IOSR Journal of Mobile Computing & Application (IOSR-JMCA)
e-ISSN: 2394-0050, P-ISSN: 2394-0042.Volume 3, Issue 4 (Jul. - Aug. 2016), PP 52-61
www.iosrjournals.org

Data Scientist Skills

Nur Amie Ismail, Wardah Zainal Abidin
(Advance Informatics School, University of Technology Malaysia, Kuala Lumpur)

Abstract: *Decision making is one of the most important aspects in order to enhance service delivery to citizens and businesses, gain more profit, and help stakeholders to strategize their business functions. Nowadays, most of the stakeholders make decisions based on the data that is precise, concise, appropriate, and accurate. Even though Big Data Analytic (BDA) tools and software can assist in this matter, skills and competency of the personnel that handle and manage the data is more crucial and important. Thus, the aim of this paper is to identify data scientist skills from global best practices and examine the most important data scientist skills required by Information Technology (IT) personnel. From our findings, we found 44 data scientist skills and the top 5 (five) skills are business, statistic, machine learning, communication, and analysis.*
Keywords: *Data Science, Data Scientist, Data Scientist Skill*

I. Introduction

Nowadays with the vast amounts of data available in the world, companies across industry are focusing on exploiting data for their competitive advantage. Hence, they realized that they need to hire more data scientists or equip their employees with data scientist skills. Data scientist is an expert who is capable to extract meaningful value from the data and also manage the whole lifecycle of data [1]. Data scientists also help to bridge the communication gap between business and IT functions, proposing meaningful measures, modelling the data, visualizing the output, sharing the technique, and automating the process [2]. According to McKinsey Global Institute, in United Stated of America (USA) alone, they need another 140 - 190 thousand of data scientist by 2018. Whereas in Malaysia, Multimedia Development Corperation (MDeC) have a set an ambitious target to produce 1,500 data scientists by 2020. According to [3], currently in Malaysia there are only eighty (80) data scientist across the country. In order to increase number of data scientist, various programs have been arranged such as Big Data conference, trainings, certification, and Massive Open Online Courses (MOOC). However, programs that have been arranged still insufficient to cater for meeting the ambitious target. Thus, this paper will identify data scientist skills from global best practices and examine the most important data scientist skills required by IT personnel in order to be recognized as Data Scientist.

This paper is organized as per following sequence. Section 2 explained review methodology that has been used in this study. Section 3 briefly explained data science definition, data science fundamental concept, and the difference between Data Scientist and Data Analyst. Section 4 discusses the data scientist skills from global best practices followed by section 5 about finding. Finally, the conclusion and future works is in Section 6.

II. Review Methodology

To get clear and better understanding on the research topic, research review has been conducted from various resources such as books, articles, journals, and web sites. List of computerized databases used in this paper are Association for Computing Machinery (ACM) Digital Library, IEEE, Science Direct, Springer Link, Wiley Online Library, ERIC, Gartner, and Google Scholar. The resources that are reviewed are within the period of 2011 to 2016. Fig. 1 below shows the review methodology that has been used in this paper:

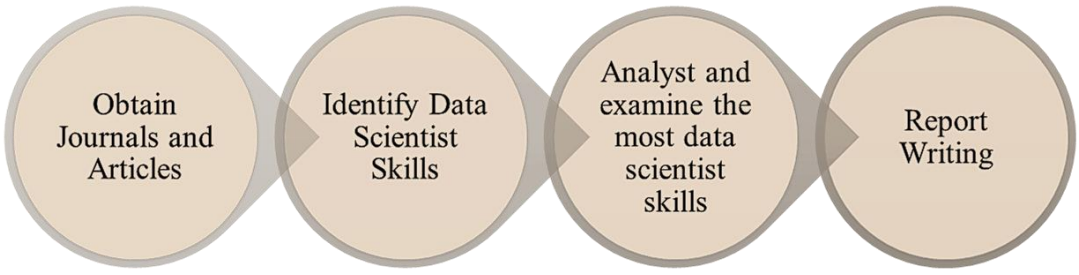

Fig. 1: Research review methodology

III. Data Science Definition, Data Science Fundamental Concepts And Difference Between Data Scientist And Data Analyst

3.1 Data Science Definition

There are several definitions of data scientists from several authors as listed in the table 1.

Table 1: Data Science Definitions

No.	Definition
1	Set of fundamental principles that support and guide the principle extraction of information and knowledge from data [4].
2	Data science is the study of the generalizable extraction of knowledge from data [5].
3	Data science is a combination of statistic, computer science, and information design [2].

From table 1, we can summarize that, data science is combination of field of study related to extraction and transformation of data.

3.2 Data Science Fundamental Concepts

According to [2], the fundamental concept of data science is extracting useful knowledge from data to solve business problems that can be treated systematically by following a process with reasonably well-defined stages. Data-science results requires careful consideration of the context in which they will be used in the relationship between the business problem and the analytics solution. This often can be decomposed into tractable sub problems via the framework of analyzing expected value. IT can be used to find informative data items from within a large body of data. Other than that, entities that are similar with respect to known features or attributes often are similar with respect to unknown features or attributes, data might not generalize beyond the observed data and to draw causal conclusions, an attention to the presence of confounding factors possibly unseen ones.

3.3 Difference Between Data Scientist And Data Analyst

According to [6], Data Analyst focus on the movement and interpretation of data, typically focus on the past and present. Where Data Scientist focus on summarizing data and to provide forecasting based on pattern identified from past and current data. [7] define and differentiate between Data Scientist and Data Analyst as describe in table 2.

Table 2: Data Scientist vs Data Analyst

Data Scientist	Data Analyst
Building statistical models that make decisions based on data. Each decision can be hard, e.g. block a page from rendering, or soft, e.g. assign a score for the maliciousness of a page that is used by downward systems or humans.	Writing custom queries to answer complex business questions.
Conducting causality experiments that attempt to attribute the root cause of an observed phenomenon. This can be done by designing A/B experiments or if A/B experiment is not possible apply epidemiological approach to the problem.	Conceiving and implementing new metrics on capturing previously poorly understood parts of the business / product.
Identifying new products or features that come from unlocking the value of data; being a thought leader on the value of data. A good example of that is the product recommendations feature that Amazon first made available to a mass audience.	Addressing data quality issues, such as data gaps or biases in data acquisition. Working with the rest of engineering to instrument incremental new data acquisition.

IV. Data Scientist Skills

In order to explore the list of data scientist skills, this study has a global reach and perspective as well includes the Malaysian public sector. Basically the data scientist incorporates advanced analytical approach using sophisticated analytic and data visualization software or tools in order to discover patterns of the data. The data scientist then will do data migration and integration, data cleaning, analyzing and deliver the outcomes. According to [8], the data scientist must be able to write in different programming language such as Python, R, Java, Ruby, Clojure, Matlab, Pig, and SQL. Other than that, the data scientist need to understand about Hive, Hadoop, and Map Reduce. They also suggest that the data scientist must be familiar with Natural Language Processing (NLP), machine learning, conceptual modelling, statistical analysis, predictive modelling, and hypothesis testing. Even

though the data scientist has to learn new skills as explained above, at least they should have the capabilities in communication skills, querying the database, understand about business strategy, able to design simple prototype for top management, and have good understanding in system architecture.

Educational data scientist is rarely sighted breed especially within business and government. In order to tackle this scenario, we need to produce more graduates and also equip the employees with necessary skills in data science. [2] suggest that the data scientist should have skills in data mining, data modelling, data visualization, and machine learning. According to [9], the data scientist uses advanced analytics such as predictive analysis, data visualization and modelling, and machine learning to predict what is going to happen in the future and give recommendations to enhance existing business process. They also defines that the data scientist is a combination of three(3) main fields which are computer science, statistics, and domain knowledge.
Fig. 2 shows the relationship and skills for each area.

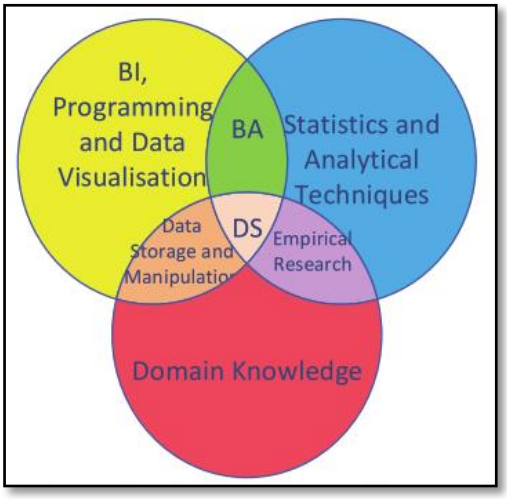

Fig. 2: Data scientist skills (Ayankoya et al., 2014)

[5] emphasizes that machine learning is the most important skill and necessary for all data scientists. In machine learning, the data scientist should master of all 3 class of skills as illustrated in table 3 as per below. Other than machine learning, data scientist also required knowledge in text mining, markup language like XML, mathematics, and artificial intelligence (AI).

Table 3: Three (3) class of skill in Machine Learning

No.	Class	Skills
1	Statistic	Bayesian statistic and probability.
2	Computer Science	Data structure, algorithm and distributed computer.
3	Correlation and Causation	Modelling.

Data scientist is an expert that has the ability to manipulate and extract knowledge and turn it into meaningful value [1]. According to their study, currently there is no accepted and effective data science professional curriculum. [10] found that the two (2) top skills companies are looking for in a data scientist are programming and statistical. The details of these two (2) skills illustrate in fig. 3.

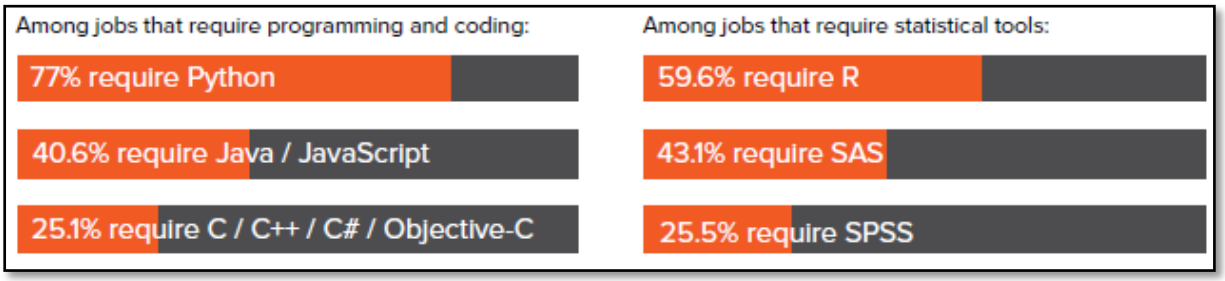

Fig. 3: Programming and Statistics

According to [11], data science is now being benchmarked against practices that employed on highly skilled professionals. Data Scientist uses scientific methods to discover knowledge and patterns of the data.

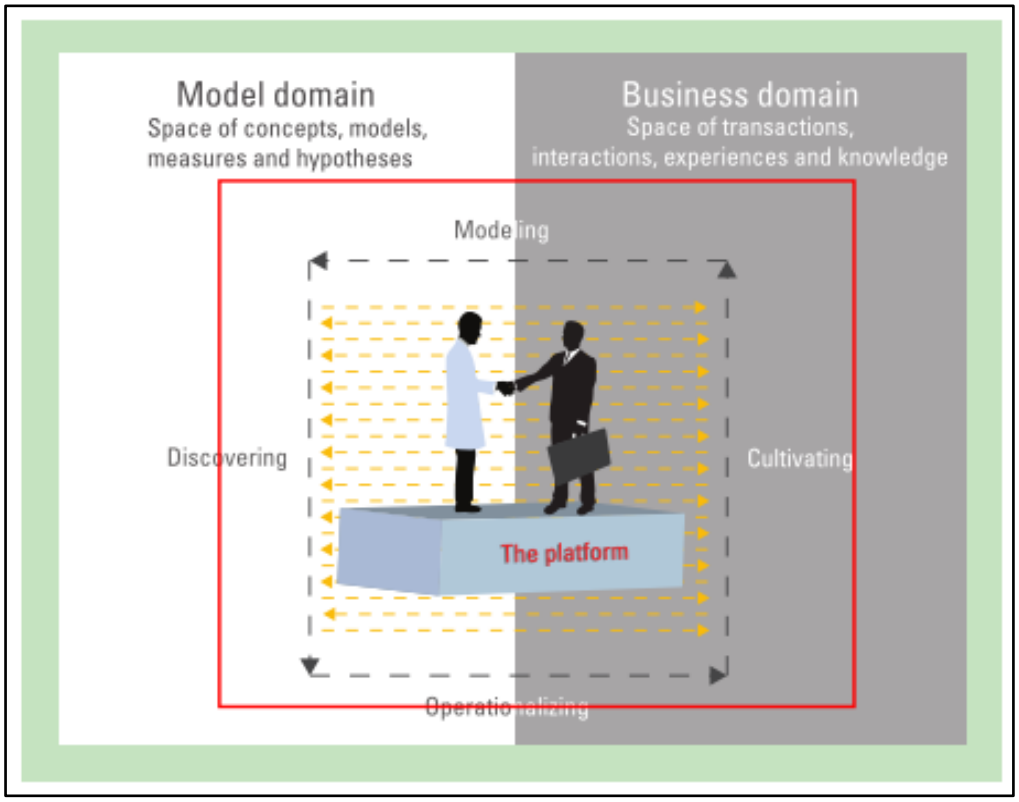

Fig. 4: The data science benefits-realization process(Viaene, 2013)

Fig. 4 illustrates on how to use data to improve the business. This process involves modelling, discovering, operationalizing, and cultivating the knowledge. The data scientist must have a pretty good skill in business domain, analysis, and communication. While, according to [12], data scientist is the sexist job in this century. Sexy in the sense of having a rare quality in high demand. Data scientist is urgently needed by organizations because they know how to use the analysis of big data to make effective decisions. Among the skill that they should consider are programming language, computer science, mathematics, economics, probability, and business. In the O'Reilly book, Analysing the Analysers by [13], they have made a survey over more than 200 data scientists to discover and analyze what data skills needed by the data scientist. They found 22 generic skills shown in fig. 5.

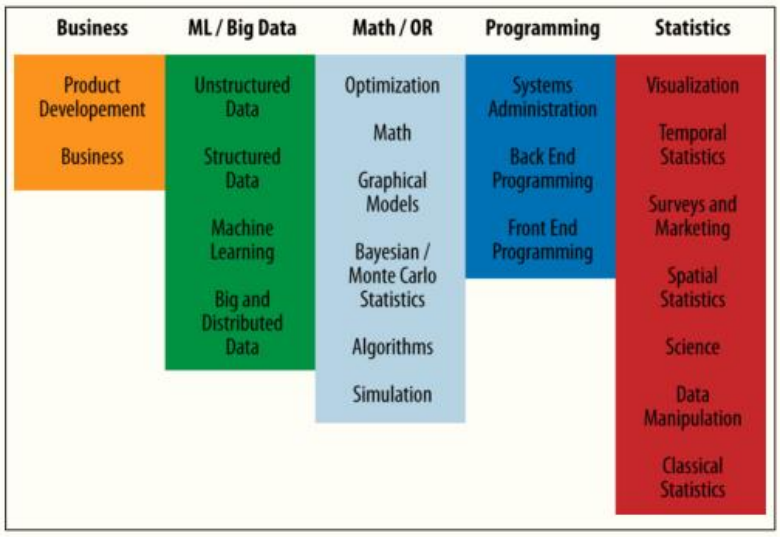

Fig. 5: Data Scientist generic skills

Malaysia Public Sector have started their BDA project since 2014 led by MAMPU. Since data science is still new in Malaysia, they do not have internal expertise in this area. Therefore, they have hired external consultants to develop the project. Even though they use external parties; knowledge transfer, training, and

technology updates are given to Government IT officers. Soon, Malaysian Government do realize that the importance of having internal expertise in this field. Hence, skills of data scientist are identified to enhance Government IT officer competency and knowledge. According to [14], the skills that are required for data scientist consist of model and analysis, data processing, statistic, business domain, soft skill, and technical skill as illustrated in Fig. 6.

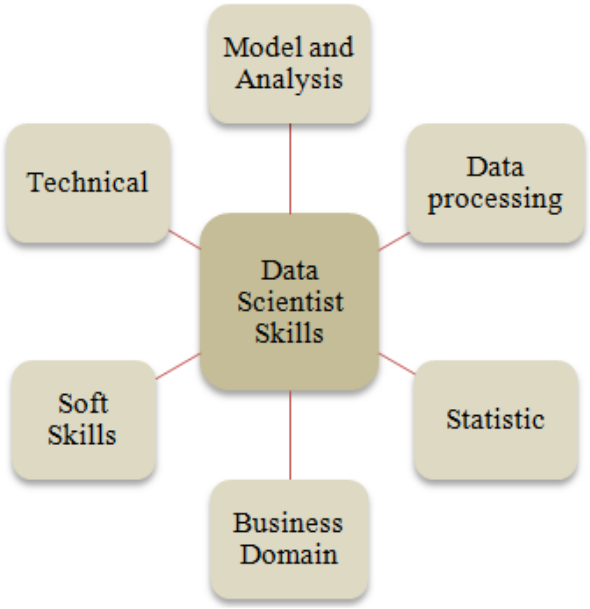

Fig. 6: Data scientist skill (Suhailis, 2016)

In 2013, [15] have announced the Digital Malaysia Roadmap, which encompasses a plan that addresses three ICT areas which are to access, adoption and usage ICT services. One of the goals in the roadmap is to improve Big Data literacies in Malaysia. Therefore, in October 2013, MDeC have conducted a survey to 17 experts in Big Data. The participants come from different background such as telecommunication company, universities, marketing agency, software development companies, and others. Based on their survey, the top five skills needed are:

(i) Big and Distributed Data (eg: Hadoop, MapReduce)
(ii) Algorithms (eg: computational complexity, CS theory)
(iii) Machine Learning (eg: decision trees, neural nets, SVM, clustering)
(iv) Back-End Programming (eg: JAVA/Rails/Objective C)
(v) Visualization (eg: statistical graphics, mapping, web-based dataviz)

In the last few years, the interest in data science field has soared. Most of the companies in USA are seeking and recruiting employees who have skills related to data science. From the perspective of [16], she emphasizes that the data scientist must have both technical skill and non-technical as listed in the table 4 below:

Table 4: List of skill needed to recruit employee in data science

No.	Type of skill	Skills
1	Technical Skills	Analytics**,** SAS, R, Python, Coding, Hadoop, SQL, and Database.
2	Non-Technical Skills	Intellectual curiosity, business acumen, and communication skills.

In United Kingdom, data science is among the most rapidly emerging field based on trend in ICT market. The key to success in business nowadays is to understand customer's preferences, needs, and behavior. Thus, data scientist plays an important role to do a prediction and make decision in this particular area. [17] concludes that data scientist need multi-faceted skills illustrated in fig. 7.

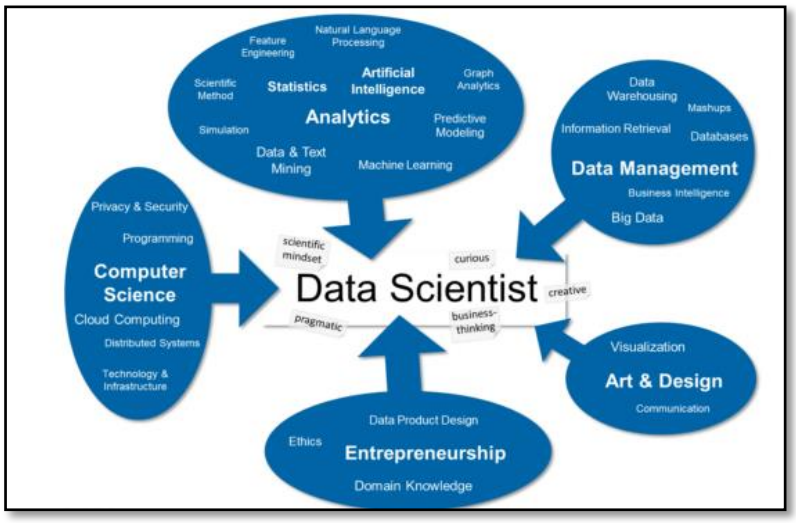

Fig. 7: Data Scientist Skills (Stadelmann et al., 2013)

In Japan, the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) has initiated a three-year project namely "Data Science Training Network" to develop the data scientists. This projects involves various stakeholders such as from the universities, the industry, and the Government. The Government of Japan realize that data science is an important area in order to increase their efficiency, generate more income, make prediction, and assist in decision making. Based on finding by [18], the constraint of employer is to find talented and highly skilled in data science. The result obtained from the first 12 months of the project revealed that to become successful data scientist, the mandatory skills are:

(i) Deep analytics skills: Machine learning, database, and programming.
(ii) Service providing skills: Communication skills and business.
(iii) Service receiving skills: Decision making.

In India, a professional is well equipped with software and tools to assist and accomplish their tasks in the office such Business Intelligence (BI) and expert system [19]. This software is widely used to help the management to strategize their business vision and mission, learn from previous trend and pattern of data, and also prevent damage and error. He also lists some of the skills required once the employee enter data science area. The skills are R programming, Python, Java, Ruby, Hadoop, Analysis, Data Mining, Machine Learning, and Statistic. Gartner, Inc. is an American research and advisory firm that provides ICT updates and best practices. Basically, best practice is defined and provided by Gartner for the purpose of benefitting in terms of efficiency optimization, reduce costs and risks, and enhance the effectiveness in the organization. Gartner has released an article that explains the relationship between IT skill, domain understanding, and data science shown in fig.8. According to [20], to avoid failure in Big Data project, team member must possess different skills through some programs like training and hands-on that can extend their current experience.

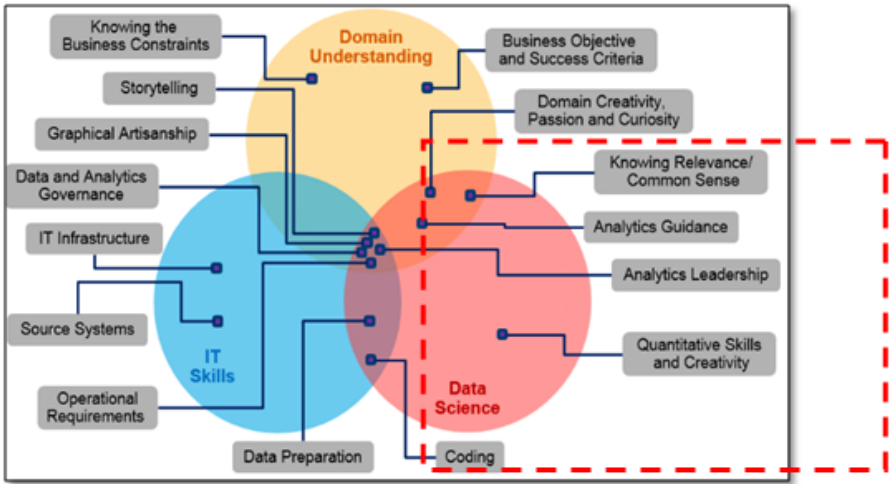

Fig. 8: Big Data Analytics Skill Model (Gartner, 2015)

Creativity, leadership, common sense, passion, and curiosity equally important with the technical skills and basically complimenting each other.

[21] reviews some of the skills required in order to become a data scientist. The most and top skill is knowing business strategy and function of the organization. Usually, IT personnel can easily adapt and learn new skills but lacking in term of aligning business with IT. Secondly they should know about the statistic. With some statistical technique, problem can be classified, translated thus providing with the recommendations. Some of the tools for statistic are R programming and SAS. Besides that, data scientist should have the capability to write in different programming languages like Java, Python, C++, and C #. Other than having programming skills, mastery in database also important. In database, they must know how to integrate, migrate, and load the data. Some of emerging tools in database are Hadoop, Hive and Mahoot. Last but not least, visualization and communication skills is important because it enables those who aren't professional data analysts to interpret data.

[22] emphasized that data scientist is broadly applied within different organizations making it difficult to provide a complete and non-controversial list of required skills. [22] suggested at high level, a data scientist needs a mastery in data warehousing, data analysis, data transformation, and communication skills.

V. Findings And Discussion

From the research review, this study found 44 skills from18 papers which is includes technical and non-technical skills as portray in Table 5. Whereas table 6 shown data scientist skills that categorized into five (5) IT domain areas adapted from [17].

Table 5: Data Scientist Skills

No	Skill	Frequency Skills Appear in the papers	Percentage (%)
1	Business	11	61.11
2	Statistic	10	55.56
3	Communication	9	50.00
4	Machine Learning	9	50.00
5	Analysis	8	44.44
6	Programming (General)	7	38.89
7	Data Modeling	5	27.78
8	Hadoop	5	27.78
9	Database (General)	4	22.22
10	Python	4	22.22
11	R Programming	4	22.22
12	Data Visualisation	3	16.67
13	Java	3	16.67
14	Mathematic	3	16.67
15	Natural Language Processing	3	16.67
16	SQL	3	16.67
17	Algorithm	2	11.11
18	Business Intelligence	2	11.11
19	Data Mining	2	11.11
20	Hive	2	11.11
21	Map reduce	2	11.11
22	Probability	2	11.11
23	SAS	2	11.11
24	Simulation	2	11.11
25	Other soft skills:	2	11.11
26	Artificial Intelligence	1	5.56
27	C/C++/C#	1	5.56
28	Clojure	1	5.56
29	Cloud computing	1	5.56
30	Computer Science	1	5.56
31	Data Manipulation	1	5.56
32	Data Transformation	1	5.56
33	Data Processing	1	5.56
34	Data warehousing	1	5.56

35	Decision making	1	5.56
36	Distributed System	1	5.56
37	Economic	1	5.56
38	Ethics	1	5.56
39	Mahoot	1	5.56
40	Matlab	1	5.56
41	Pig	1	5.56
42	Privacy and Security	1	5.56
43	Ruby	1	5.56
44	System architecture	1	5.56

Table 6: List of data scientist skill based on IT Domain Area

No.	IT Domain Area	Skills
1	Computer Science	i. C/C++/C# ii. Cloud computing iii. Distributed System iv. Java v. Programming(general) vi. Privacy and Security vii. R Programming viii. System architecture ix. Clojure x. Hadoop xi. Mahout xii. Pig xiii. Python
2	Analytics	i. Algorithm ii. Artificial Intelligence iii. Analysis iv. Machine Learning v. Mathematic vi. Matlab vii. Map reduce viii. Natural Language Processing ix. Probability x. SAS xi. Statistic xii. Simulation
3	Data Management	i. Business Intelligence ii. Database iii. Data Mining iv. Data Manipulation v. Data Modelling vi. Data Visualization vii. Data Transformation viii. Data Processing ix. Data Warehousing x. SQL xi. Hive
4	Art and Design	i. Communication ii. Decision making iii. Ethics
5	Entrepreneurship	i. Business ii. Economic

VI. Conclusion

From analysis of this study, the top five (5) skills are business, statistic, machine learning, communication, and analysis as shown in fig. 9.

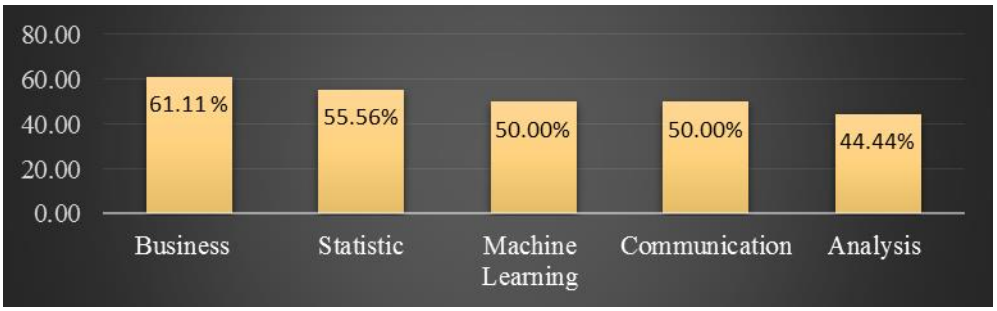

Fig. 9: The top five (5) data scientist skill

Based on fig. 9, business has the highest percentage (61.11%) of frequency appeared among the papers. The data scientist should understand their business objectives, environment, and strategies so that they know where and how to maximize usage of data in the organization. The second highest is statistic (55.56%). Basically, statistic is used to design and interpret experiments, build models, and make prediction. Then, machine learning is in the third place (50.00%). Machine learning is a method of data analysis that automates analytical model building. Machine learning allows computers to find hidden insights without being explicitly programmed where to look. Other than technical skill, non-technical such as communication (50.00%) also the characteristics or skills required in this field. This skill will assist the data scientist to understand stakeholders, to lead in decision making process, and get retention. Finally, skill in analysis (44.44%). Out of 44 skills, other non-technical skills such as economic and ethics also emphasized by some authors.

This study has aimed to identify data scientist skills from global best practices and examine the most important data scientist skills required by IT personnel. As a conclusion, even though there are many skills required by IT personnel in order to become a good data scientist, they have to make sure it is aligned with their organization needs and purposes. For future work, we recommend other researchers to explore and develop full set of data scientist curriculum. The curriculum will be a guideline and succession planning in order to prepare experts in data science.

References

[1] Manieri, A., Demchenko, Y., Brewer, S., Hemmje, M., Riestra, R., & Frey, J. (2015). Data Science Professional uncovered How the EDISON Project will contribute to a widely accepted profile for Data Scientists. In 2015 IEEE 7th International Conference on Cloud Computing Technology and Science Data. doi:10.1109/CloudCom.2015.57

[2] Shum, S. B., Hall, W., Keynes, M., Baker, R. S. J., Behrens, J. T., Hawksey, M., & Jeffery, N. (2013). Educational Data Scientists : A Scarce Breed. Retrieved from http://simon.buckinghamshum.net/wp-content/uploads/2013/03/LAK13Panel-Educ Data Scientists. pdf

[3] Patrick, S. (2015). Malaysia needs 1,500 data scientists by 2020. Retrieved from http://www.thestar.com.my/tech/tech-news/2015/04/24/data-scientists-needed-to-make-sense-of-the-numbers/

[4] Provost, F., & Fawcett, T. (2013). Data Science Its Relationship Data-Driven Decision Making, 1(1), 51–59. doi:10.1089/big.2013.1508

[5] Dhar, V. (2013). Data Science and Prediction, 56. doi:10.1145/2500499

[6] Perumal, S. (2015). Data scientist. Retrieved from http://www.slideshare.net/SevugaPerumal1/a-free- orientation-on-statistical-data-analysis-is-conducted-on-saturday-25072015-at-10-am-and-it- has-2-hours-duration

[7] Boulis, K. (2014). What is difference between a data analyst and a data scientist? Retrieved from https://www.quora.com/What-is-difference-between-a-data-analyst-and-a-data-scientist

[8] Soumendra Mohanty, M. J. and H. S. (2013). Big Data Imperatives Enterprise Big Data Warehouse, BI Implementations and Analytics. Apress.

[9] Ayankoya, K., Box, P. O., Calitz, A., Box, P. O., Greyling, J., & Box, P. O. (2014). Intrinsic Relations between Data Science , Big Data , Business Analytics and Datafication, 192–198. doi:10.1145/2664591.2664619

[10] CrowdFlower. (2015). 2015 Data Scientist Report.

[11] Viaene, S. (2013). Data Scientists Aren't Domain Experts. IEEE Compu ter Society.

[12] Patil, T. H. D. and D. . (2012). Data scientist the sexiest job of the 21st_century. Harvard Business Review.

[13] Harlan D. Harris, Sean Patrick Murphy, and M. V. (2013). Analyzing the Analyzers: An Intro Survey of Data Scientist and Their Work. (M. Loukides, Ed.) (First Edit). United States of America: O'Reilly. Retrieved from http://oreilly.com/catalog/errata.csp?isbn=9781449371760

[14] Suhailis, A. (2016). Garis Panduan Data Raya Sektor Awam.

[15] MDEC. (2014). Big Data in Malaysia : Emerging Sector Profile.

[16] Burtch, L. (2014). 9 Must-Have Skills You Need to Become a Data Scientist. Retrieved fromhttp://www.kdnuggets.com/2014/11/9-must-have-skills-data-scientist.html

[17] Stadelmann, T., Stockinger, K., Braschler, M., Cieliebak, M., Baudinot, G., & Ruckstuhl, A. (2013).

[18] Applied Data Science in Europe Challenges for Academia in Keeping Up with a Highly Demanded Topic.

[19] Maruyama, H. (2013). Developing Data Analytics Skills in Japan : Status and Challenge, 1–6.

[20] Retrieved from https://datascientist.ism.ac.jp/
[21] Gehl, R. W. (2015). Sharing , knowledge management and big data : A partial genealogy of the data scientist. doi:10.1177/1367549415577385
[22] Sicular, S. (2015). Big Data Analytics Failures and How to Prevent Them, 1(August).
[23] Rao, A. (2014). The 5 Dimensions of the So- Called Data Scientist.
[24] Polich, K. (2016). How to hire for the right big data skill set.