

From a statistical modelling perspective, big data derived small data collection is a special kind of sampling. "Sampling is the process of randomly collecting some data or samples when collecting all or analysing all is unreasonable" (National Research Council, 2013, p. 120) (Conover, 1999). Sampling is also a kind of big data reduction. For example, Google Scholar should be a sampling, because Google Scholar cannot collect all the data of scholars on the Internet. There are two core parts for any sampling towards data analysis based on statistical inference. One is to collect what kind of data. The second is how to collect data. The former is related to what kind of data are important for the designed research. In other words, importance of data is related to data analysis. The latter has been discussed in terms of statistical sampling. Statistical sampling includes random sampling and non-random sampling (National Research Council, 2013, p. 120).

For the importance of data, not all data need be taken for any decision making and rule-seeking as well as statistical inference (National Research Council, 2013, p. 128). Just as focusing on main problems with main solutions, one can also seek the important data for any decision making and statistical inference. For example, if one likes to do research on data analysis of social networking services, then one might collect the unstructured data from the Web or online social networking platforms, taking into account the big data derived small data analysis. Therefore, it is a big issue for a research to identify which data set is important to meet the objectives of the research.

For this research, what kind of data is important for examining the spectrum of big data analytics? The possible answers are: One is data from Google Scholar (https://scholar.google.com/), another is data from SCOPUS (https://www.scopus.com), because both have big data of scholars' publications. If we analyse the data on the first 100-200 scholars' profiles in the area with highest citation, then we can know the relationship among big data analytics and its related disciplines or research fields. If we analyse the data from SCOPUS, then we can use the key words of first 200 latest big data analytics papers of SCOPUS to know the relationship between big data analytics and related research topics or research fields within big data analytics.

Below is a simple example for big data derived small data collection. For example, age modelling and predicting from the Internet is developed as a software (Kemelmacher-Shlizerman, 2017). The developer searches the images of Google, for example, with "age five", and then analyses all the visual (image) results searched using Google and develops the software using similarity-based reasoning. The key idea behind is big data derived small data collection: The searched images of "age five" is a small data whereas the Google images are certainly big data. Therefore, this software is based on big data derived small data approach.

It should be noted that the above-mentioned software should have the function of backward reasoning, that is, if one is already elder and likes to get the image for her/his child time, for example, 5-year-old image. The software should be also further developed based on case-based reasoning, because the philosophy of case-based reasoning is "similar problems have similar solutions" (Sun, Finnie, & Weber, 2004).

This research uses Google Scholar to collect data of scholar profiles and research fields from top to the 150th in a descending order, based on the Number "cited by". The collection process is as follows.

1. Access a scholar profile at Google scholar, for example, http://scholar.google.com.au/. The scholar's research fields include "big data analytics"
2. Click "big data analytics" near to the photo (right bottom) of the scholar and get https://scholar.google.com.au/citations?view_op=search_authors&hl=en&authuser=1&mauthors=label:big_data_analytics.
3. Select first 150 scholar profiles from here: first 10 on this page, then continue >. Every scholar's profile consists of up to five research fields based on the rule of Google Scholar. For each scholar profile, collect the five (up to 5) research areas including big data analytics, and put them in the database.
4. For example, a scholar at Google Scholar has 4 research areas: Data mining, big data analytics, database systems, and information retrieval.

Using this method, this research collects up to 150 x 5 data items, each of them is a research field of a top 150 scholar of big data analytics. This research collects about 750 research fields, some scholars have only 3 or 4 research fields in the profile. The 150 x 5 data items are a small data, but it is derived from the big data of Google Scholar (http://scholar.google.com). Therefore, it is a big data derived small data collection. This data collection is, in essence, a big data reduction for the proposed research.

3.3 Big Data Derived Small Data Analysis

Big data derived small data analysis is important both for big data approach and big data analytics as a discipline. First of all, big data has basically been controlled by many global data giants such as Facebook, Google, Tencent, Baidu and Alibaba rather than by an individual scholar. It is expensive for a scholar to collect data and analyse the collected data. Sometimes, it is also very expensive for a company like Cambridge Analytica to collect data working together with Facebook, because Cambridge Analytica paid big price through its bankruptcy (Baker, 2018).

Secondly, sampling is the process of collecting some kind of data when collecting it all or analyzing it all is unreasonable (National Research Council, 2013, p. 120), as mentioned above. Sampling is a phrase of any statistical modelling or inference. This implicates that the majority of statistical inference based on sampling is reasoning based on incomplete knowledge or data. Therefore, any statistical modelling or inference is a kind of big data derived small data analysis and reasoning (National Research Council, 2013, p. 120).

For example, polls including National Election Poll is based on samples of population to measure the opinions of the whole population. To this end, the absolute size of the sample is important, but the percentage of the whole population is not important. A poll with a random sample of 1,000 people has margin of sampling error of ±3% for the estimated percentage of the whole population. In order to reduce the margin of error to 1% the poll needs a sample of around 10,000 people. In practice, a sample size of around 500–1,000 is a typical for political polls, taking into account cost (American Association for Public Opinion Research, 2018). Therefore, 1000 samples with 1 out of maximal 330,000 (The population of USA is 330 million) could lead to a satisfactory result for modelling and predicting voting in the national election. This case is a big data derived small data analysis. Because 330,000$\approx 2^{18}$, and 1 out of maximal 330,000 means that 1: 2^{18}, the mentioned sampling means that the big data from exabyte level (EB) size has been reduced to terabyte (TB) level size. Everyone could buy a TB portable hard drive to process the data with a TB level size, and relieve from the big data anxiety. This means that polls are a successful application of big data derived small data approach and big data-driven small data analytics, because just as a few thousands of people's questionnaire or phone interview through random sampling might decide the political election of a country.

Thirdly, from a data processing viewpoint, the largest data analyses could be performed in large data centers of a few global data monopolies running specialized software such as Hadoop over HDFS to harness thousands of cores to process data distributed throughout the cluster (National Research Council, 2013, p. 55). This means that individuals have to use big data derived small data analysis to analyse data.

Finally, any research in general and research publication in special is, in essence, based on big data derived small data analysis, because an average research publication consists of 30 references, which has only up to 30 MB of data from a data volume viewpoint (Wu, Buyya, & Ramamohana, 2016). In the big data world, the data with 30 MB is relatively small (Strang & Sun, 2015).

As an application of big data derived small data analysis, this research merges x analytics to big data analytics. For example, "health data analytics" is merged to "big data analytics". In such a way, big data analytics could also include health data analytics, teaching analytics, cognitive analytics mentioned by the scholars under investigation. Similarly, this research also merges y learning to machine learning, for example, deep learning is merged to "machine learning', because deep learning is a part of machine learning.

This research will use Microsoft EXCEL to analyse the collected data and present a spectrum of "big data analytics", which represents "Big data analytics" and its relationships with related research fields.

4 SPECTRUM OF BIG DATA ANALYTICS

This section proposes a spectrum of big data analytics based on the proposed big data-derived small data approach. First of all, it looks at data representations.

4.1 Data Representations

This subsection will present data representations of the research fields of a scholar, taking into account of cognitive behaviors of the scholar.

For a scholar, denoted as, s, s's up to five research fields can be represented as a set:

$$s = \{r1, r2, r3, r4, r5\} \quad (1)$$

The equation (1) reflects that five research fields of a scholar have the same importance for measuring his or her research activities and performance because the Google Scholar has not regulated that the first one research field is the most important, the fifth is least important. A scholar is only recommended to fill in up to five research fields when creating his or her Google Scholar profile. This research will use equation (1) to look at the top 10 research fields relating to big data analytics based on the big data derived small data analysis.

s's up to five research fields can be also represented as a 5-ary vector:

$$s = (r1, r2, r3, r4, r5) \quad (2)$$

The equation (2) reflects that five research fields of a scholar have priority for measuring his or her research activities and performance. $r1$ is the most important research field, the $r5$ is the least important research field.

s's up to five research fields can be represented as a dependence chain:

$$r1 \rightarrow r2 \rightarrow r3 \rightarrow r4 \rightarrow r5 \quad (3)$$

The equation (3) reflects that five research fields of a scholar have dependence relationship. For example, if $r1$ = data mining, $r2$ = big data analytics, then $r1 \rightarrow r2$ means that big data analytics is dependent on data mining, in other words, data mining determines big data analytics. In such a way, the scholar could focus on the research of data mining and apply his research to big data analytics. From a cognitive perspective, the scholar may recognise that data mining is more important than big data analytics.

We can use equation (2) and (3) to prioritize research fields relating to big data analytics and present a spectrum of big data analytics with respect to the degree of the importance.

4.2 Top 10 research fields Associating with Big Data Analytics

This subsection uses equation (1) to examine the top 10 research fields associating with big data analytics based on the big data derived small data analysis.

To this end, this research summarizes all the number of the occurrences of related research fields mentioned by the top 150 scholars for each of the five research fields. The summary is listed in the following Table 1.

Table 1. Top 10 research fields associating with big data analytics

No.	Research Fields	Occurrence No.
1	Big data analytics	150
2	Data mining	40
3	Machine learning	33
4	Data science and systems	30
5	Artificial intelligence	19
6	Distributed computing and systems	13
7	Cloud computing	13
8	Information retrieval	11
9	Social media and computing	8
10	Wireless networking computing	7
10	Computational science	7
12	Internet of things (IoT)	4
12	Software engineering	3
12	Operations research	3
12	Bioinformatics	3
12	Algorithm and algorithm theory	2
16	Numerical linear algebra	2

Table 1. demonstrates that the top 10 research fields associating with big data analytics consist of data mining, machine learning, data science and systems, artificial intelligence, distributed computing and systems, cloud computing, information retrieval, social media and

computing, wireless networking computing, and computational science, based on the number of the occurrences of related research fields.

The research fields of scholars have been merged in the following way.

For No. 1, big data analytics also include hospitality analytics (1) (where (x) means that x occurrences), text analytics (1), social analytics (3), business analytics (1), health data analytics (1), teaching analytics (1), cognitive analytics (1), healthcare data analytics (1), video analytics (1), as a research field of the scholar, although s/he might mention big data analytics as her / his research field.

For No. 2. data mining also includes text mining (3), graph mining (1), mobility Mining (1), social media mining (1), mining urban data (1).

For No. 3, machine learning includes deep learning (5). This means that deep learning has drawn much attention in machine learning and computer science.

For No. 4, data science and systems include database, data systems, data management, data visualization, data warehouse, NoSQL Data warehouses (1), multimedia databases (1), data privacy (1) and text warehousing.

For No. 5, artificial intelligence includes intelligent systems (2), intelligent virtual agents (1), and evolutionary algorithms (1).

For No. 6, distributed computing and systems include distributed base (2), distributed systems (4), distributed processing (1), and distributed computing (5).

For No. 7, information technology and systems include information retrieval (2), theory (2), information technology (1), information security, system safety (1), and music retrieval (2).

For No. 8, cloud computing also includes could architectures (1), and cloud data storage (1).

For No. 9, social media and computing includes social media (2), social networks, social set analysis (1), online networks (1), social networks (1) and social computing (1).

For No. 10, wireless networking computing also includes wireless communications, wireless networks (1), interconnection networks.

For No. 11, computational science also includes computational intelligence (2), computational statistics (1), and computational social science (1).

The rest includes the Internet of Things (IoT), software engineering, operations research including process optimization (1), bioinformatics, algorithms and algorithm theory, and numerical linear algebra. IoT has been a hot topic for big data (IEEE Big Data 2018, 2018). However, it seems the association between big data analytics and IoT is still very weak. In other words, big data analytics for IoT should draw more attention in the community of academia. Operations research, algorithms and algorithm theory, and numerical linear algebra are the foundations of big data analytics (National Research Council, 2013). Bioinformatics is an application field of big data analytics.

Mathematics, optimization, and statistical modelling (National Research Council, 2013) as well as visualization technology underpin the research and development of big data analytics (Sun & Wang, 2017).

4.3 Priority Analysis of Research Fields Associating with Big Data Analytics

This section uses equation (2) and equation (3) to prioritize research fields associating with big data analytics and present top 10 research fields.

As mentioned previously, the equation (2) reflects that a scholar has priority over the five research fields for measuring his or her research activities and performance. $r1$ is the most important research field, then is weighted to 5, the $r5$ is the least important research field, and weighted to 1. The others $r2$, $r3$, and $r4$ are weighted to 4, 3, and 2 respectively. That is, the degree of importance (*I*) associating with big data analytics is

$$I = 5 * r1 + 4 * r2 + 3 * r3 + 2 * r4 + 1 * r5 \qquad (3)$$

For example, for big data analytics, there are 38 occurrences as research fields 1 and 2 respectively, 29 occurrences as research fields 3 and 4 respectively, and 16 occurrences as research fields 5. Then, the degree of importance with respect to Big Data Analytics is

$$503 = 5 * 38 + 4 * 38 + 3 * 29 + 2 * 29 + 1 * 16$$

Using the weighted method, we analyse the spectrum of big data analytics, big data analytics and its impacts on other disciplines. For example, big data analytics will impact the other three research areas of a scholar based on the mentioned degree of importance. At the same time, we can use the weighted method to measure the importance of big data analytics in a scholar's research areas. For example, if big data analytics is the 4th research area of scholar *P* and the 1st research area of scholar *Q*. Then from a research viewpoint, big data analytics is more important to *Q* than to *P*.

Based equation (3), the occurrences of research fields have been aggregated and calculated. The result on priority rank of research fields relating to big data analytics is illustrated in the following Table 2.

Table 2. Priority rank of research fields relating to big data analytics

No.	Research Fields	Occurrence No.
1	Big data analytics	503
2	Data mining	144
3	Machine learning	131
4	data science and systems	98
5	Artificial intelligence	59
6	Distributed computing and systems	54
7	Cloud computing	47
8	Information retrieval	40
9	Computational science	32
10	Wireless networking computing	27
10	Social media and computing	18
12	Internet of Things (IoT)	12
12	Software engineering	12
12	Operations research	9
12	Algorithm theory	4
12	Numerical linear algebra	4
16	Bioinformatics	3

Table 1 and Table 2 demonstrate that the top 10 research fields associating with big data analytics are same based on the mentioned two methods. The top 7 research fields associating with big data analytics are the same, the order is also same using the mentioned two methods. The order of the following three research fields: social media and computing, wireless networking computing, and computational science have been changed into: computational science, wireless networking computing, social media and computing, if the weighted method replaces the set-element-number count method.

We normalize each of the top 10 research fields with respect to big data analytics by dividing 1.44 and obtain the pivot chart on the importance of top 10 research fields with respect to big data analytics, as shown in Figure 1.

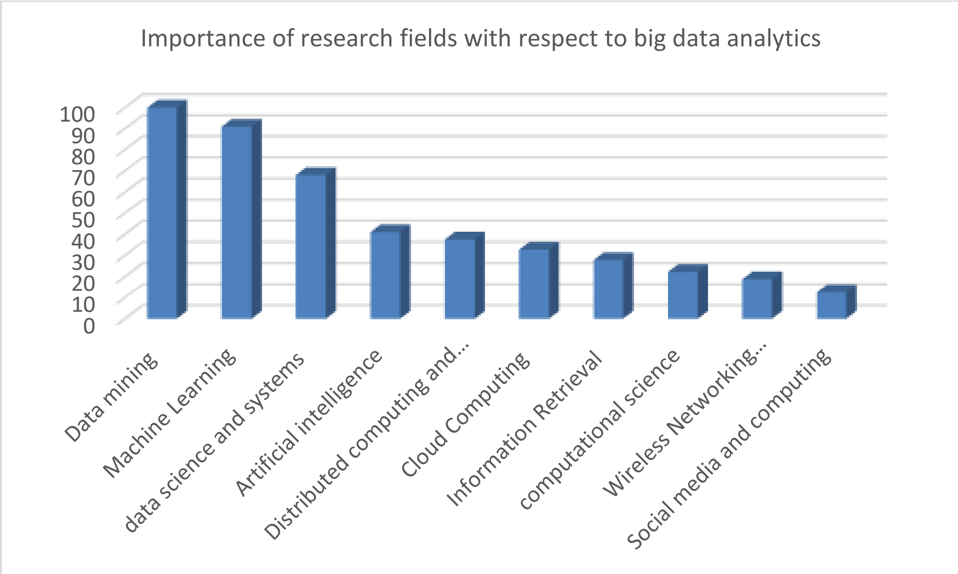

Figure 1. Importance of top 10 research fields with respect to big data analytics

Figure 1. demonstrates that data mining, machine learning, data science and systems, artificial intelligence, distributed computing and systems, cloud computing, information retrieval, computational science, social media and computing are the top 10 research fields with close relationship with big data analytics.

4.4 Distribution of Big Data Analytics Scholars across the Countries

This section explores scholars whose research area is in big data analytics and their distribution worldwide.

Based on the data collected from the Google Scholar, this research conducts a statistical analysis on the distribution of scholars focusing on big data analytics across the world. The result of data analysis demonstrates that the top 150 scholars of big data analytics are from 30 different countries. They are USA, China, Australia, UK, Canada, Belgium, Greece, India, Italy, South Korea, Qatar, Singapore, Spain, Germany, UAE, Azerbaijan, Denmark, Egypt, France, Iraq, Ireland, Japan, Luxembourg, Malaysia, Netherlands, New Zealand, Philippine, PNG, Switzerland, and Turkey. The top 10 countries ranked by the number of the big data analytics scholars are USA, China, Australia, UK, Canada, Belgium, Greece, India, Italy, South Korea. The number of scholars of big data analytics distributed across these 10 countries are illustrated in the following Figure 2.

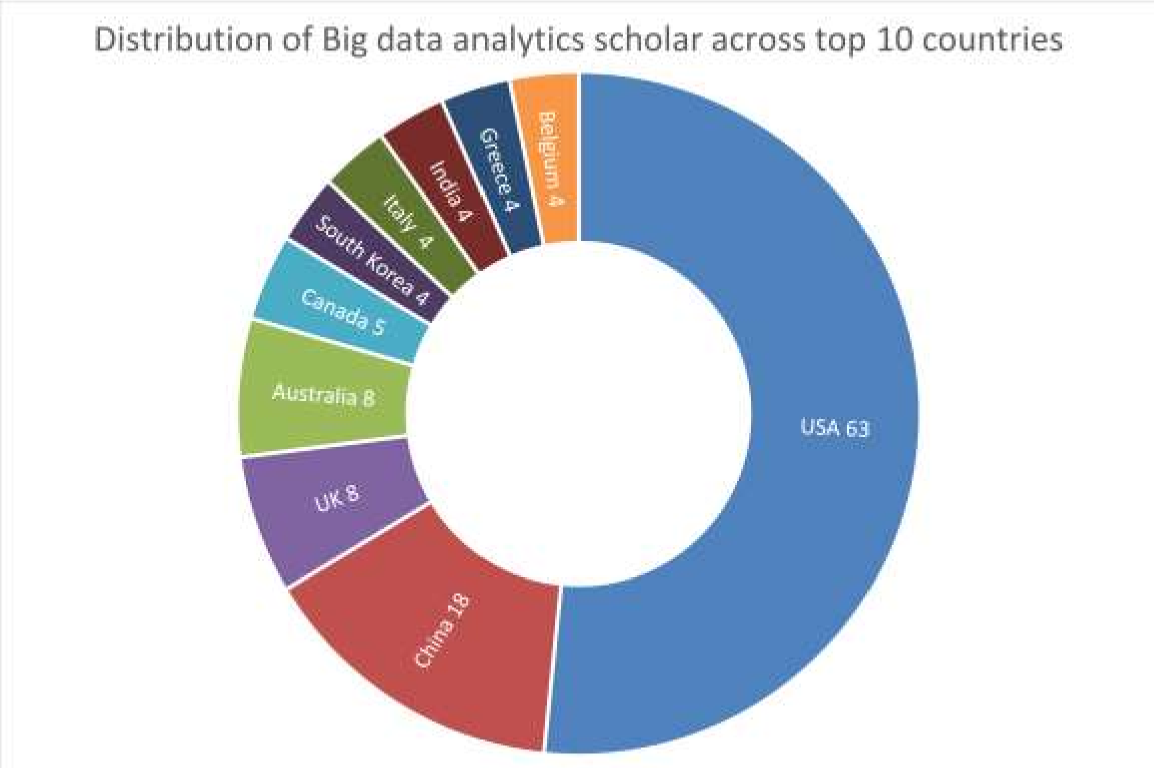

Figure 2. Top 10 countries ranked with the number of big data analytics scholars

The analysis also demonstrates that there are 122 scholars of big data analytics working in the top 10 countries, accounting for 81%, while the rest, about 19%, working in the other 20 countries. The number of scholars of big data analytics of USA and China has taken 54%. This implies that the two largest economies in the world, USA and China, have dominated the research and development of big data analytics. This result is similar to the result of http://www.guide2research.com/scientists/. Guide2research lists the Top 1000 computer science and electronics scientists with H-Index, of them more than 600 scientists listed are working in USA (Guide 2 Research, 2018). This reflects that economy has a big impact on the research and research outcome of big data analytics. However, big data is ubiquitous in every corner of the world, big data analytics should draw more attention as a science and technology worldwide.

5 DISCUSSION AND IMPLICATIONS

This section will discuss the related work, examine the theoretical, technical and social implications of this research, and explore the limitation of the research.

5.1 Discussion

This research is similar to http://www.guide2research.com/scientists/ that lists the 1000 Top H-Index for Computer Science and Electronics scientists having H-Index>=40 provided by Google Scholar (Guide 2 Research, 2018). Both use the H-index of the Google Scholar to select and collect data. Both are a kind of big data derived small data analytics. The difference between this research from that is that this research focuses on only the top 150 scholars' research field data and its small data analysis, the Guide 2 Research is on ranking scholars based on H-index directly.

Chong and Shi classify big data analytics into the following categories based on research focuses (Chong & Shi, 2015): Big data acquisition and storage, big data programming model, big data analysis, benchmark and application. None of them appears in Figure 1. The reasons behind these differences might be that the research fields (the scholars' research areas relating to big data analytics) might be at higher level than the proposed categories of big data analytics as a taxonomy (Chong & Shi, 2015). The mentioned categories are so concrete that no selected scholars of Google Scholar propose any of them as a research field. An interesting question arises: What is the relationship between a scholar's research fields and key words of her or his publications? We will address it in the future work.

This research considers big data analytics as the core of big data (Sun, Sun, & Strang, 2018), although big data analytics has been classified by IEEE Big Data 2018 into the categories of big data applications and big data privacy and security (IEEE Big Data 2018, 2018). To some extent, big data would be trash without big data analytics.

5.2 Theoretical, technical, and social implications of this research

Nowadays, some researchers and practitioners have not cared about big data and big data analytics, because the big data have been controlled by a few global data giants. It is hard to do any research on big data and big data analytics without big data for the researchers in the area of big data analytics. Therefore, the technical implication of this research is that the proposed big data derived small data approach could relieve the big data anxiety and then attract more and more researchers and practitioners to undertake the research and application of big data and big data analytics.

At the same time, the proposed spectrum of big data analytics could facilitate the integration of big data analytics and other mentioned research fields from a system integration viewpoint. It also differentiates big data analytics from other mentioned disciplines such as machine learning and data mining.

The proposed distribution of big data analytics scholars has demonstrated that USA and China have dominated the research and application of big data analytics. This implies that every country should invest more to big data and big data analytics in order to improve its global competition. Otherwise, one will have more disadvantages in global competition in big data and big data analytics-driven industry.

5.3 The limitation of this research

Google Scholar in this research is only used to analyse the relationship among big data analytics and others at the disciplinary level. In order to analyse the deep relationship within big data analytics at the internal level, which can be considered as association analysis of big data analytics, we will search wikiCFP and SCOPUS. For wikiCFP, we search all the CFP in the name of big data analytics. For SCOPUS, we search for the first 200 papers, as a big data derived small data collection and analysis, titled "big data analytics" with highest citations, and collect its related five key words, because keywords is a measurement for the relationships between big data analytics with related discipline and also research areas within big data analytics. It will be done as a future research.

6 CONCLUSION

Big data analytics is playing a pivotal role in big data, analytics, artificial intelligence (Schalkoff, 2011), management, governance (Sun, Sun, & Strang, 2018). This article presented a big data derived small data approach. It then used the proposed approach to analyse the top 150 scholar profiles of Google Scholar including big data analytics as one research field ranking with Google citations, and proposed a spectrum of big data analytics. The research demonstrates that the spectrum of big data analytics mainly includes data mining, machine learning, data science and systems, artificial intelligence, distributed computing and systems, and cloud Computing, taking into account degree of importance. The research also showed that as the largest economies in the world, USA and China, have dominated the research and application of big data analytics. This article also examined the technical, theoretical, and social implications of this research. The proposed approach and findings of this research will generalize to other researchers and practitioners of big data analytics, machine learning, artificial intelligence, and data science.

In the future work, we will search for SCOPUS using “big data analytics” to select 200 articles with the title including big data analytics and analyse the association of the key words of each paper and find the relationship among them. It also compares with the proposed result of this research in order to extend the spectrum of big data analytics to one with three-level structure.

7 ACKNOWLEDGMENT

We gratefully thank Google Scholar for its openness of research data. We also gratefully thank Hebei University of Science and Technology, Chongqing Normal University, and Federation University Australia for providing excellent research environments for completing this research.

8 REFERENCES

American Association for Public Opinion Research. (2018). *Margin of Sampling Error/Credibility Interval.* Retrieved June 29, 2018, from https://www.aapor.org/Education-Resources/Election-Polling-Resources/Margin-of-Sampling-Error-Credibility-Interval.aspx

Betser, J., & Belanger, D. (2013). Architecting the enterprise with big data analytics. In J. Liebowitz, *Big Data and Business Analytics* (pp. 1-20). Boca Raton, FL: CRC Press .

Cao, L. (2017). Data science: challenges and directions. *CACM 60 (8)* , P59-68.

Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big imppact. *MIS Quarterly, Vol. 36 No. 4, December* , 1165-1188.

Chong, D., & Shi, H. (2015). Big data analytics: a literature review. *Journal of Management Analytics, 2*(3), 175-201.

Conover, W. J. (1999). *Practical Nonparametric Statistics (3rd Edition).* New York: Wiley & Sons, Inc.

EMC. (2015). *EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* John Wiley & Sons.

Guide 2 Research. (2018). *Top H-Index for Computer Science and Electronics.* Retrieved July 9, 2018, from Guide 2 Research: Top H-Index for Computer Science and Electronics

Henke, N., & Bughin, J. (2016, December). *McKinsey Global Institute.* Retrieved from The Age of Analytics: Competing in a Data Driven World: MGI@mckinsey.com

IEEE Big Data 2018. (2018). *CFP IEEE Big Data 2018.* Retrieved from http://cci.drexel.edu/bigdata/bigdata2018/index.html

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms.* Hoboken, NJ: Wiley & IEEE Press.

Kemelmacher-Shlizerman, I. (2017, Nov 28). *Modeling People From Visual Data.* Retrieved June 26, 2018, from TEDxVienna: https://www.youtube.com/watch?v=2osANHAz284&t=245s

Kumar, B. (2015). An encyclopedic overview of 'big data' analytics. *International Journal of Applied Engineering Research, 10*(3), 5681-5705.

Laney, D., & Jain, A. (2017, June 20). *100 Data and Analytics Predictions Through.* Retrieved August 04, 2018, from Gartner: https://www.gartner.com/events-na/data-analytics/wp-content/uploads/sites/5/2017/10/Data-and-Analytics-Predictions.pdf

Liang, T.-P., & Liu, Y.-H. (2018). Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study. *Expert Systems With Applications, 111*, 2-10.

Loshin, D. (2013). *Big Data Analytics: From Strategic Planning to Enterprise Integration woth Tools, Techniques, NoSQL and Graph.* Amsterdam: Elsevier .

Manyika, J., Chui, M., & Bughin, J. e. (2011, May). *Big data: The next frontier for innovation, competition, and productivity.* Retrieved from McKinsey Global Institute: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses.* Wiley & Sons (Chinese Edition 2014).

National Research Council. (2013). *Frontiers in Massive Data Analysis.* Washington DC: The National Research Press.

NIH. (2018, July 04). *NIH Strategic Plan for Data Science.* Retrieved July 6, 2018, from National Institutes of Health: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (3rd Edition).* Upper Saddle River: Prentice Hall.

Sathi, A. (2013). *Big data analytics: Disruptive technologies for changing the game.* Boise, ID, USA: MC Press: IBM Corporation.

Schalkoff, R. J. (2011). *Intelligent Systems: Principles, Paradigms, and Pragmatics.* Boston: Jones and Bartlett Publishers.

Straetgy Analytics. (2018). *Straetgy Analytics.* Retrieved July 20, 2018, from Straetgy Analytics: https://www.strategyanalytics.com/

Strang, K., & Sun, Z. (2015). Analyzing relationships in terrorism big data using Hadoop and statistics. *Journal of Computer and Information Systems (JCIS), in Press*.

Sun, Z., & Wang, P. P. (2017). Big Data, Analytics and Intelligence: An Editorial Perspective. *Journal of New Mathematics and Natural Computation, 13*(2), 75–81.

Sun, Z., Finnie, G., & Weber, K. (2004). Case base building with similarity Relations. *Information Sciences (Elsevier), 165*(1-2), 21-43.

Sun, Z., Strang, K., & Firmin, S. (2017). Business Analytics-Based Enterprise Information Systems. *Journal of Computer Information Systems*, 57(2): 169-178 DOI: 10.1080/08874417.2016.1183977.

Sun, Z., Strang, K., & Li, R. (2018). Big Data with Ten Big Characteristics. *Proceedings of 2018 The 2nd Intl Conf. on Big Data Research (ICBDR 2018), October 27-29* (pp. 56-61). Weihai, China: ACM.

Sun, Z., Sun, L., & Strang, K. (2018). Big Data Analytics Services for Enhancing Business Intelligence. *Journal of Computer Information Systems (JCIS), 58*(2), 162-169 . doi:10.1080/08874417.2016.1220239

Sun, Z., Zou, H., & Strang, K. (2015). Big Data Analytics as a Service for Business Intelligence . *I3E2015, LNCS 9373* (pp. 200-211). Berlin: Springer.

Wiktionary. (2018, July 8). *Spectrum.* Retrieved July 20, 2018, from https://en.wiktionary.org/wiki/spectrum

Wu, C., Buyya, R., & Ramamohana, K. (2016). *Big Data Analytics = Machine Learning + Cloud Computing.* Retrieved July 20, 2018, from https://arxiv.org/ftp/arxiv/papers/1601/1601.03115.pdf