VOL 1 (2017) NO 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Data scientists' skills in detecting archetypes in Iran

Hamideh Iraj[#], Babak Sohrabi[#]

[#]*Faculty of Management, University of Tehran, Jalal ale ahmad highway, Chamran, Tehran, 1411713114, Iran*
E-mail: hamideh.iraj@ut.ac.ir, bsohrabi@ut.ac.ir

Abstract*— The use of data-driven decision making and data scientists is on the rise in Iran as companies have rapidly been focusing on gathering data and analyzing it to guide corporate decisions. In order to facilitate the process and understand the nature and characteristics of this transformation, the current study intends to learn about data scientists' skills and archetypes in Iran. Detecting skills archetypes has been done via analyzing the skills of data scientists which were self-expressed through an online survey. The results revealed that there are three archetypes of data scientists including high level data scientists, low level data scientists and software developers. The archetypal patterns are based on levels of data scientists' skills rather than the type of dominant skills they possess which was the most frequent pattern in previous studies.

***Keywords*— Data science; Data scientists; Archetypes; skills**

I. INTRODUCTION

Data science is defined as the extraction of knowledge from large volumes of data. It is a newly-established field in its embryonic phase of development. It is a newly-established field in its embryonic phase of development. Since data science employs techniques and theories from many areas [1], understanding the job market plays an important role for planners, policymakers and educational institutions on the supply side and industries on the demand side. Gaining an insight from data is necessary considering the interdisciplinary nature of the field as well as the inherent ambiguity in defining skills, roles and responsibilities [2, p. 3]. To begin with, we need to understand the current situation i.e. what skills and archetypes current data scientists possess.

The job market analysis helps rein in transparency, eliminate ambiguity and create a common understanding for discussion and communication. Harris et al. mentioned some cases to elucidate how the miscommunication leads to lost opportunities both for employers and job seekers [2, pp. 1-7]. The current study used data science techniques to learn about data scientists' archetypes in Iran and guide educational policy for information technology majors. The study is the first of its kind in the country and hence; helps create a lingua franca for students and faculty members to improve the data science education.

II. RESEARCH BACKGROUND

A few studies detected data scientists' archetypes. In 2010, Liberatore and Luo found three analytical roles for data scientists: research analysts, application analysts and user analysts. Research analysts develop new models or methods for their own and other organizations. Application analysts apply and customize existing models whereas user analysts verify and interpret results [3].

Kandel et al. studied 35 data analysts from 25 organizations to know their analytics processes, tools and skill archetypes. In their qualitative study, the researchers found that data scientists' skill patterns are defined by programming and Information Technology skills rather than statistics proficiency. Hacker, scripter and application were three archetypes they discovered during the course of the research. To them, hackers are the most proficient programmers that work independently of Information Technology department in the early stage of analytics prior to modeling. Scripters model and visualize data with statistical packages. Unlike hackers, they have more statistical and less technical proficiency. The last group, application users, work with spreadsheets or analysis applications such as SAS or SPSS [4].

Harris et al. studied data science professionals' skills and self-image regardless of their titles, education and experiences. During the study, the researchers used 22 data scientists' skills and four predefined roles consisting of data developer, data researcher, data creative and data businessperson as input. Finally, they came up with five skill

groups: (1) business, (2) machine learning/big data, (3) math/operations research, (4) programming and (5) statistics [2, p. 13].

Later, Maruyama conducted a similar study in Japan. He came to the point that the data science environment in Japan differs from that of the Silicon Valley. He collected the research data through a survey and interviews and then clustered data scientists into four following groups:

1. Young aspiring data analysts with little or no experience.
2. Mid-career engineers, working in R&D of large manufacturing enterprises and as part of their daily job, they analyze data generated within their own departments. Their career path in those firms is well-defined.
3. Employees of small and medium businesses, typically service business. They choose data analytics because of the flexibility in work. This group has high female ratio compared with other groups.
4. Seasoned service professionals in an IT service company or a consulting firm. They offer data analytics services to their clients.

The results revealed that data scientists in Japan are segmented based on the type of their jobs (whether working on the data of their own organization or those of customer's) and also their long-term career paths whereas in the United States, they are recognized by their activities and skills usage. Furthermore, data scientists' jobs in Japan are service-oriented whereas in the Silicon Valley, they are product-oriented. As such, Maruyama concluded that data scientists in Japan need analytics, service providing and service receiving skills [5]:

Accenture Company, in its 2013 report, mentioned five roles for a data science team including software engineer, systems architect, quantitative analyst, business analyst and, visualization designer. These roles are based on skills such as advanced analytics, business acumen, communication & collaboration, creativity, data integration, data visualization, software development and systems administration (Harris, Shetterley, Alter, & Schnell, 2013).

As mentioned earlier, the current study intends to explore data scientists in Iran to detect their skill archetypes and compare them with the findings of previous researchers. The skill archetypes bring in new insight for cultivating data scientists and help in mutual understanding and improving communication.

III. Research Methodology

Considering the statistical nature of the study, the current research has followed a quantitative approach. An unsupervised algorithm was used to cluster data scientists based on their skills. By unsupervised, the researchers mean there is no label for the data being analyzed. . [6, p. 445] In other words, skills archetypes are discovered from the scratch.

3.1 Research Questions

The idea behind the current research is to learn about the status quo and answer the following questions:

1. What do data scientists call themselves? Or by which names they are known?
2. What skill sets or archetypes data scientists have? Or, how can we segment them based on their skills?

Answering the research questions is the first step in learning the job market for data scientists in Iran. It helps unravel the vagueness between industries on the one hand and universities and university graduates on the other.

3.2 Data Collection

To collect the required research data, an online questionnaire was published on Google Forms between May 11th and June 27th, 2015. Data professionals in Iran were requested to participate in the survey regardless of their titles. The researchers also shared the survey link on social networks to ensure diversity in responses thus showing credible archetypal patterns. The number of valid and usable responses was 75. Reliability of the survey instrument was measured by Cronbach's Alpha for 22 skills [7] which was 0.72.

The questionnaire and skills for the current study were adopted from [2, p. 13]. The skills were measured on a 0-100 percent scale and were self-expressed by participants namely Algorithms, Back-End, Bayesian/Monte-Carlo Statistics, Big and Distributed Data, Business, Classical Statistics, Data Manipulation, Front-End Programming, Graphical Models, Machine Learning, Math, Optimization, Product Development, Science, Simulation, Spatial Statistics, Structured Data, Surveys and Marketing, Systems Administration, Temporal Statistics, Unstructured Data and Visualization.

3.3 Data Analysis

Archetypal patterns gathered during the course of study were later identified by clustering. The clustering is a process of partitioning a set of observations into subsets in a way that members of a cluster are similar to those of the same cluster but dissimilar to members of other clusters. Further, the clustering models discover previously unidentified groups in the data. It is also known as unsupervised learning because the labels of clusters are not pre-determined. [6, pp. 444-445]. Here, the input for the data analysis is the participants' 22 skills and the output is patterns observed in the data. Technically, these are called clusters but mentioned as archetypes in this study.

IV. Results

The analysis and visualization processes were done using R statistical package and Excel, respectively. The results are explained in the following subsections.

4.1 Data Scientists Titles

Answering the second research question, the survey revealed that a data scientist in Iran is remembered with different names such as programmer (21.33%), manager (21.33%), data analyst (12.00%), business intelligence professional (6.67%), data miner (4.00%), Information Technology expert (4.00%), business analyst (2.67%), statistician (2.67%) and others (26.67%).

The list has some implications: First, programmers are the main candidates for doing data science jobs in Iran. Second, managers i.e. decision-makers and consumers of data science, are playing a role in building their data tools

including spreadsheets, self-service Business Intelligence systems or small-scale data products.

4.2 Data Scientists Archetypes

Answering the second research question, the clustering of the 22 skills of data scientists was done by the k-means algorithm using fpc package. The fpc package offers the k-means clustering with multiple runs [8]. Moreover, other clustering methods such as k-medians on flexclust package [9] and ewkm on rattle package [10, pp. 190-191] yielded similar results. Fig.1 illustrates a heatmap of the data where columns and rows constitute 22 skills and 75 observations, respectively. Yellow and red colors represent high and low skill levels.

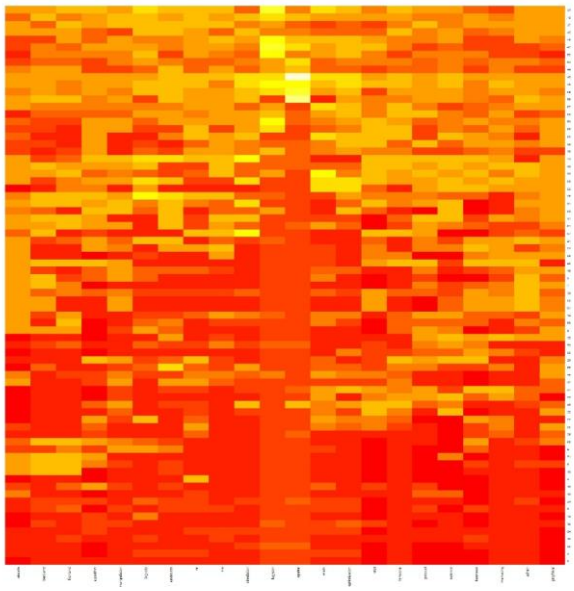

Fig. 1 Heatmap of the Data

An interesting pattern is observed in fig. 1: Those with higher skills (lighter colors) have higher skills in almost all levels and vice versa. The pattern for the middle group is quite different. They have higher skill levels in a subset of the 22 skills and lower level in the remaining skills. Fig. 2 highlights the clustering centroids.

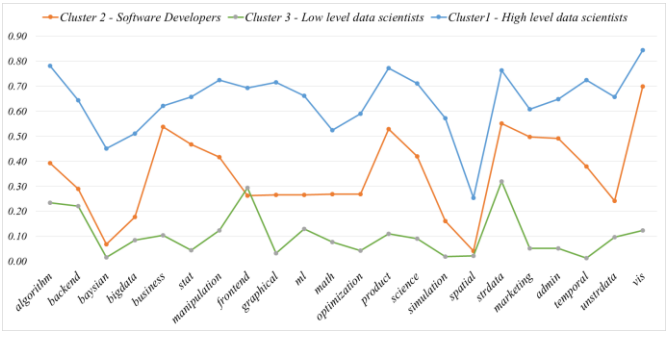

Fig. 2 K-means Clustering Centroids

The researchers found that data scientists in Iran are segmented into the following three clusters:

- Cluster 1: Members of this cluster are data scientists with higher skills. They outperform their peers in clusters 2 and 3 by any of the given 22 skills. This cluster has the highest ratio of PhDs (16%) compared with other clusters.
- Cluster 2: Members of this cluster are software developers with higher skill in back-end and front-end programming. They are experts in structured data with a fair amount of experience in product development, systems administration and visualization. This cluster has the highest ratio of women (36.36%), the highest ratio of technical and project managers (24%) and senior managers (28%). Moreover, it is very similar to programming cluster in [2, p. 11] and hacker archetype in [4].
- Cluster 3: Members of this cluster are data scientists with lower skills. They fall behind their peers in Clusters 1 and 2 by most of the 22 skills. This cluster has the highest ratio of people majoring in computer sciences (76.47%).

The results reveal that data scientists in Iran are clustered based on their skill levels rather than the type of dominant skills they possess. Such pattern (high-level and low-level data scientists) were not found in the literature.

The three clusters in Fig. 2 correspond to the areas in Fig. 1. For a group of the records at the top of the Fig. 1, high values (yellow) are observed for almost all the variables. This corresponds to high-level data scientists. For a second group of the records at the bottom of the Fig. 1, low values (red) are observed for almost all the variables. This corresponds to low-level data scientists. Finally, software developers have a different pattern: They have expertise in some skills (yellow) and lower skill levels in the rest of skills (red).

4.3 Model Evaluation

The clustering evaluation was done using three steps: assessing clustering tendency, determining the number of clusters and measuring clustering quality, which are described below:

- Assessing clustering tendency

It determines whether a set of observations has a non-random structure. Applying Hopkins statistic test iteratively using the 0.5 threshold, the researchers found that the data have statistically significant clusters. [6, pp. 484-486] The results were calculated using the clustertend package in R [11] as follows:

- Observations: 75
- Samples: 15
- Iterations: 1000
- Mean: 0.46
- Standard Deviation: 0.02
- Minimum: 0.39
- Median: 0.46
- Maximum: 0.51
- Numbers above 0.5 threshold: 5

As the Hopkins test statistic goes beyond 0.5 thresholds in only 0.005 of iterations, data has statistically meaningful clusters and is distributed non-uniformly.

- Determining the number of clusters

The correct number of clusters plays an important role in the clustering since it affects the entire process. However, the result is often ambiguous as it depends on the shape and

scale of data and the clustering resolution required by the user. [6, p. 486] In the current study, only the number of three clusters yielded meaningful and interpretable results.

- Measuring clustering quality

The silhouette coefficient evaluates the quality of clustering based on the compactness of each cluster and the separation of each cluster member to other clusters. This coefficient is between -1 and 1. The more the coefficient, the better clustering results i.e. more compact and well-separated clusters. [6, pp. 489-490]. In the current study, the silhouette coefficient was calculated with cluster package in R [12]. The results are as follows:

- High-level data scientists: 15 members, silhouette coefficient = 0.27
- Software developers: 33 members, silhouette coefficient = -0.01
- Low-level data scientists: 27 members, silhouette coefficient = 0.3

The silhouette coefficients for the three clusters match the heatmap illustrated in Fig. 1. High-level and low-level data scientists' clusters are compact and well-separated: They have relatively high (yellow) and low values (red) for all the 22 variables respectively. As such, members of the two clusters are closer to each other in the 22-dimensional space compared with the "software developers" cluster that consists of high values (yellow) in a subset of skills and low values (red) in the remaining skills.

4.4 Spread of Data Scientists Skills

To explore data scientists' skills further, the spread of the given 22 skills was calculated for each individual in terms of standard deviation. A [13, pp. 39-43] of these standard deviations per cluster is depicted in Fig. 3. The width of the boxes corresponds to the number of cluster members.

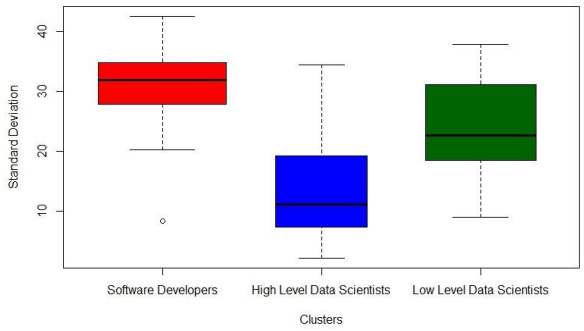

Fig. 3 The spread of data scientists' skills per cluster

The researchers interpreted Fig. 3 in two ways: First, the medians are far apart so these three clusters have different spread in their 22 skills; software developers on top of the three, low-level and high-level data scientists at the next levels. Therefore, software developers have the largest spread in their skills. This means they are focused on a narrower subset of skills. Similarly, high-level data scientists have the least spread and finally low-level data scientists fall in between.

Second, the interquartile range (IQR) for these three clusters was investigated. It is a measure of variability in data defined as the difference between the third quartile and the first quartile Considering the interquartile range, high-level data scientists and low-level data scientist are close to each other and larger than the range for software developers. In other words, the spread (interquartile range) of the spread of (standard deviation) skills for software developer cluster falls behind those of high-level and low-level data scientists. This phenomenon shows that software developers have similar strong and weak points i.e. there are a certain number of skills in which software developer are experts and the rest of skills in which they lack expertise.

V. Interpretation and Discussion

In this section, the results and observed patterns will be interpreted.

5.1 Interpreting Results

The data shows that an important factor differentiating data scientists' skills in clusters 1 (high-level data scientists) and cluster 3 (low-level data scientists) is the type of their work. Cluster 3 members had a larger proportion of consultancy jobs (58.82% in cluster 3 versus 64% in cluster 1) Meanwhile, Cluster 3 members had the highest ratio of people working for a holding company or a department of a large corporation (41.18% in Cluster 3 versus 36% in Cluster 1).

It seems that professionals of a holding company or a large corporation have less diversity in their projects and their work becomes a routine after a while. Therefore, a probable scenario is that the type of work affects the diversity of data science projects and this, in turn, affects the diversity and amount of skills data scientists have. In other words, a consultancy job in a small company needs more quantitative skills than a job in an analytics department of a large corporation.

5.2 Interpreting Patterns

Three patterns were observed in the data:

1. **Similarity in skill patterns between clusters 1 and 3**

As Fig. 2 shows, an interesting pattern between Cluster 1 (high-level data scientists) and Cluster 3 (low-level data scientists) is their similarity: the two clusters experience many ups and downs simultaneously.

A probable explanation for this similarity is the fact that none of the data science university programs has so far been designed in Iran and data scientists gained all of their skills in the course of their works. Consequently, their strengths and weaknesses are quite similar. For example, both clusters have low levels of simulation, optimization, Bayesian and spatial statistics since these skills are merely learned theoretically. Further, both have high skills in working with structured data which is taught in schools and is being widely used in industries.

2. **The high discrepancy in skill levels among three clusters**

As illustrated in Fig. 2, the discrepancy among these three clusters centroids is high. For example, the difference in "graphical models" skill between high-level data scientists and software developers is 45% whereas the difference in the same skill between software developers and low-level data scientists is 24%. This demonstrates that a data scientist should escalate his/her skills considerably to enter a higher

cluster. It is worth mentioning that this discrepancy is very low in generally low level skills such as Bayesian/Monte-Carlo and Spatial Statistics.

A probable reason for this phenomenon may be similar to pattern 1. The lack of data science skills in university programs affects this high discrepancy. Consequently, universities can narrow this gap by prioritizing the data science education.

3. The role of software developers in data science teams

The highest skill levels for members of the "software developers" cluster are business (54%), product development (53%) and visualization (70%). This elucidates that the members of this cluster are involved in creating data products rather than implementing analytical methods. The low level of "machine learning" skill (27%) reinforces this hypothesis, since machine learning is a key in building predictive models. In addition, the low level of unstructured data skill evinces that the members of this cluster are not employed for building unstructured data models.

This phenomenon also implies that this cluster is playing its traditional role in software development so that companies can educate the group of data scientists and assign more diverse roles and let it grow data science skills.

5.3 Implications for Curriculum Design

Considering the three archetypes, an important question comes in: How can universities help students to go to higher level of skills? The journey starts by finding entry level jobs as a low-level data scientist and continuing to become a high-level data scientist.

Basic education in data science skills plays the main role in starting a career in data science. Students of programming and statistics backgrounds are much more likely to apply for data science jobs. However, this is insufficient to meet the demand. Many universities started data science programs to fill the gap, especially at the master's level. For university graduates, other methods emerged to facilitate the process including Massive Open Online Courses (MOOCs) and bootcamps. MOOCs played a key role in data science education. Data science courses gained popularity shortly after the advent of MOOCs and helped to educate thousands of students. MOOCs can be used at personal, group or organizational level to maximize efficiency and provide a social learning environment. In the meanwhile, bootcamps tried to bridge the gap by collaboration with different industries and offering practical training. Examples of data science bootcamps are "Insight Data science Fellows Program" [5], "ASI Data Science & Business Analytics" [14] and "The Data Incubator" [15].

Members of the software developer cluster have a different roadmap: In order to become a high-level data scientist, they should widen their expertise in statistics and business. This can be accomplished by the same tools previously mentioned.

Becoming a high-level data scientist requires years of learning and experience and cannot happen overnight. However, creating a group of data scientists for learning and collaboration makes a big change. Universities in many countries established campus-wide data science research centers where researchers from different disciplines can collaborate in interdisciplinary projects. This approach distinguishes data science from machine learning research centers. You can find business and marketing data scientists, bioinformatics researchers, computational physicists, environmental science specialists and many other groups working on quantitative data-driven projects.

In the context of Iran, university courses and MOOCs are suitable starting points and can even work out in the short run but other methods require thorough needfinding, planning, budget and resource allocation and evaluation.

VI. CONCLUSION

The current study confirms the idea in (Maruyama, 2014) that data scientist archetypes depend on the industry structure and may vary from country to country. As such, the study contributes to the body of knowledge by, first, doing data scientist archetype detection, which was the first of its kind in Iran and second, introducing the new pattern i.e. skill levels.

The current study has the following applications for companies, job seekers and universities. For companies, learning about current data scientists archetypes helps in first, talent acquisition i.e. to define whom they want to hire, with what skill archetypes. As mentioned earlier, failing to do so may result in lost opportunities. Second, in educating and retaining employees, it would be helpful to know which archetype and skills the data scientist has and how educating will change his/her skills set and career path.

For job seekers, learning about data scientists' archetypes helps in finding a map of skills to develop. Job seekers can develop their skills based on a selected archetype or they can devise a new map of interests leading to a job in data science. Universities and educational institutions can study about data scientist archetypes and figure out what kind of data scientists they are cultivating with which archetypes, strengths and weaknesses. This works as a feedback to evaluate and improve educational programs. They can also use the insight in planning data science programs from scratch

REFERENCES

[1] T. Stadelmann, K. Stockinger, M. Braschler, M. Cieliebak, G. Baudinot, O. Dürr and A. Ruckstuhl, "Applied Data Science in Europe: Challenges for Academia in Keeping Up with a Highly Demanded Topic," in European Computer Science Summit, Amsterdam, Netherlands, 2013.

[2] H. D. Harris, S. P. Murphy and M. Vaisman, Analyzing the Analyzers An Introspective Survey of Data Scientists and Their Work, O'Reilly Media, 2013.

[3] M. Liberatore and W. Luo, "The Analytics Movement: Implications for Operations Research," Interfaces, vol. 40, no. 4, p. 313–324, 2010.

[4] S. Kandel, A. Paepcke, J. M. Hellerstein and J. Heer, "Enterprise Data Analysis and Visualization: An Interview Study," IEEE Visual Analytics Science & Technology (VAST), 2012.

[5] H. Maruyama, "Developing Data Analytics Skills in Japan: Status and Challenge," The Institute of Statistical Mathematics, 2013.

[6] J. Han, M. Kamber and J. Pei, Data mining : concepts and techniques, 3rd ed., Morgan Kaufmann Publishers, 2011.

[7] B. J. Fraser, D. F. Treagust and N. C. Dennis, "Development of an instrument for assessing classroom psychosocial environment at universities and colleges," Studies in Higher Education, vol. 11, no. 1, pp. 43-54, 1986.

[8] C. Hennig, "fpc: Flexible Procedures for Clustering," 2015. [Online]. Available: http://CRAN.R-project.org/package=fpc.

[9] F. Leisch, "A Toolbox for K-Centroids Cluster Analysis," Computational Statistics and Data Analysis, vol. 51, no. 2, pp. 526-544, 2006.
[10] G. J. Williams, Data Mining with Rattle and R: The art of excavating data for knowledge discovery, Springer, 2011.
[11] L. YiLan and Z. RuTong, "clustertend: Check the Clustering Tendency," 2015. [Online]. Available: http://CRAN.R-project.org/package=clustertend.
[12] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert and K. Hornik, "cluster: Cluster Analysis Basics and Extensions," 2015. [Online]. Available: https://cran.r-project.org/web/packages/cluster/index.html.
[13] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley Publishing Company, 1977.
[14] ASI team, 2015. [Online]. Available: http://www.theasi.co/.
[15] The Data Incubator team, 2015. [Online]. Available: https://www.thedataincubator.com/.
[16] J. G. Harris, N. Shetterley, A. E. Alter and K. Schnell, "The Team Solution to the Data Scientist Shortage," Accenture Institute for high performance, 2013.