Issues in Information Systems
Volume 17, Issue IV, pp. 137-144, 2016

IDENTIFYING DATA SCIENCE AND ANALYTICS COMPETENCIES BASED ON INDUSTRY DEMAND

Ashraf Shirani, San Jose State University, ashraf.shirani@sjsu.edu

ABSTRACT

Data science and analytics are among the most in demand and fast growing disciplines. However, due to the fact that the field straddles boundaries of many disciplines and its skillset continues to evolve, it is often difficult to delineate its specific skillset and competencies. Separately, the industry propelled by rapid and accelerating innovations in technology and business models is creating demand for data science skills that many academic institutions are not adequately prepared to meet. As a result, there are indications of an impending gap between the workplace demand and supply of certain skills. Competency based education and various variations of it are being increasingly explored as one viable approach to addressing the issue. Based on the author's survey of the industry demand for data science and analytics, this paper attempts to identify and categorize competencies in this field into a simple taxonomy that should be of help in data science program and curriculum design.

Keywords: Data Science, Analytics, Big Data Analytics, Competency Based Education; Problem Based Learning

INTRODUCTION

Demand for data science and analytics skills has grown more than three times in the past five years and continues to increase at a fast pace, according to the labor market analytics firm, Burning Glass technologies (2016). Describing what exactly are data science skills, however, is somewhat nuanced. This is partly due to the fact that unlike many other disciplines, boundaries of data science are not clearly demarcated and overlap multiple fields ranging from mathematics, statistics, and computer science to data and information management and visualization, to Apache Hadoop and programming languages. This in turn poses a challenge to various stakeholders in the field including program administrators, course designers, and learners. One objective of this research is to help provide clarity in this regard by developing taxonomy derived from the current marketplace demand for data science skills.

There are indications that the key stakeholders in the US higher education system – students and employers –are becoming increasingly dissatisfied with the system. This is partly attributable to the declining value of the college degree due to increasing costs and declining benefits from higher education (Vedder and Denhart, 2014). A 2013 report from the Center for College Affordability and Productivity (Vedder et al., 2013) found significant underemployment among US college graduates. The study highlights that there are now "more college graduates working in retail than soldiers in the U.S. Army, and more janitors with bachelor's degrees than chemists. In 1970, less than 1% of taxi drivers had college degrees. Four decades later, more than 15% do."

Studies that have focused particularly on the skills gaps, report increasing skill gap in the information technology sector. Skill gap refers to the extent of mismatch between the skills required for a job and those possessed by the graduates. Based on the Bureau of Labor Statistics 2014 data, the skills gap in science and engineering fields was among the highest in the US (Bessen, 2014). Also, a joint World Economic Forum (WEF) and Boston Consulting Group study (2016) found significant skill gaps worldwide and concluded that "too many students are not getting the education they need to prosper in the 21st century and countries are not finding enough of the skilled workers they need to compete." WEF proposed a shift towards marketable education that prepares graduates for work on day one on the job. The gap is especially acute in the information technology sector and a number of scientific and engineering fields, according to the study.

The widening gap between the supply and demand of certain skills is a manifestation of the gradual reversal of the roles of higher education and industry: the typical cycle of innovations originating from university research labs and

then moving onto the marketplace is slowly reversing direction as far as the information, communication, and computing technologies are concerned. Relatively more agile corporations enabled by new business models, sharing economy, and crowd-sourcing are setting the pace of technological change. Universities with their decades old business models and administrative structures are slow to change and are playing catch up with the industry. Take big data analytics, for example: various technologies under the open source Apache Hadoop umbrella are the standard that IT industry leaders have been busy implementing for a number of years whereas only a handful of universities are beginning to explore and offer education in these fields.

Competency-Based Education

Partly in response to the aforementioned trends, an instructional framework known as competency based education (CBE) with origins in the medical filed, is experiencing resurgence in higher education in general. CBE is an approach to teaching and learning aimed at proving knowledge and skills that prepare students for work in their field of study through projects and tasks that mirror authentic organizational work (Hoogveld et al., 2015). In one sense, this is somewhat like the "direct method" of learning a foreign language in which learners are immersed into their target language by not using any language other than that language in the learning process.

A prominent feature of CBE is problem-based learning (PBL). In PBL, the emphasis is placed on accomplishing a typical task in the industry whereby the learners learn the concepts and skills necessary to complete the task in the context of specific organizational environment rather than getting introduced to the concepts in abstract before actually applying them to a specific task. A number of studies have found that PBL improves quality of education in general, and is particularly more effective in providing marketable education than the traditional approaches (Koenen et al., 2015). PBL's focus typically is on complex, realistic tasks in the context of professional work environment (Van-Merrie-nboer & Kirschner, 2001). Creating such tasks, on the other hand, requires identifying relevant competencies (Hoogveld et al., 2005). The term "competency" has been defined variously, though a common theme has emerged that includes knowledge, skills, attitudes, and behaviors necessary to meet complex demands of a task in the particular context (OECD, 2005). For the purpose of this study we define competencies as personal attributes including attitudes and behaviors, plus professional knowledge and skills necessary to successfully perform a give task or job. The objective of this study is to identify current set of competencies in data science and analytics derived from their demand in the workplace. Doing so would be of help in designing undergraduate and graduate curricula and courses in data science.

DATA SCIENCE AND RELATED DISCIPLINES

Beginning perhaps with the Harvard Business Review article describing data scientist as the "sexiest job of the 21st century" (Davenport & Patil, 2012), it appears that data science has become the standard-bearer of all things analytics. It has created a great enthusiasm among students, academic institutions, and employers alike. At the same time, though, other related and competing terms have created some ambiguity and they need to be clarified.

A visual conceptualization of the constituent disciplines in the overall data science and analytics field is shown in Figure 1 and briefly described below. The two different colors and shades of the components in the diagram indicate core components (blue color), and supporting disciplines.

Data mining: Much before the term data science became popular, data mining represented the leading technology for discovering trends, patterns, and associations in historic data. It also offered tools and techniques for projecting the trends into the future and a precursor of predictive analytics (Fayyad et al., 2002).

Business intelligence (BI) is a set of analytical techniques, software tools, and processes for gathering, organizing, storing, reporting, and analyzing data for organizational decision support (Wixom et al., 2010). BI is a niche analytical framework with roots in relational databases and data warehousing. Data sources for BI projects typically include structured, organizational data along with relevant extra-organizational information. Data is usually modeled in dimensional format to facilitate aggregation and summarization and is optimized for query processing, reporting,

and visualization. BI software tools facilitate visibility into the organization's operations and provide organizational decision support. BI initiatives also enable business performance management.

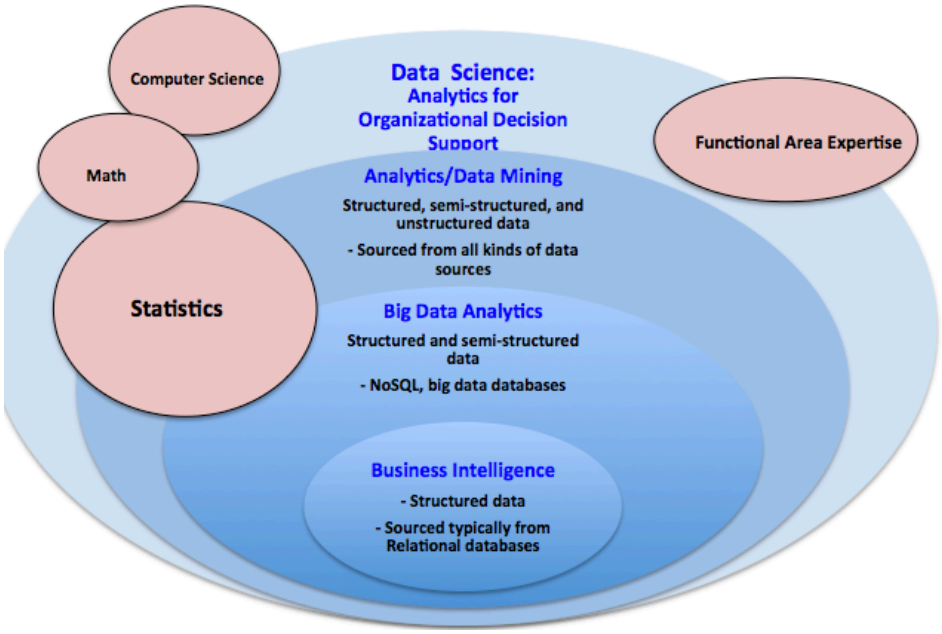

Figure 1. Data Science as an Umbrella Term

Big data analytics: Big data and NoSQL databases have gained immense significance in data management and analytics in recent years. The term "big data" is commonly defined by its attributes starting with the letter "V": volume, variety, velocity, and others. Big data systems and applications aim to efficiently store and analyze data in a timely manner – such as 4-5 terabytes of data generated by the New York stock Exchange each day, or the Hadron Collider near Geneva, Switzerland, that produces about 30 petabytes or more of data per year (White, 2015). Variety refers to the format and structure of data – the data being generated in large quantities may be structured, semi-structured or unstructured and may come in varying file formats. As for velocity, it is becoming increasingly necessary to capture, store, and analyze real-time streaming data originating from a variety of sources ranging from satellites to sensors and mobile devices. Distributed computing, based on the Hadoop framework, map-reduce algorithm and its variants are the primary means for enabling big data processing and analytics. Some of the underlying data management and analysis concepts and techniques originate from the relational database model though many other are based on distinctly different algorithms, programming languages, and technologies.

Many organizations foresee big value from big data analytics. To accomplish that objective, however, big data first must be transformed into an integrated and consistent format (LaValle et al., 2011). What this requires, essentially, is to find ways to effectively make data warehousing, BI, and big data technologies work together.

Data science: Provost et al. (2013) describe data science as the "principles, processes, and techniques for understanding phenomenon" and to extract useful knowledge from data to support organizational decisions. Furthermore, data science is not about processing and analysis of data - though these steps are necessary for enabling decision science's goal, which is to provide data driven decision support. The pivotal difference between data science and other data driven analytical methodologies including data mining and data analytics is thus whether organizational decision support is part of the overall analytical effort or not. Moreover, as this market-demand based field study indicates, "data science" and "data scientist" are now often used as umbrella terms for all things analytics –and for the purpose of this study we use these terms synonymously.

METHODOLOGY AND RESULTS

A high-level categorization of skills may be represented as a "T-shaped" configuration in which the top part of the letter "T" represents breadth of skills and the vertical part indicates depth (Harris et al., 2013). One typical approach to identifying skills is to focus on the breadth of knowledge and skills necessary to extract information and insights from data and build data products; deeper skills in one of the constituent areas of data science, functional area of application, or both may enhance data science competencies including the ability to extend current knowledge and create new knowledge. We extend this concept to a π-shaped configuration (Figure 2). Although this study touches upon some of the deep level aspects, the focus is on the broad-based skills since doing so is helpful for career planning and also for designing undergraduate and graduate curricula and courses in data science.

Figure 2. Breadth and Depth of Skills

Survey of Open Positions

Three job sites were searched for decision science open positions during the months February through first week of May 2016. Two of these sites (kdniggets.com and r-bloggers.com) post openings specific to data science and analytics, whereas the third site, dice.com, specializes in IT positions. The former two are also the preferred sites of many data scientists for the latest news, views, software polls, and tutorials in the field. Table 1 below lists the sites and search terms used in the given order (more to less frequent).

Table 1. Job Sites and Search Terms

Job Sites	Search Terms
- kdnuggets.com - r-bloggers.com - dice.com	- Data science; data scientist - Analytics - Analyst; data analyst - Data engineer; data manager

The text analysis and competency identification process followed the following steps:

- Detailed position descriptions and requirements for each of the 30 positions were copied and pasted into a single text document. The document was then imported into a data and text mining software, *RapidMiner*, for analysis (term identification, frequencies, and ranking).
- Stop-words (words such as such as *the*, *is*, *at*, *which*, and *on* that do not particularly contribute much towards term identification) were removed.
- The case of the text was transformed to all lower case to avoid any case sensitivity.

- One-word terms were first identified, though some of these were less meaningful as single words. We then retrieved two- and three-word n-grams to obtain more meaningful terminology. Table 2 below lists terms and their ranks for the first 50 one- and two-word terms in the ads. These terms convey a general sense of skills needed and their priorities in the analytics workforce. The following tentative conclusions can be drawn from this analysis:

- Expected and/or required foundational skills and backgrounds include Statistics, Mathematics, and computer programming. Also prior work experience is often expected.
- Specific data analytics skills mentioned include data mining; R, SQL, Python, C, and Java programing; and quantitative skills in general.
- Big data skills include Hadoop and Spark frameworks.
- "Soft" skills were among the most frequently expected and included both personal, inter-personal, and teamwork abilities such as drive, insightfulness, problem solving, communication, people skills, and working in teams.

Table 2. Top 50 Relevant Terms by Rank Order

	Term		**Term**		**Term**		**Term**		**Term**
1	Data	11	Analytical	21	Statistics	31	Computer science	41	Distributed
2	Experience	12	Data mining	22	Teams	32	Hadoop	42	Software development
3	Business	13	Problem solving	23	Modeling	33	Statistical analysis	43	Group
4	Team	14	Understanding	24	Performance	34	Mathematics	44	Ideas
5	Skills	15	SQL	25	Agile	35	Solving	45	Interpersonal
6	R	16	Python	26	Big data	36	People	46	Java
7	Mining	17	Algorithms	27	Communication skills	37	Spark	47	Leadership
8	Software	18	Communication	28	Drive	38	Stakeholders	48	Leading
9	Statistical	19	Programming	29	Experience working	39	Business intelligence	49	Metrics
10	Insights	20	Quantitative	30	Complex	40	C	50	Model

Additional views of the text processing output were generated to better understand the underlying content. Sorting the output alphabetically by n-grams, for example, was of much use. As an example, the n-grams beginning with the word "*ability*" indicate emphasis on a number of "soft" skills at least as much as on many of the technical skills. To translate this analysis to a workable set of competencies, however, required inspection of each document and hand coding by the author.

Table 3. Occurrences of the Term "Ability" (Stop-words Reinserted):
An Example of How Sorting and Interpreting the Terms Provided Useful Insights

Term
Ability
Ability to communicate
Ability to communicate complex…
Ability to convey
Ability to convey complex…
Ability to convey message
Ability to effectively…
Ability to effectively analyze
Ability to evaluate
Ability to evaluate business

Using their pedagogical knowledge of the field combined with insights from term-frequency analysis and detailed inspection of each of the position announcements, the authors organized the identified competencies and skills into a taxonomy as shown in Table 4 below.

Table 4. A Taxonomy of Data Science and Analytics Competencies

<table>
<tr><td rowspan="2">(Traditional) Data Science and Analytics</td><td colspan="2">"Big" and Relational Data Analytics</td></tr>
<tr><td>Big Data Analytics</td><td>Relational Data Management & Analytics</td></tr>
<tr><td colspan="3">Advanced Skills</td></tr>
<tr><td>- Data products development (with R Shiny package)
- Deep Learning: recurrent neural networks; reinforcement learning; natural language processing.
- Ensemble learning; random forests; uni-directed graphical models</td><td>Enabling Frameworks and Languages
- Apache Hadoop, Spark, Hive, Yarn, TensorFlow
- Java, Scala, or Python
- Spark SQL, Hive SQL
⬇⬇⬇
- Streaming data analytics
- Temporal and geospatial data analytics
- Network analytics
- Text and latent semantic analysis</td><td>- Data warehousing: dimensional modeling; extraction, transformation, and loading (ETL); dimensional data visualization – dashboards and scorecards
- Advanced SQL</td></tr>
<tr><td colspan="3">Introductory to Intermediate Skills</td></tr>
<tr><td><u>Required</u>: Math and statistics background</td><td colspan="2"><u>Required</u>: Database and programming background</td></tr>
<tr><td><u>Classification and Regression</u>: Neural networks; classification and regression trees; support vector machines; genetic algorithms; linear regression; nonlinear regression; logistic regression; Bayesian classification
<u>Association</u> – Link and sequence analysis: (Apriory algorithm; graph-based techniques)
<u>Cluster Analysis</u>: (k-means, hierarchical)
<u>Text Mining, Web Analytics</u></td><td>- Hadoop, MapReduce, HDFS concepts
- NoSQL Databases: HBase, Cassandra, MongoDB,
- Hive & Hive SQL
- Spark programming
- Scala, Java, and/or Python programming</td><td>- Procedural extensions to SQL (e.g., PL/SQL)
- Intermediate SQL
- The relational data model; normalization; ER diagrams</td></tr>
<tr><td colspan="3">Foundational Competencies</td></tr>
<tr><td>"Hard" Skills
- Math and statistics background (linear algebra, calculus, probability, statistics)
- R programming</td><td colspan="2">"Hard" Skills
- Intro to programming: Java, Python, Scala, or C++/C#
- Intro to Structured Query language (SQL)</td></tr>
<tr><td colspan="3"><u>"Soft" Skills</u>
- Communication
- Teamwork
- Problem-solving
- Critical thinking</td></tr>
</table>

PEDAGOGICAL IMPLICATIONS

The primary goal of this study was to seek input from the workforce marketplace in order to provide guidance for data science and analytics program and curriculum design. Findings of the study culminate in the taxonomy shown in Table 4. This taxonomy suggests two different tracks leading to specializations in the field – the one on the left comprising of the left-most column represents the traditional analytics track, whereas the one in the right-most column is the traditional data management and analysis track, often pursued by information systems majors. The column in the middle on big data analytics is most possibly of interest and use in both tracks.

The big data management and analytics field is growing rapidly in terms of frameworks, in-memory and distributed processing capabilities, data extraction and preparation methods and technologies, and analytical techniques. According to a recent Forrester Research (2015) study, over 60% of the US and European companies had either implemented or were planning to implement big data by the end of 2016. It is thus essential that the traditional data analytics disciplines be augmented with relevant big data skills. Likewise, information systems graduates seeking data analyst and similar positions would be well advised to develop these skills to be able effectively support data science teams.

And finally, equally or perhaps more importantly, there are strong indications from the marketplace that data scientists need to have personal and inter-personal attributes and abilities to formulate, model, and solve organizational problems, communicate results, and rationalize recommended solutions. These skillsets are acquired through good domain knowledge, solid mathematics and statistics background, and by working in teams.

CONCLUSION

The objective of this study was to identify data science competencies based on the overall industry demand to help provide guidance in program and curricula development. The study focused primarily on broad skillset for all levels of data scientists. Though the sample included about 5% of the internship position announcements and another 5% of the senior data science and management positions, the majority (90%) positions in the sample were broad-based, mid-level general data science skills. One somewhat unexpected finding was an almost equal emphasis by the employers on soft skills as on technical skills. These skills include communication, working in teams, leadership, and application of analytics to achieve business objectives.

The overall data analytics and management field is in flux and evolving rapidly. The findings and recommendations of the present study would therefore need to be updated and augmented soon and periodically to remain current and relevant. Also, future research may attempt to provide more granularity in terms of the depth of expertise required for positions at various levels. The current study aggregated all ads into a single document before conducting text analysis rather than segregating positions by their level, title, or track into separate documents and creating term-document matrices for each. Doing so would help correlate a number of dimensions with data science competencies including type and size of the organization, functional area of business, and position level.

REFERENCES

Bessen, J. (2014). Employers Aren't just whining – the "Skills gap" is real. *Harvard Business Review,* (August 25.)

Burning Glass Technologies. *Blurring lines: How business and technology skills are merging to create high opportunity hybrid jobs.* Retrieved May, 2016, from http://burning-glass.com/wp-content/uploads/Blurring_Lines_Hybrid_Jobs_Report.pdf

Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Davenport, T., & Patil, D. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review,*

Fayyad, U., Wierse, A., & Grinstein, G. G. (2002). *Information Visualization in Data Mining and Knowledge Discovery*. San Francisco, CA: Morgan Kaufmann.

Forrester Research (2015). Global Business Technographics Data And Analytics Survey, 2015. https://www.forrester.com/Global+Business+Technographics+Data+And+Analytics+Survey+2015/-/E-SUS2955

Harris, H., Murphy, S., & Vaisman, M. (2013). Analyzing the analyzers: An introspective survey of data scientists and their work. O'Reilly Media, Inc.

Hoogveld, A., Paas, F., & Jochems, W. (2005). Training higher education teachers for instructional design of competency-based education: Product-oriented versus process-oriented worked examples. *Teaching and Teacher Education, 21*, 287-297.

Koenen, A., Dochy, F., & Berghmans, I. (2015). A phenomenographic analysis of the implementation of competence based education in higher education. *Teaching and Teacher Education, 50*, 1-1-12.

LaValle, S., Lesser, E., Shockley, R. P., Hopkins, M., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21-32.

Organization for Economic Cooperation and Development (OECD). The definition and selection of key competencies (2005).

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data*. Sebastopol, CA: O'Reilly Media, Inc.

Russom, P. (2011). *Big data analytics*. Best Practices Report, Fourth Quarter. TDWI Research. http://www.pentaho.com/sites/default/files/uploads/resources/tdwi_best_practices_report-_managing_big_data.pdf

Van-Merrie-nboer, J., & Kirschner, P. A. (2001). Three worlds of instructional design: State of the art and future directions. *Instructional Science, 29*, 429–441.

Vedder, R., & Denhart, C. (2014, 1/8/2014). How the college bubble will pop. *The Wall Street Journal,* pp. A13.

Vedder, R., Denhart, C., & Robe, J. (2013). Why are recent college graduates underemployed? University enrollments and labor-market realities. *Center for College Affordability and Productivity,*

White, T. (2015). *Hadoop: The definitive guide.* O'Reilly Media, Inc.

Wixom, B., & Watson, H. (2010). The BI-based organization. *International Journal of Business Intelligence Research, 1*(1), 13-28.

World Economic Forum, & Boston Consulting Group. *New vision for education: Unlocking the potential of technology world economic forum; Boston consulting group.* Retrieved May, 2016, from http://widgets.weforum.org/nve-2015/