Integrated Economy and Society: Diversity, Creativity, and Technology
16–18 May 2018
Naples • Italy

make learn
TIIM

Management, Knowledge and Learning
International Conference 2018
Technology, Innovation and Industrial Management

BIG DATA AND DATA SCIENCE: A SCIENTOMETRICS APPROACH

Anita Papić
Faculty of social sciences and humanities in Osijek, Croatia
apapic@ffos.hr

Endrina Eskić
Faculty of social sciences and humanities in Osijek, Croatia
eeskic@ffos.hr

Abstract:
Big data can be defined as collection of data from traditional and digital sources inside and outside of certain organization which can be used for analyzes and discoveries. This paper gives theoretical insight into the emerging field of big data and data science as multidisciplinary science with main focus on data. Data science field aroused at intersection of several other well established fields such as social sciences, statistics, information science, computer science and design. Also in the paper are described data science techniques such as data mining, machine learning and data visualization. The special emphasize is given to business intelligence (BI) which encompasses strategies and technologies that companies use for business data analytics. In the empirical part of the paper the scientometrics analyze was conducted to find out which publications publish about big data and data science the most, which regions, institutions and authors are the most productive in the field of big data and data science and in which scientific disciplines are big data and data science employed the most.

Keywords: big data, business intelligence, data science, scientometrics

1. INTRODUCTION

Data science is multidisciplinary science with focus on data. Data are ubiquitous and through data mining it can be identified patterns within data which can improve humans' condition and make commercial and social values. Big data could help us to better understand physical and biological systems as well as social and economic systems (Stanton, 2013). Nowadays every organization which wants to attract community is confronted with issue how to more efficient use their own data but also all other available data (Peng & Matsui, 2017). Manipulation with big data demands special skills and tools. Big data are often too voluminous to be stored at just one computer and to be managed by traditional database systems, statistical packages or standard graphical software. Big data are much heterogeneous, uncompleted and unstructured such as for example sensor's data or audio and visual content. Provenance and quality of big data are often questionable and also datasets must be combined to be useful. Manipulation with users' datasets opens many questions such as privacy, security and ethical issues. Data science field aroused at intersection of several other well established fields such as social sciences, statistics, information science, computer science and design (Peng & Matsui, 2017).

Big data can be defined as collection of data from traditional and digital sources inside and outside of certain organization which can be used for analyzes and discoveries. There are two groups of data encompassed within the term of big data. The first group of data are unstructured data namely these data are unorganized and cannot be managed by traditional database systems. For example, metadata, tweets and other posts of social media are unstructured data. The second group of data are multi-structured data which are induced from human-computer interactions, web applications and transactions. More and more multi-structured data will be produced thus digital age creates new communication channels and interactions, social platforms and marketing experts try to improve users experience in regard to different electronic devices. The company Gartner emphasizes three (3V) following characteristics of big data while others add two more characteristics of big data (validity and value): volume - huge amounts of data, velocity- speed of generation of information in almost real time, variety- different formats of available data.

Business intelligence (BI) encompasses strategies and technologies which companies use for business data analytics. Usual functions of BI technologies include reporting, analytics, data mining, comparing and predictive analytics. BI technologies can do big data analytics in purpose to identify and develop new strategic business opportunities. The aim of business intelligence is to enable interpretation of big data (Varcellis, 2009). Identification of new possibilities and efficient strategy conduction based upon insights in big data can companies provide comparative advantage within the market and long-term stability. Companies can use business intelligence as support in wide spectrum of business decisions from operational to strategical decisions. The main business decisions include products' positioning or prices and strategical business decisions include making priorities, goals and directions. In all cases BI is the most efficient when combines data obtained from the market where company conducts business (external data) with data from inside the company such as financial data (internal data). Internal and external data when combined can give the complete picture or in fact create "intelligence". Business intelligence tools, if often used, can empower organization to get insight into new markets, products' and services' adjustment to different segments of markets and to predict needed marketing impact. Nowadays, business intelligence includes even probabilistic simulations, optimization of key success indicators, process management control etc. Some of business intelligence techniques are data mining, machine learning and data visualization.

Data mining is computational process of pattern discovery in big datasets which includes methods at intersection of machine learning, statistics and database systems. Data mining is one step of knowledge discovery in data (KDD) within databases. The real task of data mining is semi-automatic and automatic analytics of big data and discovery of patterns such as clusters, association rules, anomalies etc. (Thuraisingham, 2000). Machine learning is the field of computer science which computers give opportunities to learn something without real programming. Machine learning was developed from study of

pattern recognition and artificial intelligence and includes algorithms' construction which can learn and make predictions (Kohavi & Provost, 1998).

This paper aims to explore the role of big data and data science within the established scientific disciplines and publications as well as the most productive regions, institutions and authors in this emerging field thus the research questions in this paper are the following:

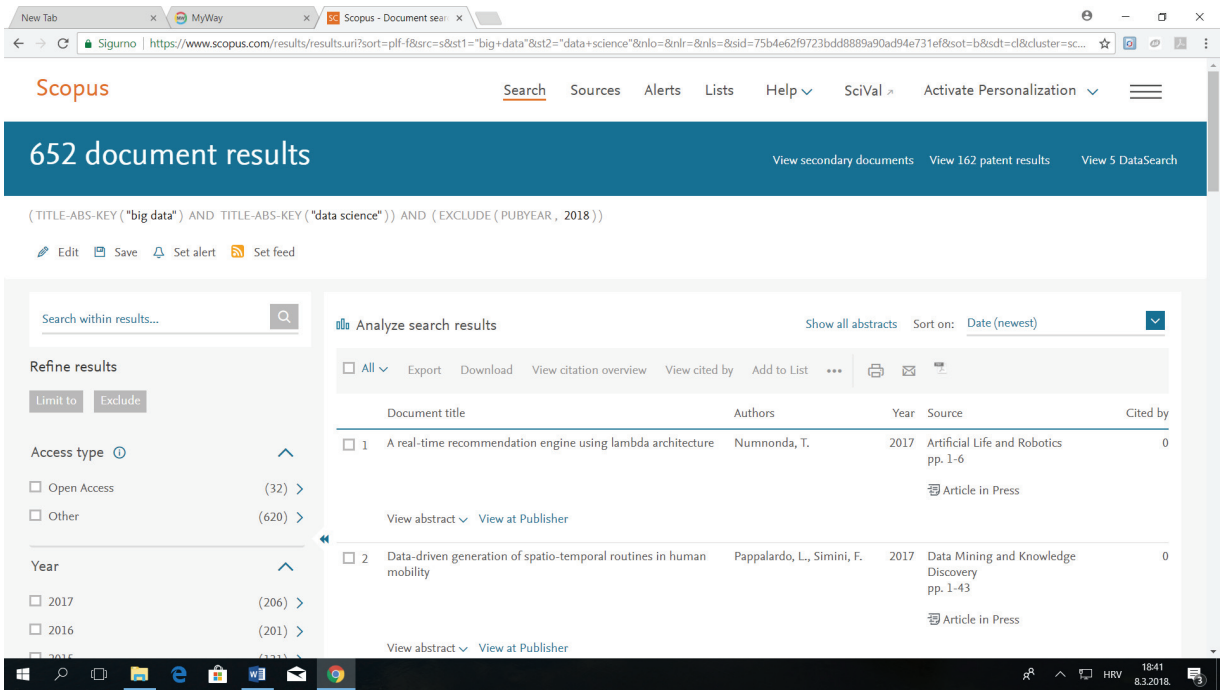
(1) Which publications publish about big data and data science the most?
(2) Which regions, institutions and authors are the most productive in the field of big data and data science?
(3) In which scientific disciplines are big data and data science employed the most?

2. METHOD

Scopus is bibliographic and citation database which indexes journals, book series and conference proceedings from all scientific areas. Scopus includes sources from all scientific areas and within Scopus database more than 130 Croatian journals are indexed. Citation data within Scopus database are available from 1996 year till today. Scopus contains more than 40 million records and even 70 percent of them have an abstract. Scopus is a product of Elsevier corporation. The method used in this paper is scientometric analysis of literature within Scopus database according to the following search strategy:

TITLE-ABS-KEY (“big data”) AND TITLE-ABS-KEY (“data science”) AND (EXCLUDE (PUBYEAR, 2018)) shown at Picture 1. Total of 652 documents according to this search strategy were found in time period from 2011 year till 2017 year (without 2018 year) and analyzed for the purpose of answering to the research questions.

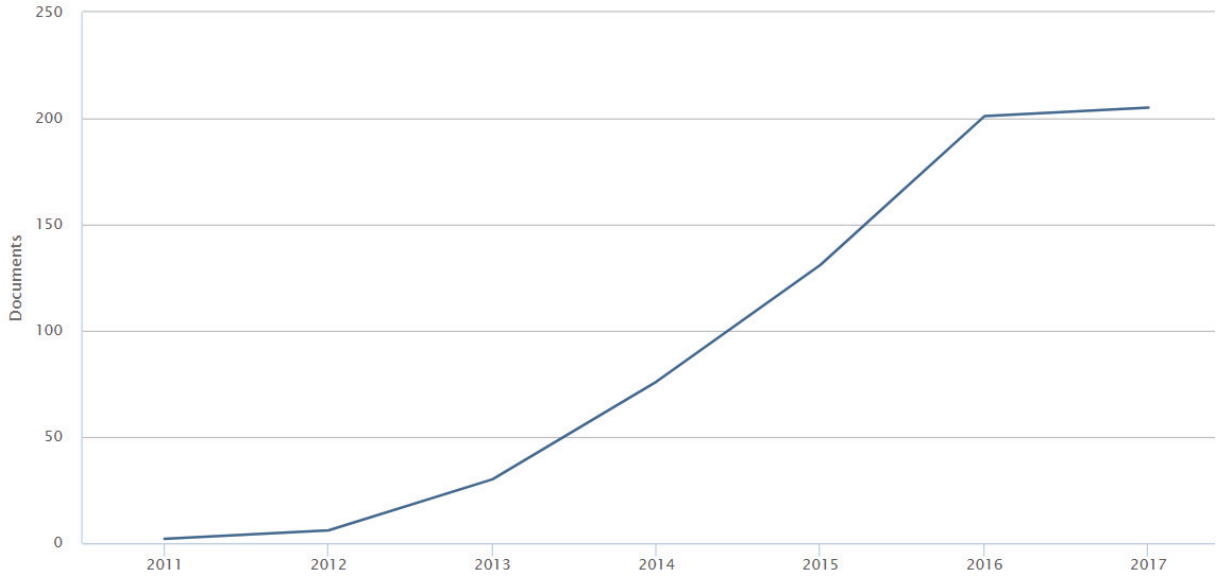
Picture 1: Search strategy TITLE-ABS-KEY (“big data”) AND TITLE-ABS-KEY (“data science”) AND (EXCLUDE (PUBYEAR, 2018))



3. RESULTS AND DISCUSSION

Big data is the emerging field which can be seen at Picture 2. Namely, within Scopus database documents about big data and data science date from 2011 not before. From 2011 year till today the number of documents enlarged for more than one hundred times. In 2011 year just 2 documents are recorded and then the exponential growth began 2012 (6 documents), 2013 (30 documents), 2014 (76 documents), 2015 (131 documents), 2016 (201 documents) and 2017 (205 documents).

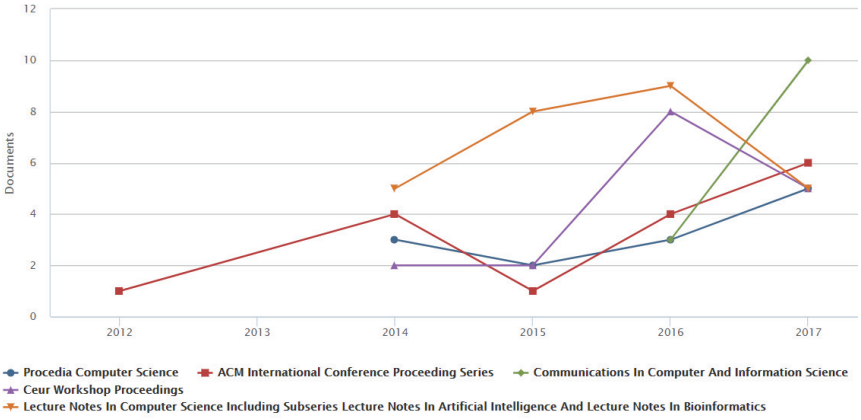
Picture 2: Documents by year about big data and data science



Source: Scopus

The publication which publishes the most on big data and data science topics is *Lecture Notes in Computer Science* (28 documents) which can be seen at Picture 3. After that follow Ceur Workshop Proceedings (17 documents), ACM International Conference Proceeding Series (16 documents), Communications In Computer And Information Science (13 documents), Procedia Computer Science (13 documents), etc.

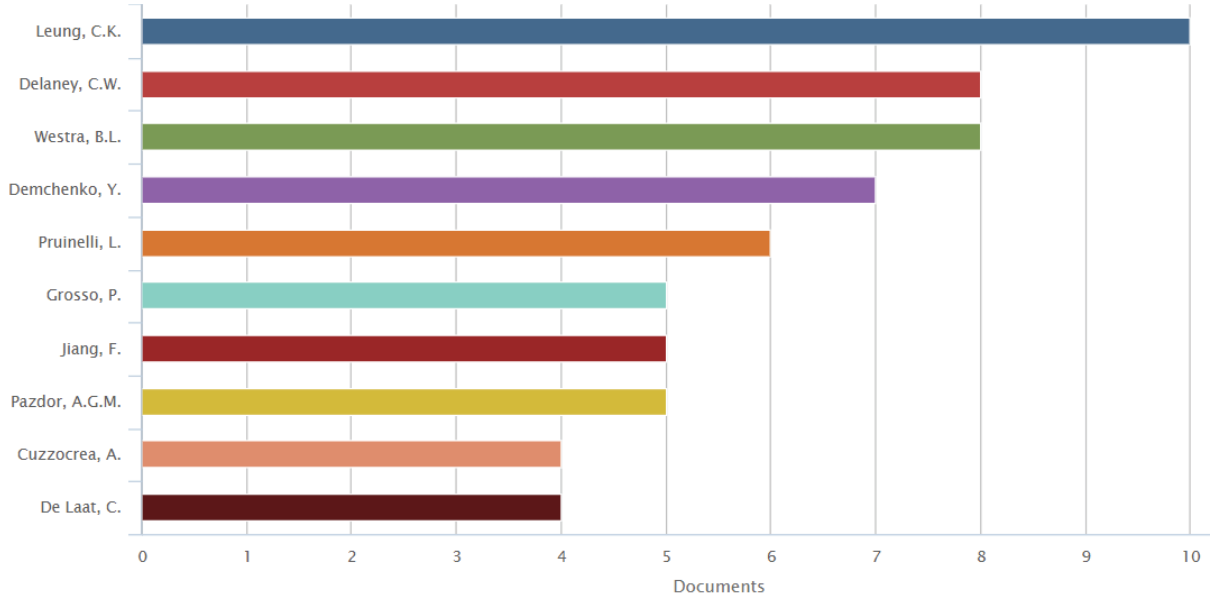
Picture 3: Documents per year by source about big data and data science



Source: Scopus

Picture 4 shows the most productive author in regard to big data and data science topics Carson K. Leung from University of Manitoba, Winnipeg, Canada with 10 published documents.

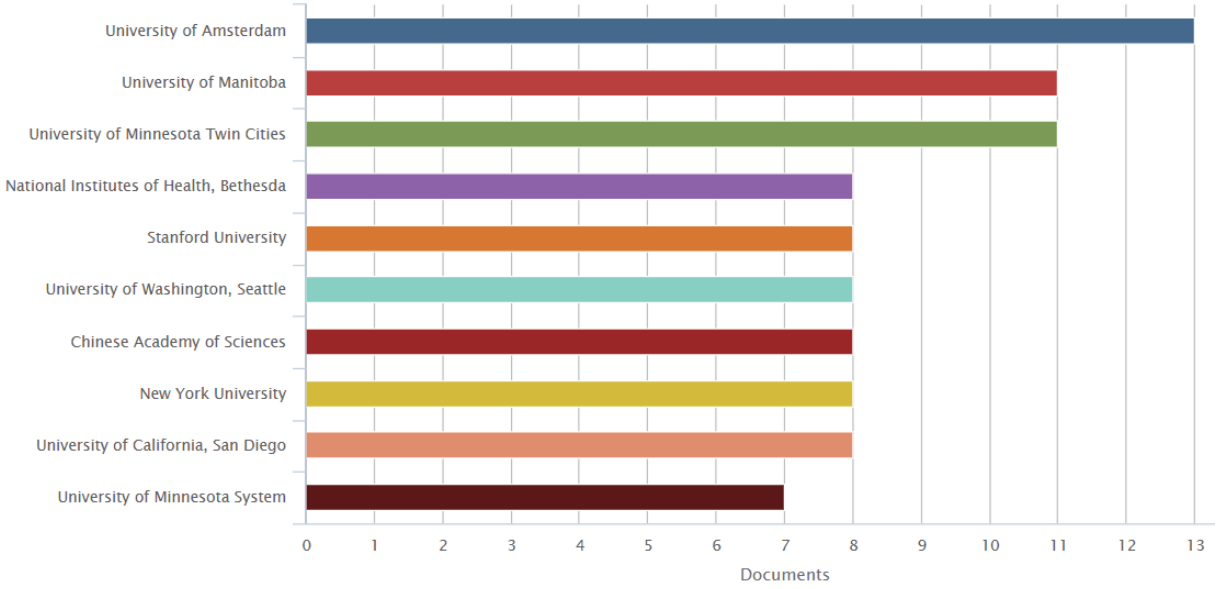
Picture 4: Documents by authors about big data and data science



Source: Scopus

Picture 5 shows that University of Amsterdam in Amsterdam, Netherlands is the most productive institution in regard to published documents on big data and data science topics.

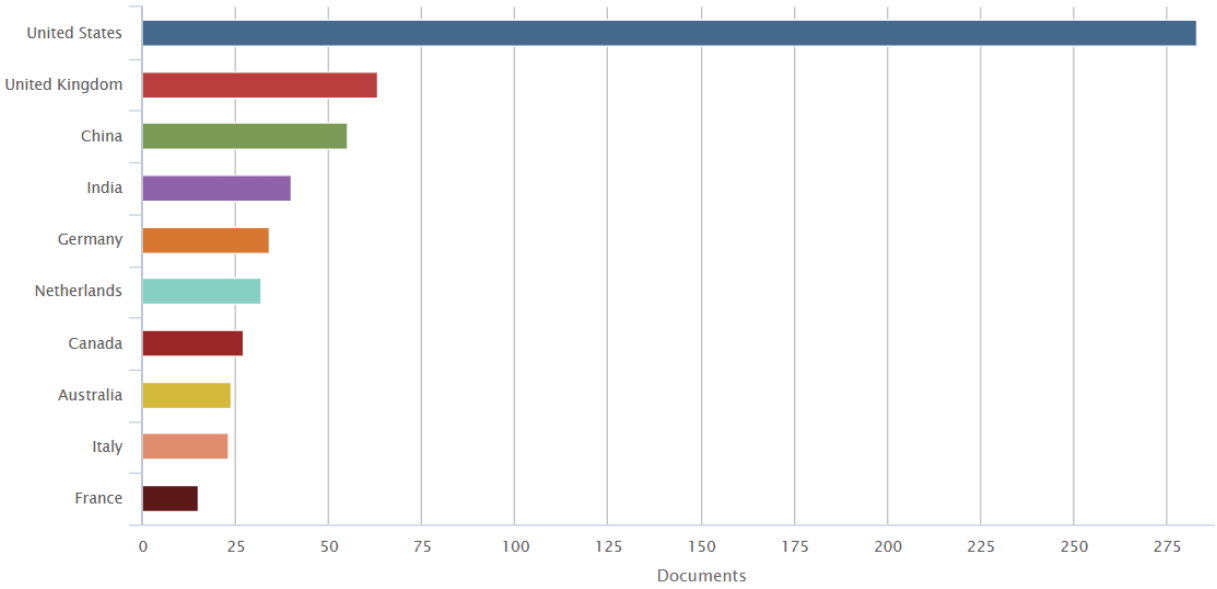
Picture 5: Documents by affiliation of authors about big data and data science



Source: Scopus

Picture 6 shows that the most productive region in regard to big data and data science is United States.

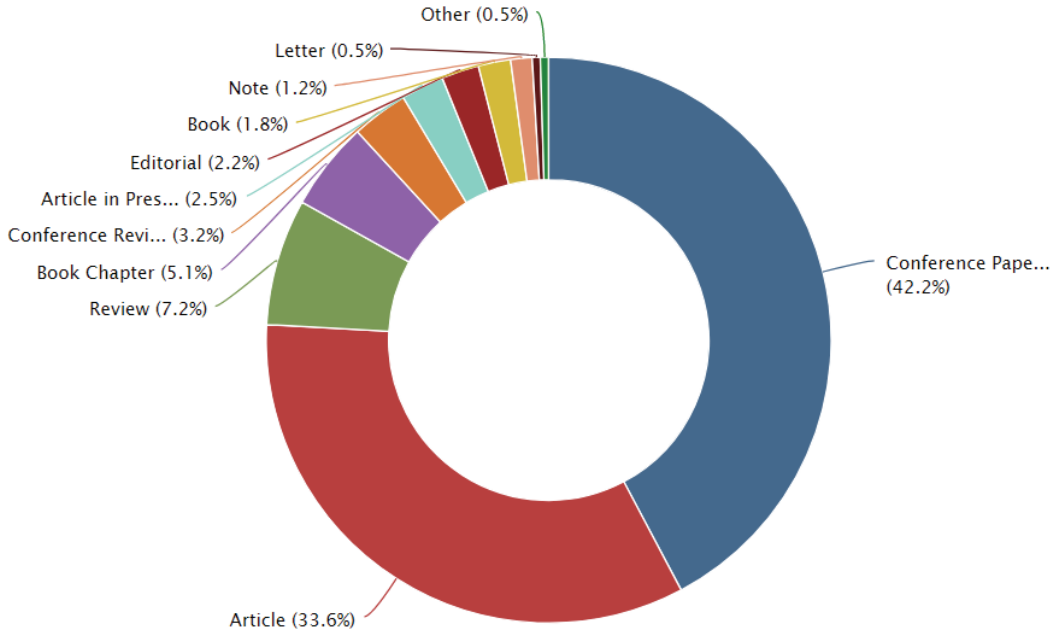
Picture 6: Documents by country about big data and data science



Source: Scopus

At Picture 7 it can be seen that the majority of papers on big data and data science are published as conference papers (42.2%) and articles in journals (33.6%).

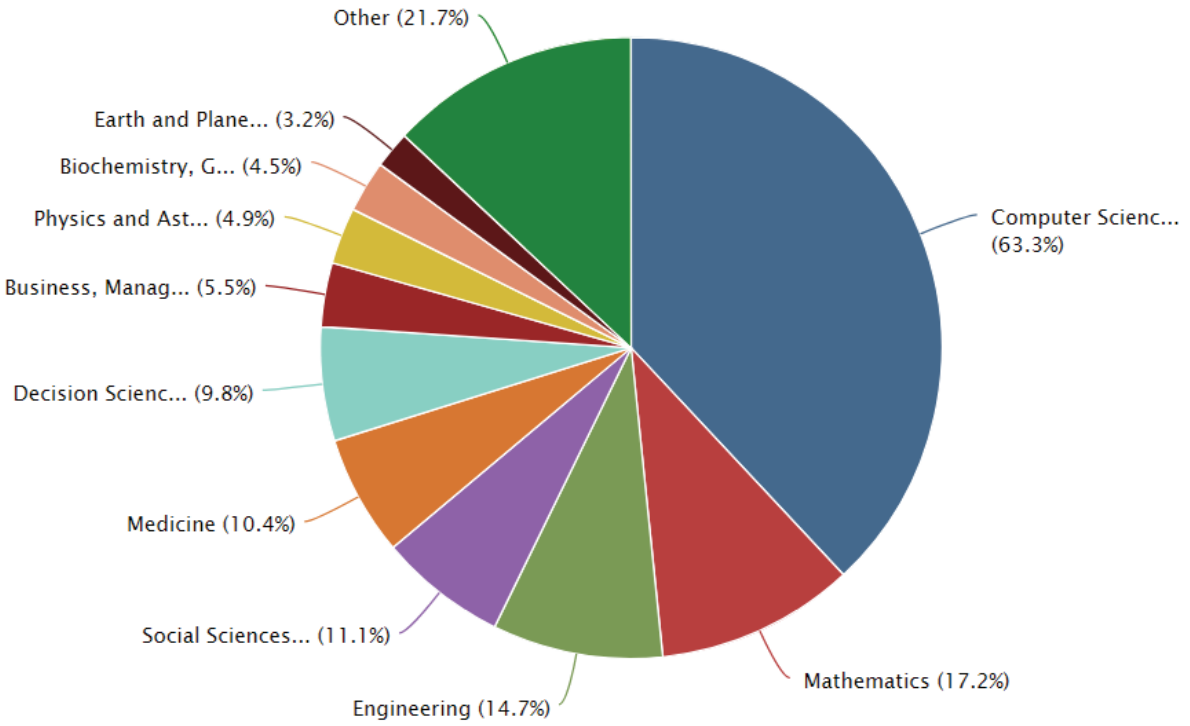
Picture 7: Documents by type about big data and data science



Source: Scopus

At Picture 8 can be seen that scientific disciplines which employ big data and data science the most are computer science (63.3%), mathematics (17.2%) and engineering (14.7%).

Picture 8: Documents by subject area about big data and data science



Source: Scopus

4. CONCLUSION

Big data could help us to better understand physical and biological systems as well as social and economic systems. Nowadays every organization which wants to attract community is confronted with issue how to more efficient use their own data but also all other available data. Manipulation with big data demands special skills and tools. Business intelligence (BI) technologies can do big data analytics in purpose to identify and develop new strategic business opportunities. The aim of business intelligence is to enable interpretation of big data. Identification of new possibilities and efficient strategy conduction based upon insights in big data can companies provide comparative advantage within the market and long-term stability. According to the obtained research results big data is the emerging field. Namely, within Scopus database documents about big data and data science date from 2011 not before. From 2011 year till today the number of documents enlarged for more than one hundred times. The publication which publishes the most on big data and data science topics is *Lecture Notes in Computer Science*. The most productive author in regard to big data and data science topics is *Carson K. Leung* from University of Manitoba, Winnipeg in Canada. *University of Amsterdam* in Amsterdam, Netherlands is the most productive institution in regard to published documents on big data and data science topics. The most productive region in regard to big data and data science is *United States*. The majority of papers on big data and data science are published as *conference papers* (42.2%) and *articles in journals* (33.6%). The scientific disciplines which employ big data and data science the most are *computer science* (63.3%), *mathematics* (17.2%) and *engineering* (14.7%). This scientometrics research provides indicators for management of science and business in regard to the emerging fields of big data and data science.

REFERENCE LIST

1. Kohavi, R. & Provost, F. (1998). Glossary of Terms. *Machine Learning* 30 (2-3), 271-274.
2. Peng, R., & Matsui, E. (2017). *The Art of Data Science: A Guide for Anyone Who Works with Data.* Skybrude Consulting, LLC.
3. Scopus. (2018, February 22). Retrieved from https://www.elsevier.com/solutions/scopus
4. Stanton, J. (2013). *Introduction to data science*. Portions.
5. Thuraisingham, B. A. Primer for Understanding and Applying Data Mining. (2018, February 22). Retrieved from https://www.utdallas.edu/~bxt043000/Publications/Journal-Papers/DS-DM/J71_A_Primer_for_Understanding_and_Applying_Data_Mining.pdf
6. Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making:* Data, information and knowledge. Wiley.