IJDC | *General Article*

Education for Real-World Data Science Roles (Part 2): A Translational Approach to Curriculum Development

Liz Lyon
School of Information Sciences
University of Pittsburgh

Eleanor Mattern
University Library System
University of Pittsburgh

Abstract

This study reports on the findings from Part 2 of a small-scale analysis of requirements for real-world data science positions and examines three further data science roles: data analyst, data engineer and data journalist. The study examines recent job descriptions and maps their requirements to the current curriculum within the graduate MLIS and Information Science and Technology Masters Programs in the School of Information Sciences (iSchool) at the University of Pittsburgh. From this mapping exercise, model 'course pathways' and module 'stepping stones' have been identified, as well as course topic gaps and opportunities for collaboration with other Schools. Competency in four specific tools or technologies was required by all three roles (Microsoft Excel, R, Python and SQL), as well as collaborative skills (with both teams of colleagues and with clients). The ability to connect the educational curriculum with real-world positions is viewed as further validation of the translational approach being developed as a foundational principle of the current MLIS curriculum review process.

Received 20 October 2015 ~ *Accepted* 24 February 2016

Correspondence should be addressed to Liz Lyon, School of Information Sciences, University of Pittsburgh. Email: elyon@pitt.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v11i2.417

Introduction and Context

This paper reports on the findings of a study that is framed as an analysis of requirements for real-world data science positions. The study is the outcome of an exploration of current and future curriculum developments within the graduate MLIS program in the School of Information Sciences (iSchool) at the University of Pittsburgh. The study examines a suite of data science roles based on recent job descriptions and maps their requirements to the current curriculum. The study was conducted in two parts, each using the same methodology. Part 1 investigated three specific data science roles: data librarian, data archivist and data steward, and the findings have been reported elsewhere (Lyon, Mattern, Acker and Langmead, 2015). The current study forms Part 2 of the analysis and examines three further data science roles: data analyst, data engineer and data journalist. We address the three research questions also explored in Part 1:

1. What are the skills, competencies, knowledge, experience and education required for the distinct data science roles?
2. How do these data science role requirements map to current curriculum topics and course offerings?
3. What opportunities emerge for new collaborations and partnerships in developing the data science curriculum?

Literature Review

The challenges in developing workforce capacity and capability for data science and data stewardship have been well-documented (Bakhshi, Mateos-Garcia and Whitby 2014; BRDI, 2015), with an acknowledged data talent gap identified. In particular, there are new curriculum components associated with the range of emerging data science roles. A distinctive approach that draws on translational principles (Woolf, 2008) has been applied to data science education – 'from iSchool to marketplace' (Lyon and Brenner, 2015) – and is adopted in both parts of this study. This recognises the need for higher education providers in the data science arena to take a pragmatic and market-aware view to ensure continuing relevance and compatibility with current workforce demands. Prior commentary on the three data science roles explored in this study highlights the different perspectives on their associated tasks and skills; this commentary includes perspectives on building data science teams (Patil, 2011), a brief review of three data science careers (Lee, 2014), an articulation of data scientist vs data analyst (Rivera and Haverson, 2014) and data scientists vs data engineers (Walker, 2013) and a handbook about data journalism (Gray, Bounegru and Chambers, 2011).

Consideration of data science roles from an educational perspective was addressed by Stanton, Palmer, Blake and Allard (2012) reporting on a workshop; they discussed the concept of a 'T-shaped professional' where broad data knowledge is complemented by deep knowledge in one of three areas (Data Curation, Analytics/Visualisation/Preservation, Networks/Infrastructure). An 'I-shaped' model was also proposed, which included domain knowledge at the base. The paper explores

educational models and recommends a continuing education model beginning with an undergraduate degree (e.g. Computer Science, Information Science, Applied Statistics or Mathematics) or a graduate degree (e.g. MLIS). The student then moves on to acquire domain knowledge through an internship or on-the-job experience.

Methodology

The methodology applied was based on the qualitative workflow described in detail in Part 1 of this job analysis study (Lyon, Mattern, Acker and Langmead, 2015), and comprised the use of keyword searching of job banks to locate and select ten positions within the specified timeframe (i.e. the last 12 months) in each of the three data roles. The job bank used in Part 2 was indeed.com and the postings are listed in the Appendix. This step was followed by a content analysis of the job descriptions using a coding scheme for five categories: a) Education – academic qualifications; b) Experience – direct hands-on practice; c) Knowledge – understanding of/familiarity with topics/subjects/issues; d) Skills – ability to do an action well; e) Competencies – proficiency with specific tools/technologies/programming languages.

The requirements were identified, sorted and examined for patterns across the three roles. We designated requirements that appeared in at least three of the positions as 'Key Requirements.' The next step was to consider the graduate courses provided within the Masters in Library and Information Science (MLIS) Program and also by the Information Science and Technology Program in the School of Information Sciences, University of Pittsburgh, during academic year 2015-2016 to determine which options would support the requirements indicated in the job descriptions. From this mapping exercise, we were able to identify model 'course pathways' and module 'stepping stones'; it also informed our approach to meeting employer expectations in preparing iSchool students for real-world positions. The requirements mappings enabled the identification of course topic gaps and highlighted opportunities for collaboration with other Schools.

Results

Firstly, we record the prolific number of positions available in these three job categories at the point of sampling in October 2015. This is in stark contrast with at least one of the roles, the data archivist, which was analysed earlier in 2015 in Part 1 of this study. The majority of the positions in the sample were located within the private sector and came from a mix of large corporate businesses and smaller companies. There were relatively few positions within universities or other public sector bodies.

Detailed mappings of the requirements for the three roles in each of the five categories listed above are presented in Tables 1, 3 and 5. Note that competency in four specific tools or technologies was required by all three roles: a) Microsoft Excel, b) R, c) Python and d) SQL. Collaborative skills were also highlighted as a requirement in each of the three roles. Position requirements for the three roles referenced the ability to work well with both teams of colleagues and with clients; the ability to work with the latter reflecting the business/corporate nature of the employers. In addition, three of the categories (Experience, Knowledge and Competencies) were found to have overlapping content within the job descriptions, i.e. the categories were blurred with no clear

 doi:10.2218/ijdc.v11i2.417

delineation between them. The results are therefore presented based on the best semantic matching (e.g. '*understanding of...*' was interpreted as '*Knowledge of....*'). A sixth category ('*Other*') was introduced to include additional requirements that did not fall under any of the five themes listed previously. An example is '*Security Requirements*' and these are referenced under the appropriate role.

Data Analyst

The Data Analyst jobs seek candidates with a Bachelor's degree, but with no consistently specified subject domain. There was little emphasis on education within the job requirements (Table 1). In contrast, experience working as an analyst or in data analysis was repeatedly highlighted as a Key Requirement.

A broad range of additional experience is frequent in the narrative of the job descriptions, including experience with data management, data acquisition or sourcing, and statistical work. There is also relatively limited emphasis on knowledge requirements for Analyst positions, though relevant domain knowledge was cited in some job descriptions. In contrast, a relatively broad range of skills were listed, with particular emphasis on writing, attention to detail and accuracy; time management and collaborative skills were also required. Whilst a range of competencies were specified, the primary requirement was for expertise with data analysis software tools such as R, SAS, Alteryx and Stata.

Table 1. Key requirements for data analyst.

Education	Experience	Knowledge	Skills	Competencies
Unspecified Bachelor's degree	Experience in data analysis / as a data analyst	Knowledge of domain area	Written communication and documentation skills	Competence with SQL
	Experience in data management	Knowledge of data mining	Attention to detail and accuracy skills	Competence with data management products MS Excel, R, SAS, Stata, SPSS, Alteryx
	Experience of statistical work	Knowledge of specified statistical techniques	Ability to work independently	Competence with MS Access
	Experience of data acquisition		Organisational and analytical skills	Competence with Python

Table 1. Key requirements for data analyst *(continued)*

Education	Experience	Knowledge	Skills	Competencies
	Experience in data design / data modelling		Research skills	
	Experience with large datasets / data aggregation		Ability to work collaboratively in teams or with clients	
			Time management skills / ability to meet deadlines	

In the additional Other Requirements category, we observed '*Background verification check*' as a security requirement for some positions, but this was not designated a Key Requirement (i.e. it occurred in less than three Data Analyst positions).

Our recommended course pathways through the MLIS and Information Science and Technology Masters Programs for a prospective Data Analyst include the essential and desirable course stepping stones listed in Table 2. Additional courses from other University of Pittsburgh schools and departments are proposed.

Table 2. Course pathways for a data analyst.

Essential Masters courses from the iSchool	Desirable Masters courses from the iSchool	Additional courses from other Schools and Departments
Data mining	Information security and privacy	Business
Data analytics	Technologies and services for digital data	Mathematics
Spatial data analytics	Software engineering	Statistics
Information visualisation	Mathematical foundations for information science	
Cloud computing	Introduction to neural networks	
Database management	Research data management	
Algorithm design	Research data infrastructures	
Web technologies and standards	E-Business	
Social computing		

 doi:10.2218/ijdc.v11i2.417

Data Engineer

The Data Engineer positions seek candidates with a Bachelor's degree in Computer Science, Mathematics, Statistics or Information Systems as a preferred domain (Table 3). A degree in Business or Information Technology was specified in some positions. In other positions a Masters degree was preferred or an Advanced Certificate in areas such as Agile Systems, Big Data, Data Science. The three Key Requirements for experience for these positions were core data engineering/data processing/data warehousing or ETL (Extract/Transform/Load) capability. There was also an emphasis on data at scale (i.e. large IT implementations or large amounts of raw data). This experience is frequently quantified and is a primary requirement of these roles. However, there was little focus on knowledge requirements, beyond business intelligence and database technologies. The ability to work collaboratively was highlighted in many positions, alongside written communication skills and the ability to solve problems or trouble-shoot in the working environment. Whilst a broad range of technical competencies were listed, there was a strong focus on Hadoop/MapReduce and associated technologies, such as Hive and Pig. Non-relational databases were also cited, including MongoDB and neo4j, accompanying requirements for a selection of programming and scripting languages.

In the additional Other Requirements category, we observed '*TS/SCI Polygraph*' and '*Background verification check*' as security requirements for some positions, but once again these were not designated as Key Requirements, as they occurred in less than three Data Engineer positions.

Table 3. Key requirements for data engineer.

Education	Experience	Knowledge	Skills	Competencies
Bachelor's degree in Computer Science, Math, Statistics, Information Systems	Experience with data engineering / processing / warehousing / software	Knowledge of BI technologies	Ability to work collaboratively in teams or with clients	Competence with Hadoop / MapReduce / Pig / Hive
Bachelor's degree in Business, IT	Experience with large IT engagements / large amounts of raw data	Knowledge of databases / SQL	Oral presentation and documentation / written communication skills	Competence with Python
Master's degree	Experience of databases and ETL		Ability to multi-task and prioritize work	Competence with SQL
	Quantified experience e.g. 2-5+ years		Problem-solving and trouble-shooting skills	Competence with databases neo4j, MongoDB

Table 3. Key requirements for data engineer *(continued)*

Education	Experience	Knowledge	Skills	Competencies
				Competence with programming and scripting languages PERL, Java, Ruby
				Competence with data analysis products MS Excel, R, SAS
				Competence with Unix / Linux

Our recommended course pathways through the MLIS and Information Science and Technology Masters Programs for a prospective Data Engineer, include the essential, desirable and additional course stepping stones listed in Table 4.

Table 4. Course pathways for a data engineer.

Essential Masters courses from the iSchool	Desirable Masters courses from the iSchool	Additional courses from other Schools and Departments
Data structures	GIS systems	Business
Database management	Technologies and services for digital data	Mathematics
Advanced topics in database management	Web technologies and standards	Statistics
Cloud computing	Mathematical foundations for information science	
Data analytics	Research data infrastructures	
Algorithm design	E-Business	
Information security and privacy		
Software engineering		
Decision analysis and decision support systems		

Data Journalist

The Data Journalist positions seek candidates with similarly substantive and quantified experience as a journalist or reporter (Table 5). Experience with statistical work, data visualization or graphics, were also listed. However, there is no specific education requirement beyond a Bachelor's degree. Stated knowledge requirements are rare,

 doi:10.2218/ijdc.v11i2.417

although mathematics or statistics or a particular domain area relevant to the position were listed in some job descriptions. The skills requirements reflected those observed in the other two positions: oral and written communication skills, collaborative skills and an attention to detail. Time management/ability to meet deadlines was also a Key Requirement. The widest range of competencies was observed for the Data Journalist roles, encompassing programming and scripting languages, visualisation and graphics software, cartographic or mapping tools, web authoring, data analysis packages and database query methods (SQL).

Table 5. Key requirements for data journalist.

Education	Experience	Knowledge	Skills	Competencies
Unspecified Bachelor's degree	Experience as a journalist or reporter	Knowledge of mathematics or statistics	Oral and written communication skills	Competence with programming and scripting languages Python, JavaScript, Node
	Experience with statistical work	Knowledge of domain reporting area	Ability to work collaboratively in diverse teams or with clients	Competence with visualisation and graphics software Adobe Creative, D3
	Experience with data visualisation or graphics creation		Attention to detail and accuracy skills	Competence with cartography tools QGIS, ArcGIS, TopoJSON
	Experience with data acquisition		Time management skills and ability to meet deadlines	Competence with web markup and style sheets (CSS)
	Quantified experience			Competence with statistical analysis products R, MS Excel
				Competence with SQL

In the additional Other Requirements category, we observed that employers desired the submission of a portfolio, via either clippings or a link to a web-published portfolio. This was not designated as a Key Requirement as it occurred in less than three Data Journalist positions.

Our recommended course pathways through the MLIS and Information Science and Technology Masters Programs for a prospective Data Journalist, include the essential, desirable and additional course stepping stones listed in Table 6.

Table 6. Course pathways for a data journalist.

Essential Masters courses from the iSchool	Desirable Masters courses from the iSchool	Additional courses from other Schools and Departments
Data analytics	Technologies and services for digital data	English
Spatial data analytics	Data mining	Communication
Information visualisation	Cloud computing	Business
Algorithm design	Software engineering	Mathematics
Information security and privacy	Mathematical foundations for information science	Statistics
Web technologies and standards	Research data management	
Web services and distributed computing	Research data infrastructures	
E-Business		

Discussion

Whilst this is a modest study, the methodology has been effective in identifying the features and characteristics of each of the three positions investigated. The wealth of Data Analyst, Data Engineer and Data Journalist positions within the job bank searched is evidence of the continuing investment in and growth of data-driven markets and the accompanying huge demand for a workforce with the critical skills in these areas. The distribution of positions in the sample highlights the value placed on these data roles by private sector organisations; to some extent, universities and other public sector bodies appear to be slower in investing in these particular data roles.

Comparing the Data Roles

The results from this study indicate that these are three clearly differentiated data roles, but with overlapping requirements and a common core set of critical competencies and skills. The commonalities and differences in requirements have been summarised in a Venn diagram shown in Figure 1. The focus on quantified experience for a Data Engineer and a Data Journalist may reflect parallel foundations in professional practice: both fields have an established 'hands-on' approach with strong traditions of learning on-the-job. Similarly, the requirement for domain knowledge for a Data Analyst (e.g. in health or finance or aquatic sciences) and a Data Journalist may reflect their situation within a particular disciplinary field or sector, where an understanding of the established practices, politics and culture will be an advantage. Other bilateral commonalities, such as the requirement to source or acquire data for a Data Analyst and a Data Journalist, reflects the importance of being able to 'find data' from external sources, e.g. government datasets, for subsequent exploration, visualization and insight development. The focus on large volumes of data or data aggregation observed in the data analyst and

 doi:10.2218/ijdc.v11i2.417

data engineer requirements highlights the importance of working at scale; many of the roles in the sample were based in very large multi-national companies with millions of clients generating huge data volumes through retail, business or leisure transactions (in other words, big data). The relevance of statistical skills for the Data Journalist roles in addition to the Data Analyst roles was a surprise; however quantitative data is critical for both roles and statistical techniques provide the essential tools and protocols to demonstrate significance, trends and insights from the evidence base. The value of mathematics, statistics and quantitative thinking was identified in an earlier Data Science Venn Diagram (Conway, 2010).

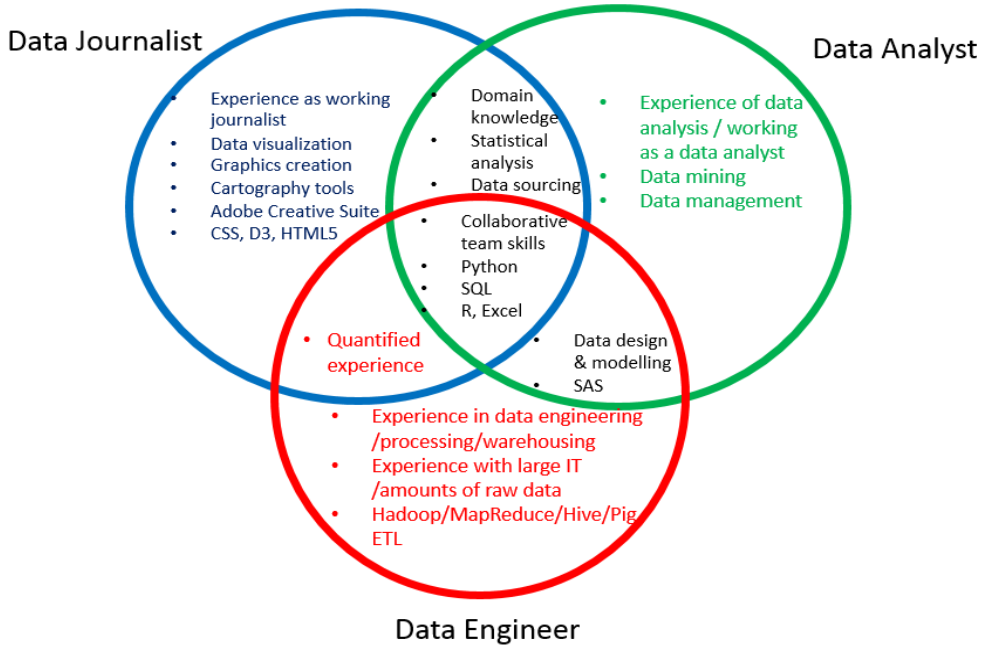

Figure 1. Data roles and requirements Part 2.

The suite of competencies required by each of the three roles (Python, R, Excel and SQL) form a foundational 'technical data toolkit' and highlight the relevance of coding and querying proficiency. Demand for Python programming expertise was found to have increased by almost 100% in big data related positions in 2014 in an analysis of big data hiring trends (Columbus, 2014). Python, R and Excel were also highlighted as key tools for data analysts for the data wrangling process – '*the process of making data useful*' (Kandel et al., 2011). However, these technical abilities need to be blended with other attributes such as research skills (Data Analyst), documentation skills (Data Engineer) and an ability to meet deadlines (Data Journalist). A blended or rounded set of skills was also highlighted as a desirable feature by UK business representatives in the Nesta Model Workers Report (Bakhshi at al., 2014).

The relative lack of commonality in requirements with the three roles previously studied is striking (Lyon, Mattern, Acker and Langmead, 2015). Whilst there are some requirement intersections (e.g. data management, relational databases and data visualization), overall these form two largely separate groups, each with three inter-related roles. However, within a data-intensive marketplace, the roles are inter-dependent: a Data Analyst, Journalist or Engineer requires high quality, curated data to work with, whilst a Data Librarian, Archivist, or Steward/Curator require the data in their care to be used, wrangled and analysed to demonstrate their value.

iSchool Curriculum Development

The paper demonstrates that the curriculum requirements for the data analyst and the data engineer roles are very well-matched to the iSchool curriculum, with potential collaborative opportunities with other academic schools, such as the School of Engineering to enrich and supplement the iSchool offer. However, it can be argued that the delivery of the educational curriculum for the Data Journalist role may be best positioned within a school providing journalism, media, communication, English or creative writing programs, with the additional collaborative opportunities arising in reverse with iSchools. There is also further scope for the development of Advanced Certificates; whilst this was not identified as a Key Requirement, it was highlighted as a requirement for some positions across each of the roles. Such qualifications provide an effective route for up-skilling of current professionals.

The model pathways described in this study appear to be similar to the concept of a 'trajectory' posited by Furst, Isbell and Guzdial (2007), who also present a 'threads' approach to reviewing the Computer Science curriculum at The Georgia Institute of Technology in Atlanta. Threads takes a view beginning with courses or modules and leading out to generic career roles such as 'Practitioner' (software engineer); in contrast the translational approach at the University of Pittsburgh begins with real-world roles and tracks back through the role requirements to the courses and modules offered by the graduate programs. Both methodologies have their value, since in each case they join up the educational offerings with career options, current workforce trends and future market demands.

Conclusions

Whilst this is a modest study, the methodology is transferable and may be applied within other iSchools and by other education providers. The findings emphasise the inter-disciplinary, blended or hybrid character of the curriculum requirements for the data science roles. Higher education providers will need to carefully customise and modify their curricula to optimally match these complex real-world requirements. However, there are significant opportunities to develop new partnerships, both across campus and beyond, to create exciting translational curricula to meet current and future data workforce capacity and capability challenges.

References

Bakhshi, H., Mateos-Garcia, J. &Whitby, A. (2014). Model workers: How leading companies are recruiting and managing their data talent. Nesta Report. Retrieved from http://www.nesta.org.uk/sites/default/files/model_workers_web_2.pdf

BRDI. (2015). *Preparing the workforce for digital curation.* National Academies Press. Retrieved from http://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation

Columbus, L. (2014). Where big data jobs will be in 2015. Forbes Magazine Technology Blog. Retrieved from http://www.forbes.com/sites/louiscolumbus/2014/12/29/where-big-data-jobs-will-be-in-2015/

Conway, D. (2010) The data science venn diagram. Retrieved from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Finzer, W. (2013). The data science education dilemma. *Technology Innovation in Statistics Education, 7*(2). Retrieved from http://escholarship.org/uc/item/7gv0q9dc

Furst, M., Isbell, C., & Guzdial, M. (2007). Threads: How to restructure a computer science curriculum for a flat world. SIGCSE'07, March 7-10, 2007, Kentucky. Retrieved from http://www.cc.gatech.edu/~isbell/papers/isbell-threads-sigcse-2007.pdf

Gray, J. Bounegru, L. & Chambers, L. (Eds). (2011). The data journalism handbook: How journalists can use data to improve the news. O'Reilly.

Kandel, S., Heer, J, Plaisant, C., Kennedy, J., van Ham, F., Riche, N.H., Weaver, C., Lee, B., Brodbeck, D., & Buono, P. (2011). *Research directions in data wrangling: Visualizations and transformations for usable and credible data.* Information Visualization. Retrieved from http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf

Lee, C.H. (2014). Three data careers decoded and what it means for you. Udacity Blog. Retrieved from http://blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html

Lyon, L. & Brenner, A. (2015). Bridging the data talent gap: Positioning the iSchool as an agent for change. *International Journal of Digital Curation, 10*(1), 111-122. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/10.1.111/384

Lyon, L., Mattern, E., Acker, A., & Langmead, A. (2015). Applying translational principles to data science curriculum development. iPres Conference Proceedings, Chapel Hill, November 2015. (In Press).

Patil, D.J. (2011). Building data science teams. O'Reilly Radar.

Rivera, R. & Haverson, A. (2014). Data scientist vs data analyst. Captech Insights Blog. Retrieved from https://www.captechconsulting.com/blogs/data-scientist-vs-data-analyst

Stanton, J., Palmer, C.L., Blake, C., & Allard, S. (2012). Interdisciplinary data science education. Special Issues in Data Management. American Chemical Society Symposium Series. Retrieved from http://pubs.acs.org/doi/pdf/10.1021/bk-2012-1110.ch006

Walker, M. (2013). Data scientists vs data engineers. Data Science Central Blog. Retrieved from http://www.datasciencecentral.com/profiles/blogs/data-scientists-vs-data-engineers

Woolf, S.H. (2008). The meaning of translational research and why it Matters. *JAMA 299*(2), 211-213. Retrieved from http://jama.jamanetwork.com/article.aspx?articleid=1149350

 doi:10.2218/ijdc.v11i2.417

Appendix

Table 7. Job postings examined [30]

Data Analyst	Data Engineer	Data Journalist
Data Analyst (Wayne State University)	Senior Big Data Engineer (Trulia)	Data Journalist (Hanley Woods Media)
Data Analyst (Bith Group Technologies)	Data Engineer (IBM Analytics)	Data Journalist (Dow Jones)
Data Analyst (Hsssoft)	Data Engineer (Deloitte Consulting)	Data Journalist (Bloomberg)
Data Analyst (Biogensys)	Data Engineer (Shape Up)	Data-Drive Journalist (Inter-American Development Bank)
Data Analyst (Norwalk Community Health Center)	Data Engineer (Hulu)	Data Visualization Specialist/Data Journalist (voxgov)
Data Analyst (Terranova Consulting)	Data Engineer (Boston Consulting Group)	Data Reporter (Thomson Reuters)
Data Analyst (University of Washington Institute for Health Metrics and Evaluation)	Data Engineer (AWS)	Web Developer/Visual Journalist (ESPN)
Data Analyst (Tufts University Energy Metabolism Lab)	Data Engineer (IMS Health)	Newsperson/Data Journalist (Associated Press)
Data Analyst III (UTAH State University)	Data Engineer (Stepping Up Solutions)	Visual Journalist (Sports Media Network)
Data Analyst (IT America)	Big Data & Data Management Engineer (American Express)	Data Journalist (United Software Group)