Industrial Requirements for Jobs in Data Science

Gokul Bhojraj
Business Intelligence
Dalarna University
Borlänge, Sweden
e-mail: h18gokka@du.se

Yeswanth Markonda
Business Intelligence
Dalarna University
Borlänge, Sweden
e-mail: h18yesma@du.se

Pär Eriksson
Data and Information Science
Dalarna University
Borlänge, Sweden
e-mail: pei@du.se

***Abstract*- Data science jobs are among the most sought-after positions in the world. Yet, many existing and emerging workers don't have complete idea about the full skillset employers need. While this isn't a new problem, we have tried to find out the most important requirements that Industries are expecting and give an insight for the people looking jobs in Data Science. To make the prediction we have analysed the employers job advertisements and processed it to find out the key requirements that are expected from them using Natural Language Processing. The result shows that the most sought-after person has a PhD or bachelor's degree majoring in engineering with skills in Machine Learning and research with working knowledge using tools like Excel and Python.**

Keywords- Data science; Data Scientist; Data Analyst; Job requirements; Natural Language Processing; Visualization; Prediction; Analysis; Data Cleaning; classification

I. INTRODUCTION

Data Science has been established as an important emergent scientific field and paradigm driving research evolution in such disciplines as statistics, computing science and intelligence science, and practical transformation in such domains as science, engineering, the public sector, business, social science, and lifestyle. The field encompasses the larger areas of artificial intelligence, data analytics, machine learning, pattern recognition, natural language understanding, and big data manipulation. It also tackles related new scientific challenges, ranging from data capture, creation, storage, retrieval, sharing, analysis, optimization, and visualization, to integrative analysis across heterogeneous and interdependent complex resources for better decision-making, collaboration, and, ultimately, value creation.

We have considered the job advertisements in Indeed.com for the year 2018 to analyse the landscape of jobs requiring data science and analytics competencies and skills. We have used Natural Language processing method to process the description about the jobs and found the major requirements from them. This data-driven predictions and strategies has provided major factors that classify the data science jobs in market and the requirements for each of it.

There are two different markets for data science and analytics jobs. Across the ecosystem, we see two broad families: analytics-enabled jobs and data science jobs.

Data analysts sift through data and seek to identify trends. What stories do the numbers tell? What business decisions can be made based on these insights? They may also create visual representations, such as charts and graphs to better showcase what the data reveals.

Data scientists are pros at interpreting data, but also tend to have coding and mathematical modelling expertise. Most data scientists hold an advanced degree, and many went from data analyst to data scientist. They can do the work of a data analyst, but are also hands-on in machine learning, skilled with advanced programming, and can create new processes for data modelling. They can work with algorithms, predictive models, and more. Competencies and skills needed for data science jobs are different. They're often the aptitudes that entrepreneurs and innovators most desire. Candidates for these roles have strong credentials (Skills and education) in programming and applied data science.

II. OBJECTIVES

The important objective were a) To find the Data Science job market share b) To find the different categories in Data Science Job market c) To find important requirements like Skills, Tools, Major, Degree for Data Analyst and Data Scientist Jobs d) To visualize the results in an easy understandable manner. e) Draw different conclusions from the results that we found.

III. METHODOLOGY

We have used the Dataset obtained from jobs advertisements posted in Indeed.com (USA) of year 2018 size 6963. The Dataset included the attributes Position, Company Name, Reviews, Location, and Description about the jobs. We performed the following processing steps:

a) Data cleaning: We have cleaned the data by removing irrelevant attributes, later based on the data we found the two major job positions in the Data Science Market Namely, Data Scientist and Data Analyst. We categorised the jobs into Data Scientist, Data Analyst, Engineer and Others based on the position title given by the employers and classified the companies into small, medium, large according to the reviews of the company. After cleaning the data, we found 6953 jobs that were usable and we created new attributes mainly Job category, size of the company.

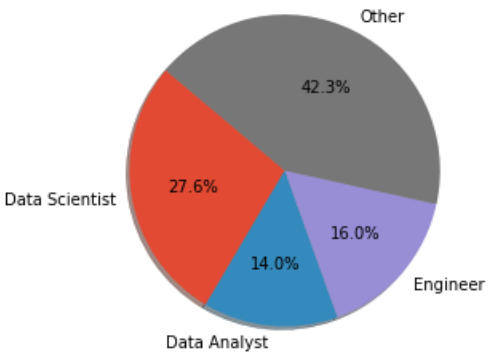
b) Natural Language processing: Significant work has been done through natural language processing technique to process the descriptions provided by the employers for the jobs. We used Python libraries for implementing the Natural language processing techniques.

We followed the standard steps used in Natural Language processing (NLTK) such as sentence segmentation, word tokenization, text lemmatization, dependency phrases, named entity recognition (NER). We considered the frequency of each of the requirement to find the importance of that requirement to that job as per the descriptions. Find the summation of that requirement for each of our category to find the importance of it for the job position.

Later, we extracted the main requirements from the description namely skills, tools, major and degree for each of the categorised jobs. We consider the highly important requirements and visualize the analysed data in a simpler way using different data representations.

IV. RESULTS

After analysing the data of size 6953, we found the following share of jobs for Data Science, which is 41.6% which is a tremendous share in the entire job market.



Job Market

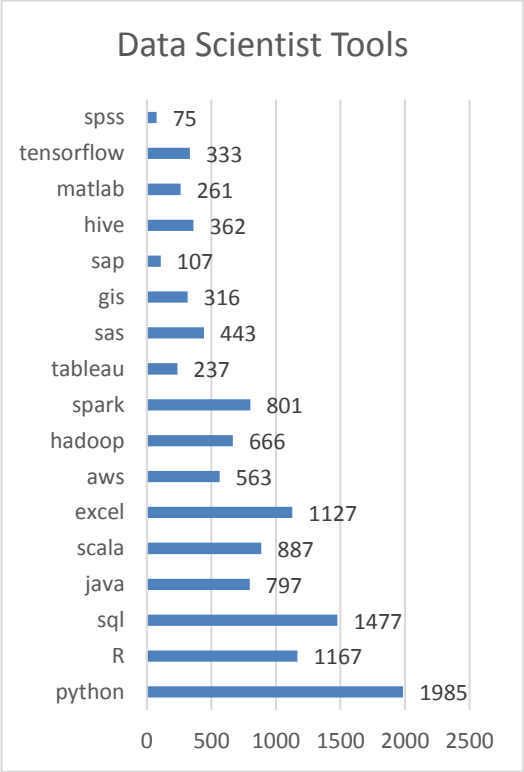
Position	
Title	**Count**
Data Scientist	1919
Data Analyst	975
Engineer	1115
Other	2944
Total	**6953**

We found that among Data scientist and Data Analyst, Data Scientist share the higher percentage of the jobs with the total of 27.6%, Data Analyst with 14.0%, Engineer with 16.0% and other jobs with 42.3%. Hence, we found that Data Science is an important career among rest.

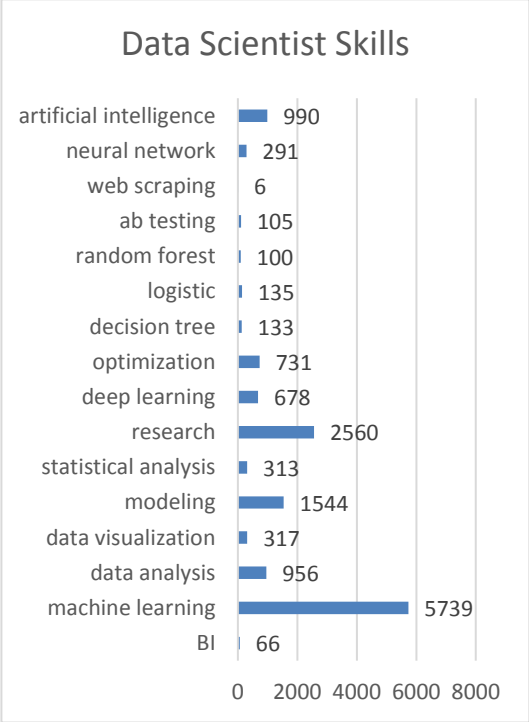
1) Requirements for Data Scientist position:

The following are the requirements expected for a Data Scientist by employers categorized into Tools, Skills, Degree asked, Major preferred.

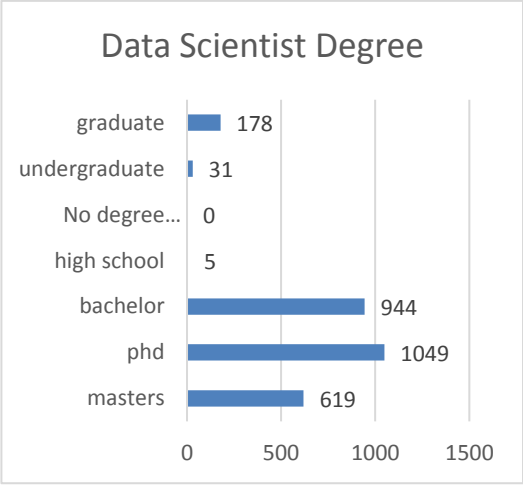
Tools: We found that the main tool that employers ask are Python, Sql, R, excel and followed by the rest as shown in the graph.



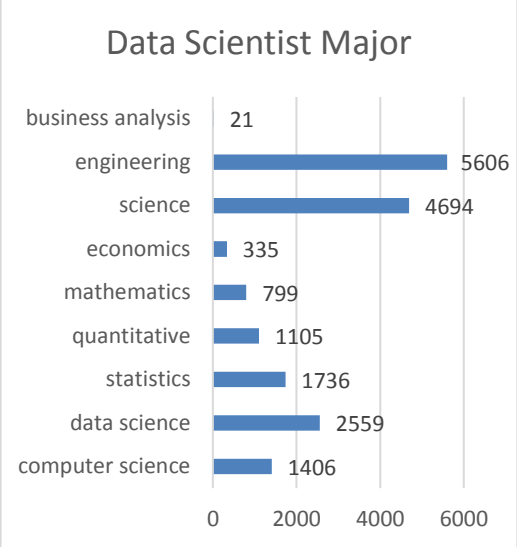
Skills: We found that the main skills that employers ask are Machine Learning, Research, Modelling, Artificial Intelligence and followed by the rest as shown in the graph.



Degree: We found that the Degree that employers ask are PhD, Bachelors, master's and followed by the rest as shown in the graph.



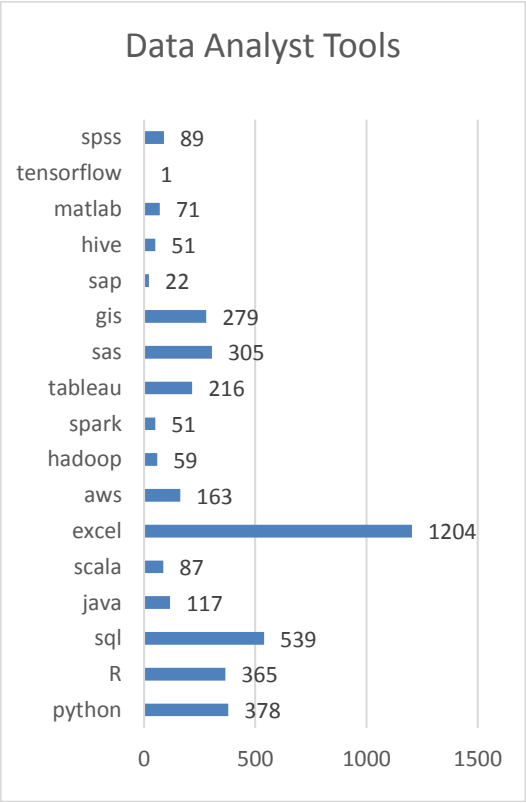
Major: We found that the Major that employers ask are Engineering, Science, Data Science, Statistics and followed by the rest as shown in the graph.



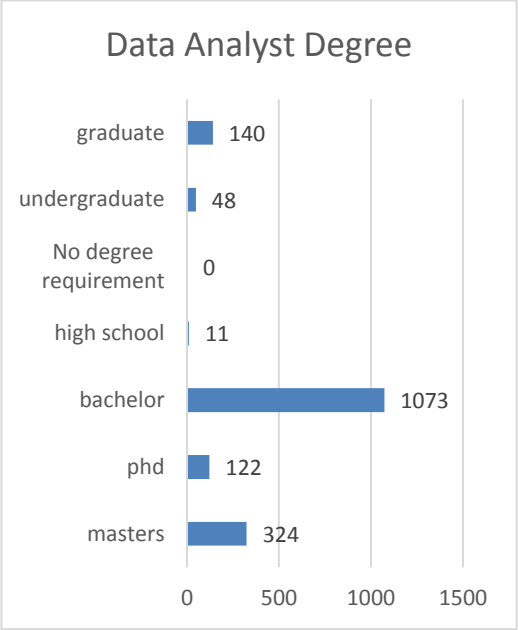
2) Requirements for Data Analyst position:

The following are the requirements expected for a Data Analyst by employers categorized into Tools, Skills, Degree asked, Major preferred.

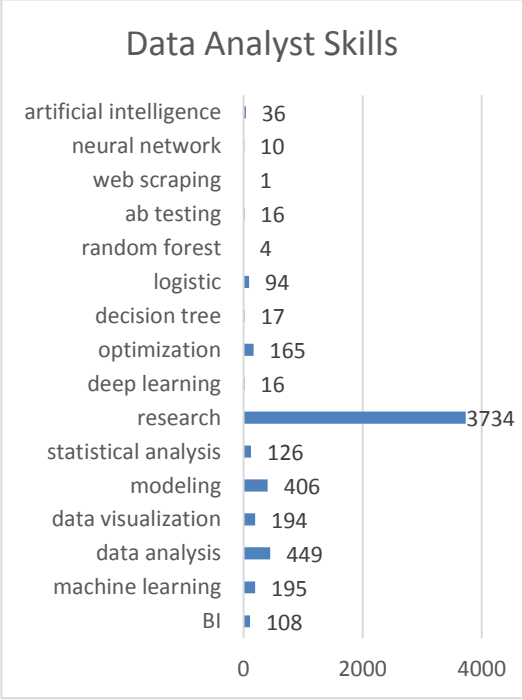
Tools: We found that the main tool that employers ask are Excel, Sql, Python, R and followed by the rest as shown in the graph.



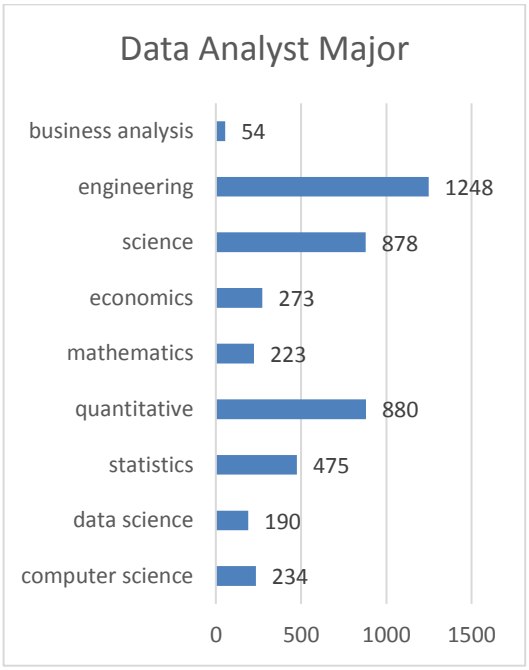
Degree: We found that the Degree that employers ask are Bachelors, master's and followed by the rest as shown in the graph.



Skills: We found that the main skills that employers ask are Research, Data Analysis, Modelling, Machine Learning and followed by the rest as shown in the graph.



Major: We found that the Major that employers ask are Engineering, Quantitative, Science, Statistics and followed by the rest as shown in the graph.



V. APPLICATION

Our research knowledge can be used in following areas:

Academic: It can be used by the universities while designing the courses to make students well prepared to fit in the data science job.

Job Seekers: People who are looking for the jobs in data science can use this knowledge to acquire the relevant skills and tools.

Employers: Employers can use this knowledge to have a proper idea while providing descriptions for a job in data science.

General: To have an idea about trending jobs, data science fields, data science job market and comparion between data scientist and data analyst.

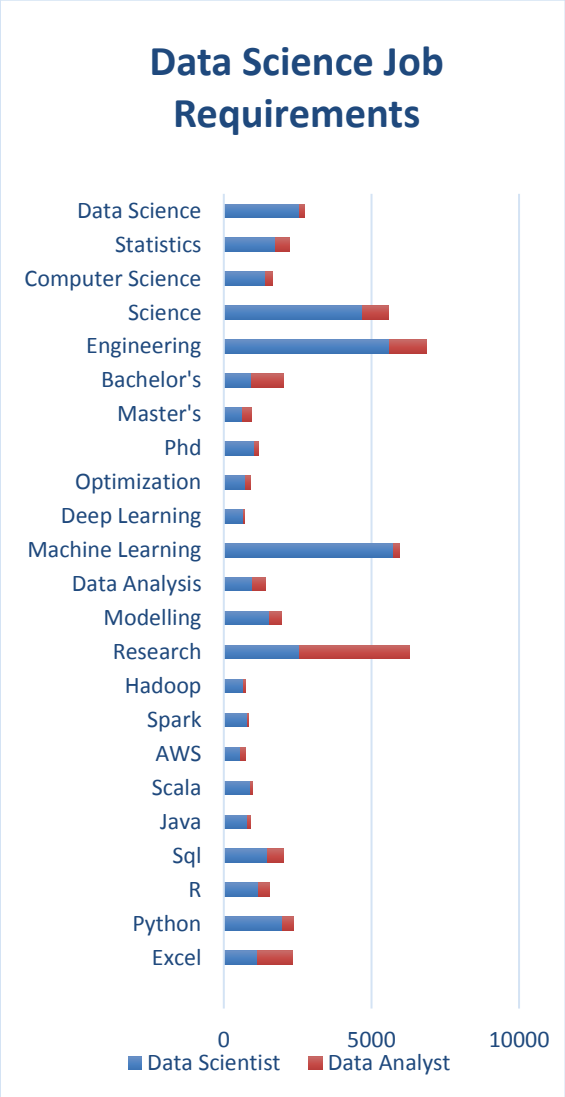
VI. CONCLUSION

We found that the data science is the most valuable and highly required job position for any company to grow in the competitive world, data science jobs share a huge percentage when compare to other jobs. We found that employers not really classify data science jobs into data scientist and data analyst based on any criteria. They just give a title which is more suitable for them and there are no specific criteria to divide among the two jobs.

Data science jobs requires high education qualifications and knowledge in various tools and should possess several skills in different domains.

In data science there is no exact limit or no exact tools, skills, degree and major that are required. Employers just asked based on their requirements and will giving descriptions for the job's employers tried to fit various requirements that are irrelevant for data science employee such as HTML, CSS, PHP, etc.

So, the major requirements that person should have to be hired by an employer in most of the circumstance he should possess the following qualifications:
Degree: PhD, Master's and Bachelor's.
Major: Engineering, Science, Data Science, Statistics.
Skills: Machine Learning, Research, Modelling, Data Analysis, Optimization and Deep Learning.
Tools: Excel, Python, R, Java, SQL, Scala, Hadoop, AWS and Sparks.



VII. FUTURE WORK

In this analysis we have not considered experience required for the particular job position, because experience is just specified by the employer without any exact criteria. But, however, experience will also be an important parameter during the process of hiring in any company. Today, most of the employers are looking for hands on experience with various tools, sometimes employers don't care for the other requirements if the candidate have a quite good experience in the field. So, the future scope of this analysis will be done by considering experience parameter and make the analysis more accurate prediction.

REFERENCE

[1] K. Ryan, "The Role of Natural Language in Requirements Engineering," in [1993] Proceedings of the IEEE International Symposium on Requirements Engineering. IEEE Comput. Soc. Press, 1993, pp. 240–242.

[2] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. O'Reilly Media, Inc., 2009.

[3] C. Rolland and C. Proix, "A Natural Language Approach for Requirements Engineering," in Advanced Information Systems Engineering. Springer, 1992, pp. 257–277.

[4] P. Pradnya, "Overview of Predictive and Descriptive Data Mining Technique", in Computer Science and Software Engineering, 2015.

[5] C. Manning and H. Schuetze, Foundations of Statistical Natural Language Processing. The MIT Press, 2012.

[6] M. Porter, "An Algorithm for Suffix Stripping," Program: electronic library and information systems, vol. 14, no. 3, Dec. 1980, pp. 130–137.

[7] Eagle, N., and A. Pentland. (2006). "Reality Mining: Sensing Complex Social Systems." Personal and Ubiquitous Computing, Vol. 10, No. 4, pp. 255–268.

[8] White, C. (2008, July 30). "Business Intelligence in the Cloud: Sorting Out the Terminology." March 2013.

[9] Chae, B., D. B. Paradice, J. F. Courtney, and C. J. Cagle. (2005). "Incorporating an Ethical Perspective into Problem Formulation." Decision Support Systems, Vol. 40, No. 2, pp. 197–212.

[10] Nizar R. Mabroukeh, Christie I. Ezeife, "Using Domain Ontology for Semantic Web Usage Mining and Next Page Prediction", Proceeding of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, ACM, New York, pp. 1677–1680, 2009.

[11] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.A. Cichocki and R. Unbehaven. Neural Networks for Optimization and Signal Processing, 1st ed. Chichester, U.K.: Wiley, 1993, ch. 2, pp. 45-47.

[12] Low, Yucheng, et al. "Distributed Graph Lab: a framework for machine learning and data mining in the cloud." Proceedings of the VLDB Endowment 5.8 (2012): 716-727.

[13] Yanbin, Ye and Chia-Chu Chiang. "A parallel apriori algorithm for frequent item sets mining." Software Engineering Research, Management and Applications, 2006. Fourth International Conference on. IEEE, 2006.

[14] Neethu B, "Classification of Intrusion Detection Dataset using machine learning Approaches", International Journal of Electronics and Computer Science Engineering, 2012.

[15] Vincent F. Mancuso, Dev Minotra, Nicklaus Giacobe, Michael McNeese and Michael Tyworth" ids NETS: An Experimental Platform to Study Situation Awareness for Intrusion Detection Analysts", IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, New Orleans, LA, 2012.

[16] Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. ONLINE-WESTON THEN WILTON-, 23, 62-73.

[17] Hendrix, G. G., Sacerdoti, E. D., Sagalowicz, D., & Slocum, J. (1978). Developing a natural language interface to complex data. ACM Transactions on Database Systems (TODS), 3(2), 105-147.

[18] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 427-434). IEEE.

[19] McDonald, R., Crammer, K., & Pereira, F. (2005, October). Flexible text segmentation with structured multilabel classification. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 987- 994). Association for Computational Linguistics.

[20] Ahonen, H., Heinonen, O., Klemettinen, M., & Verkamo, A. I. (1998, April). Applying data mining techniques for descriptive phrase extraction in digital document collections. In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on (pp. 2-11). IEEE.

[21] N. Mlambo, "Data Mining: Techniques, Key Challenges and Approaches for Improvement", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016.

[22] Kantardzic, Mehmed, Data Mining: Concepts, Models, Methods, and Algorithms, New York: John Wiley & Sons Inc publishes, 2003.

[23] Kwak, D.; Kim, K. A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. Expert Syst. Appl. 2012, 39, 2590–2596.

[24] BERGER, A. L., DELLA PIETRA, S. A., AND DELLA PIETRA, V. J. 1996. A maximum entropy approach to natural language processing. Computational Linguistics 22, 1, 39–71.

[25] E. Meszarosova, "Python and Teaching Programming at Upper Secondary Schools", International Conference on Information and Communication Technologies in Education, 2015.

[26] Marshall, B., McDonald, D., Chen, H., and Chung, W. 2004. EBizPort: Collecting and analyzing business intelligence information. J. Amer. Soc. Inform. Sci. Technol. 55, 10, 873–891.

[27] Jermol, M., Lavrac, N., & Urbancic, T. (2003). Managing business intelligence in a virtual enterprise: A case study and knowledge management lessons learned. Journal of Intelligent & Fuzzy Systems, Vol. 14(3), pp. 121-136.