Preparing Undergraduate Data Scientists for Success in the Workplace: Aligning Competencies with Job Requirements

Dr. Duo Li, Shenyang Institute of Technology

Dr. Duo Li is an associate professor of Big Data Management and Application major at the School of Economics and Management, Shenyang Institute of Technology. Duo Li is a member of ASIST&T, and his research interests are focused on Human-Computer-Interaction, Big Data, and Data Analytics.

Dr. Elizabeth Milonas, New York City College of Technology

Elizabeth Milonas is an Associate Professor with the Department of Computer Systems at New York City College of Technology - City University of New York (CUNY). She currently teaches relational and non-relational databases and data science courses to undergraduate students. She holds a BA in Computer Science and English Literature from Fordham University, an MS in Information Systems from New York University, and a Ph.D. from Long Island University. Her research interests focus on three key areas: data science curriculum and ethics, retention of minority students in STEM degree programs, and organization and classification of big data.

Dr. Qiping Zhang, Long Island University

Dr. Qiping Zhang is an Associate Professor in the Palmer School of Library and Information Science at the C.W. Post Campus of Long Island University, where she also serves as director of the Usability Lab. Dr. Zhang holds a Ph.D. and an M.S. in informatio

Preparing Undergraduate Data Scientists for Success in the Workplace: Aligning Competencies with Job Requirements

1. Introduction

The increased use of Data Science technologies, particularly artificial intelligence and machine learning has caused an increase in demand for skilled Data Science professionals [1,2,3]. This demand is driven by the rising dependence of businesses on these technologies to inform strategic decisions [1,2,3]. The Data Science domain is multidisciplinary, encompassing skill sets, including statistics, business, and computer science [3]. Data Science professionals are expected to master this diverse skill set [3]. However, the Data Science domain is constantly and rapidly changing as new technologies are incorporated into the field [3]. This ever-evolving landscape poses a difficult challenge to universities tasked with educating the next generation of data scientists. To adequately prepare students for the dynamic demands of the Data Science domain, the data science competencies taught in university courses must align with the required skills demanded by industry.

This study analyzes the alignment between Data Science competencies taught in 136 undergraduate Data Science programs across the United States [4] and the skills required for full-time, entry-level undergraduate degree Data Science jobs as listed on Indeed. The research question answered as a result of this study is:

> *RQ: Do the Data Science competencies taught in undergraduate programs align with the required skills for entry-level, full-time Data Science jobs?*

2. Literature Review

2.1 Data Science Job Skills

Many studies have evaluated data science job requirements and the needed skills. A recent study [5] examined 11,965 paragraphs of text from 1200 USA job announcements. It determined that a balance of statistical, machine learning, and programming skills are required for data scientists. An earlier study [6] found that data scientist jobs require a focus on skills such as database management (SQL programming, intelligence design, data warehousing), programming (problem-solving, languages such as Python, Java), project management (planning, project analysis, risk reporting), data analytics (computer learning, programming, statistical modeling), and business impact (consulting, market delivery, strategic management). Results [7] from an analysis of 1050 unique records of Data Science job requirements showed that technical skills are in high demand when seeking Data Scientists. These skills include proficiency in Big Data Technologies, software development, data management, analytic methods, algorithms, programming languages, and analytic tools. In addition, the study findings [7] showed demand for soft skills (non-technical and interpersonal skills) such as information management, communications, and presentation skills. Similarly, study results from an analysis [8] of 1216 job advertisements showed a need for soft skills in posting related to data science positions. In addition, study results showed a need for design and development skills, including systems development methodologies, particularly for analytical systems. Soft skills and technical

expertise were also found to be vital skills in a study [9] that analyzed a total of 3009 job posts from LinkedIn.com, Monster.com, and Procom.com. The results from this study [9] emphasized the importance of people skills and positive personality traits in addition to various technical skills, including programming and Big Data processing. The results [9] also identified the need for STEM professionals to possess business skills and knowledge of non-STEM domains.

2.2 Data Science Curriculum

Research geared towards understanding the effectiveness of the curriculum in job preparedness shows that higher education needs to evaluate the curriculum to better prepare graduates for jobs in the Data Science domain. The needs for on-the-job skills were outlined in a study [10] in which 24 industry professionals from various industry sectors, including technology, education, transportation infrastructure, and manufacturing, were interviewed. The interview results showed that industry professionals would like a data science curriculum to help students gain broad foundational skills that can be easily transferred across jobs. These professionals indicated that as AI becomes more in demand, data scientists will need more than technical skills, and they recommend that foundational skills, which include abstract thinking, problem-solving, human-centered design, and liberal arts education, be infused into the Data Science curriculum. In a similar study [11], professionals were interviewed, and their perspectives on curriculum and readiness align with the findings of [10]. Professionals emphasized the need to teach soft skills, including written and oral communication, interpersonal skills, leadership, self-motivation, self-confidence, negotiation, and adaptability. In addition, study results [11] identified time management and listening as important skills for students to possess while in the workforce and important skills to be taught in any STEM curriculum.

3. Method

3.1 Dataset for Data Science Undergraduate Programs

One hundred and thirty-six (136) colleges were identified in an earlier study [4], and the degree program curriculum in their data science degrees was examined. Course names and descriptions were used to map the degree program curriculum for each of the 136 colleges to the Data Science Competencies identified in the earlier study [4]. This information was taken from the websites of each of the 136 colleges in the study. Course content was not analyzed for this study; only course names and descriptions were included in the analyses.

3.2 Data Science Competencies

The degree program curriculum and data science competencies used in this study were identified in an earlier study [4], which examined 136 colleges and their undergraduate Data Science degree program curriculum. The competencies detailed in Table 1 are drawn from the Data Science Task Force of the Association of Computing Machinery (ACM) report[4], which identified 11 core data science competencies shown in Table 1.

Table 1: Data Science Competencies and Sub-topics by 2021 ACM Data Science Task Force

ACM Data Science Task Force Report Competencies	
1. Analysis and Presentation • Foundational considerations • Visualization • User-centered design • Interaction design • Interface design and development	**7.** Data Privacy, Security, Integrity, and Analysis for Security • Data privacy • Data security • Data integrity • Analysis for security
2. Artificial Intelligence • General • Knowledge representation and reasoning – logic-based • Knowledge representation and reasoning – probability-based • Planning and search strategies	**8.** Machine learning • General • Supervised learning • Unsupervised learning • Mixed methods • Deep learning
3. Big Data Systems • Problems of scale • Big data computing architectures • Parallel computing frameworks • Distributed data storage • Parallel programming • Techniques for Big Data Applications • Cloud computing • Complexity theory • Software support for Big Data applications	**9.** Programming, data structures, and algorithms • Algorithmic thinking and problem-solving • Programming • Data structures • Algorithms • Basic complexity analysis • Numerical computing
4. Computing and Computer Fundamentals • Basic computer architecture • Storage systems fundamentals • Operating system basics • File systems • Networks • The Web and web programming • Compilers and interpreters	**10.** Software development and maintenance • Software design and development • Software testing
5. Data Acquisition, Management, and Governance • Data acquisition • Information extraction • Working with various types of data • Data integration • Data reduction and compression • Data transformation • Data cleaning • Data privacy and security	**11.** Professionalism • Continuing professional development • Communication • Teamwork • Economic considerations • Privacy and confidentiality • Ethical considerations • Legal considerations • Intellectual property • On automation
6. Data Mining • Proximity measurement • Data preparation • Information extraction • Cluster analysis • Classification and regression • Pattern mining • Outlier detection • Time series data • Mining web data • Information retrieval	

The research cited in [12] streamlined the 11 competencies outlined by the ACM Data Science Task Force into 7 distinct competency categories (see Table 2). This restructuring was based on a comprehensive examination of curriculum requirements from 136 colleges offering data science programs. The objective was to harmonize the competencies defined by the 2021 ACM Data Science Task Force [12] with the prerequisites of modern data science degree programs. These 7 distinct competency categories serve as the framework for the present study.

Table 2: Data Science Competency (Milonas, Li, & Zhang, 2022)

Category	Competency Categories (Milonas, Li & Zhang, 2022)	Data Science Competencies (Corresponding ACM Task Force)
1	Computing Fundamentals	4. Computing and Computer Fundamentals 9. Programming, data structures, and algorithms 10. Software development and maintenance
2	Data Management, Governance, Privacy	5. Data Acquisition, Management & Governance 7. Data Privacy, Security, Integrity, Analysis for Security 11. Professionalism
3	Data Visualization	1. Analysis and Presentation
4	Machine Learning	8. Machine learning 2. Artificial Intelligence
5	Data Mining, Big Data	6. Data Mining 3. Big Data Systems
6	Data Science in Context	11. Professionalism
7	Math and Statistics	9. Programming, data structures and algorithms

3.3 Dataset for Data Science Jobs

Data Science Job Data was collected from Indeed.com in November 2023. A total of 1512 data elements were gathered based on the search criteria identified in the following Step 1.

Step 1: Identify Matching Jobs

The search term "Data Scientist Entry Level" was used to identify matching jobs. Filtering criteria included Job Type = ”full-time,” Experience Level = ”Entry Level,” and Education = ”Bachelor’s degree.”

Step 2: Identify Job Skills

Table 3 outlines the dictionary of data science skills extracted from our analysis of 136 degree program courses. These skills align with the seven distinct competency categories highlighted in Table 2.

Table 3: Data Science Job Skill Dictionary

Category	Competency Categories	Keywords for Data Science Job Skills (Identified from Corresponding Sample Courses)
1	Computing Fundamentals	SQL Programming, Introduction to Programming, Algorithms, Data Structures, Object Oriented Programming, Software Engineering, Systems Analysis and Design, Human-Computer Interaction
2	Data Management, Governance, Privacy	Data Warehousing, SQL, Databases, Security, Fraud Detection, Network Security, Ethics
3	Data Visualization	Data Visualization
4	Machine Learning	Machine Learning, Data Modeling, Artificial Intelligence, Deep Learning
5	Data Mining, Big Data	Data mining, Data modeling, systems analysis, Big Data, Data munging
6	Data Science in Context	Capstone, Internship, Senior Project, physics, biology, chemistry, humanities
7	Math and Statistics	Calculus, discrete structures, probability theory, elementary statistics, statistics, linear algebra.

Step 3: Define Geographic Regions

Table 4 below identifies the U.S. Geographical regions used in the earlier study [12], the states in each region and the representative cities used in the Indeed search to acquire job data for the present study.

Table 4: Regions, States in Regions, and Representative Cities in Regions

Regions	**States from which Degree Program Curriculum Data was gathered**	**Representative Cities used in Indeed search to acquire job data**
Northeast	Connecticut, Massachusetts, Maine, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont	New York and Boston
West	Alaska, California, Colorado, Hawaii, Idaho, Montana, Nevada, Oregon, Utah, Washington, Wyoming	San Francisco
Midwest	Iowa, Illinois, Indiana, Kansas, Michigan, Minnesota, Missouri, North Dakota, Nebraska, Ohio, South Dakota, Wisconsin	Chicago
Southwest	Arizona, New Mexico, Oklahoma, Texas	Austin
Southeast	Alabama, Arkansas, DC, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, West Virginia	

Table 5 below summarized the total number of Data Science undergraduate programs analyzed in each region (from the earlier study [12]) and the total number of jobs found in each representative cities from Indeed.com.

Table 5: Regional Distribution of Data Science Undergraduate Programs and Data Science Jobs by Five Cities

Region	N. of Degree Programs	Representative City	N. of Job Data
Midwest	39	Chicago	222
West	21	San Francisco	411
Northeast	55	Boston New York	304 401
Southwest	7	Austin	174
Southeast	12		
Total	**136**		**1,512**

Step 4: Align Data Science Job Skills with Data Science Competencies

A Python script was programmed to scrape job data from Indeed.com, focusing on identifying the required job skills. Using keywords selected from the dictionary in Table 3, we aligned extracted job skills with corresponding competencies.

If a job description contained any selected keywords in Table 3, the matching result was counted as "1." Subsequently, the system calculated the aggregate score for each competency based on the frequency of course-related keywords. If a company required more than one programming language, the score of the analysis result was counted as "1." We used the same scoring rule for each competency for textual analysis of the results. Such a scoring process was achieved by programming with VBA in Excel.

4. Results

4.1 Percentage of Competency Coverage in Data Science Curriculum

Table 6 reported the percentage of data science competency coverage in degree programs across various regions.

Table 6: Percentage of Competencies Coverage in Degree Programs across Regions

	Competencies	MIDWEST (39)	WEST (12)	NORTHEAST (55)	SOUTHWEST (7)
1	Computer Fundamentals	87%	100%	85%	100%
2	Management/ Governance/ Privacy	74%	71%	87%	86%
3	Data Visualization	56%	62%	49%	29%
4	Machine Learning	64%	71%	56%	29%
5	Data Mining/ Big Data	51%	48%	64%	71%
6	Data Science in Context	64%	62%	78%	57%
7	Math/ Statistics	97%	100%	100%	100%

4.2 Percentage of Competency Required in Job Postings

Figure 1 compares required job skills corresponding to data science competencies from the five cities in this study (New York, Boston, San Francisco, Chicago, and Austin). Given the total number of jobs from different cities in our dataset, we normalize data by dividing competency frequency by the total number of jobs for each city.

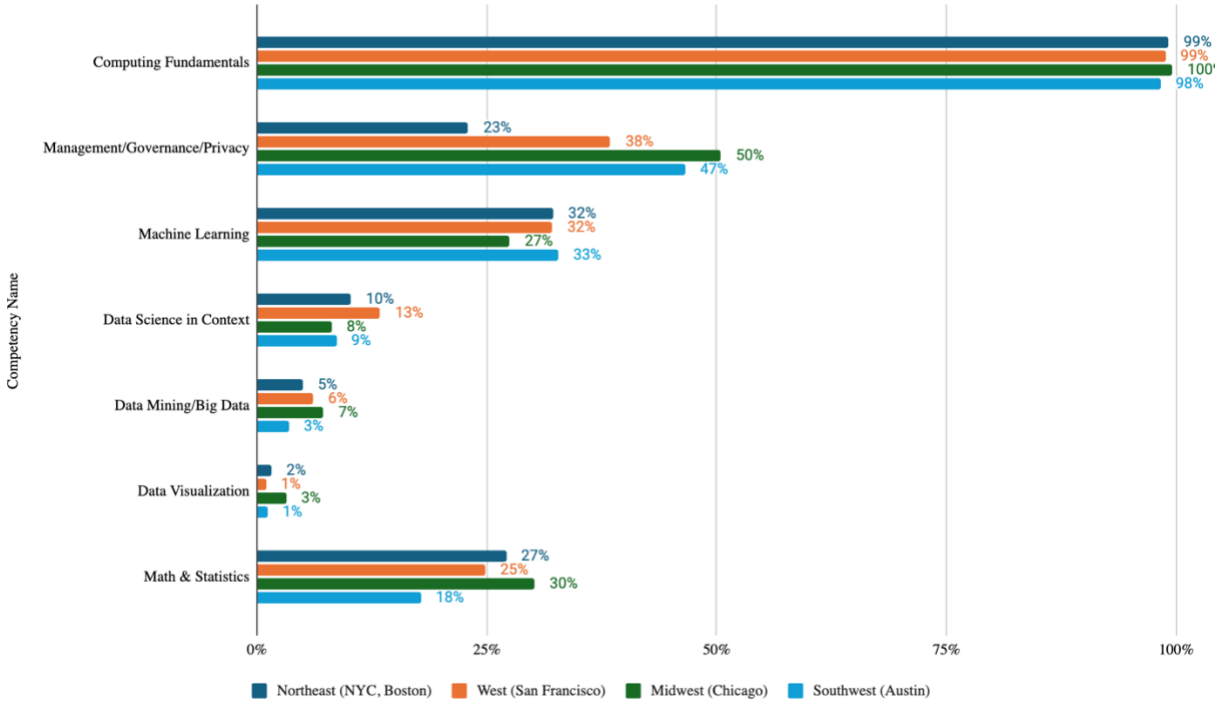

Figure 1: Comparison of Competencies Required in Data Science Jobs by Region

Our data revealed interesting results. We divide them into three competency groups.

First, *immediate demand competency group*: the top-ranking competencies (23%-100%) include 1) **Computing Fundamentals**, 2) **Data Management**, and 3) **Machine Learning**. Competencies in this group are consistently required in job skill listings in all four regions. This group represents practical skills immediately required by current job markets: programming, management, and machine learning. More than 90% of jobs require proficiency in the Computing Fundamentals competency across all five cities. This emphasizes the vital nature of possessing programming skills in today's job market. The Data Management competency is required in 23-50% of jobs. The Machine Learning competency is required in about 30% of all jobs.

Second, *low demand competency group:* Interestingly, competencies (ranging from 1%-13%) including Data Mining, Big Data, and Data Visualization are not widely required. The Data Science in Context competency is only required for 8-13% of jobs. Few jobs require the Data Mining, Big Data (3-7%) and Data Visualization (1-3%) competencies.

Third, *long-term foundational competency group:* Though the Math & Statistics competency is not as practical as programming and management, it is a foundational skill

required for people to solve real-world problems. The Math and Statistics competency is required for 27-30% of jobs, similar to management and machine learning competencies.

4.3 Overall Alignment of Job Skills with Degree Program Curriculum Competencies

Figure 2 shows the overall alignment of competencies in job skill requirements with competencies in data science degree program curriculum coverage.

Overall, the competencies covered in the data science degree program curriculum (in blue bars) in Figure 2 are more than the competencies required in job skills (in orange bars). In addition, all competency percentages are much higher in the data science degree program curriculum than in job requirements, with the exception of the computing fundamentals competency, which shows the opposite but only a slight margin (9%).

This result suggested that most current data science degree programs teach the job skills required in the field of data science. Though some competencies (e.g., low demand competency group - data science in context, data mining/big data, data visualization, and foundational group - math & statistics) are not immediately demanded in the job markets, they will prepare students for future development. It will be interesting to watch the trend and to conduct a follow-up study in the next 5-10 years to see whether such alignment will change as new technologies in the data science field progress.

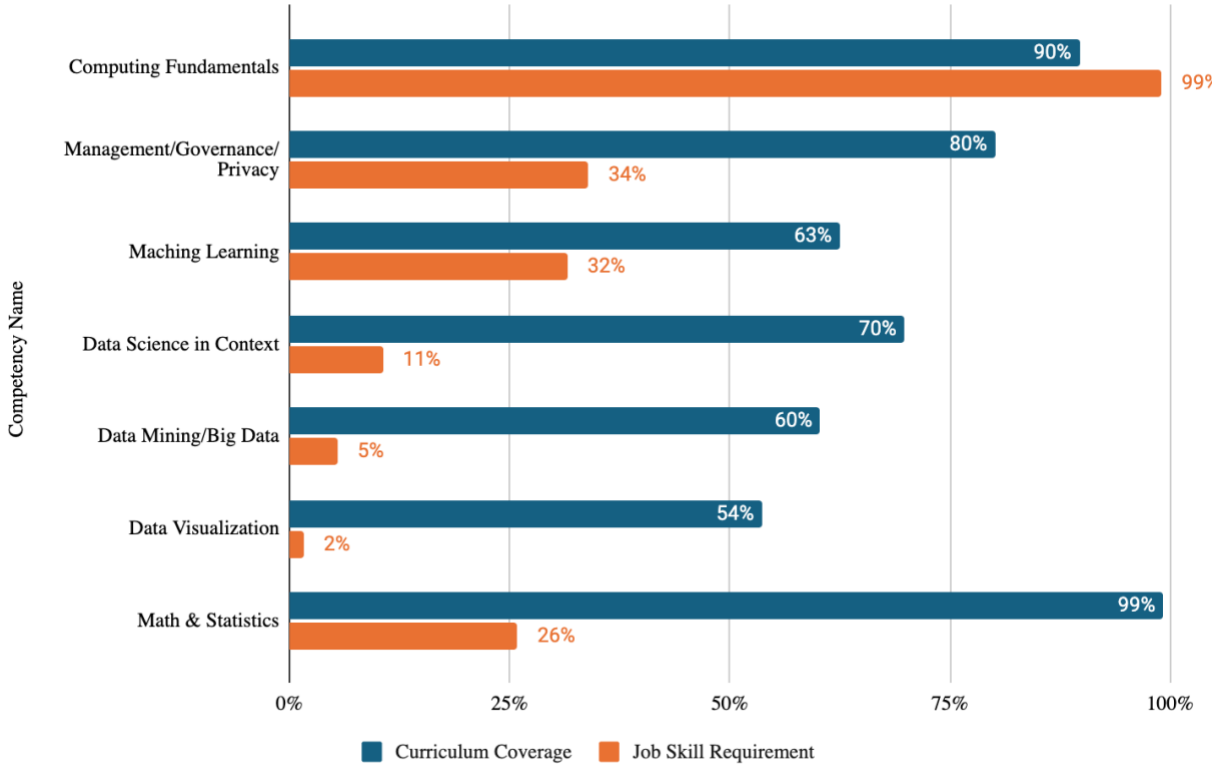

Figure 2: Overall Alignment of Competencies in Degree Program Curriculum Coverage within Job Requirements

4.4 Regional Alignment of Job Skills with Program Curriculum Competencies

In the following, we will present the results of aligning competency percentage in data science degree program curriculum coverage with competency percentage in job skill requirements for each region.

4.4.1 Region 1: Northeast Region (NYC & Boston)

The Northeast region data consisted of job data for New York and Boston. We used the job data in these two cities to represent the Northeast region. As shown in Figure 3, the competency percentage in job postings in New York and Boston was compared with that in degree program curriculum data for the Northeast region, as outlined in [12]. The findings are similar to the overall alignment results above.

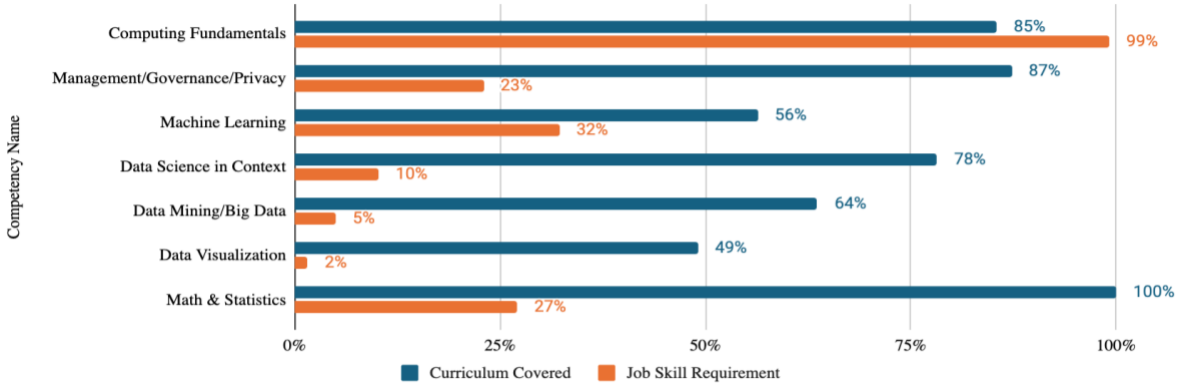

Figure 3: Alignment of Competencies in Degree Program Curriculum Coverage within Job Requirements (Northeast Region)

4.4.2 Region 2: West Region (San Francisco)

The West region data consisted of job data for San Francisco. As shown in Figure 4, the competency percentage in job postings in San Francisco was compared with that in degree program curriculum data for the West region, as outlined in [12]. The findings are similar to the overall alignment results above, except for the Computing Fundamentals competency, which is taught by all data science program curriculums in this region. Interestingly, Silicon Valley is in this region. It appears this region highly emphasizes Computing Fundamentals, which perfectly aligns with the job requirements.

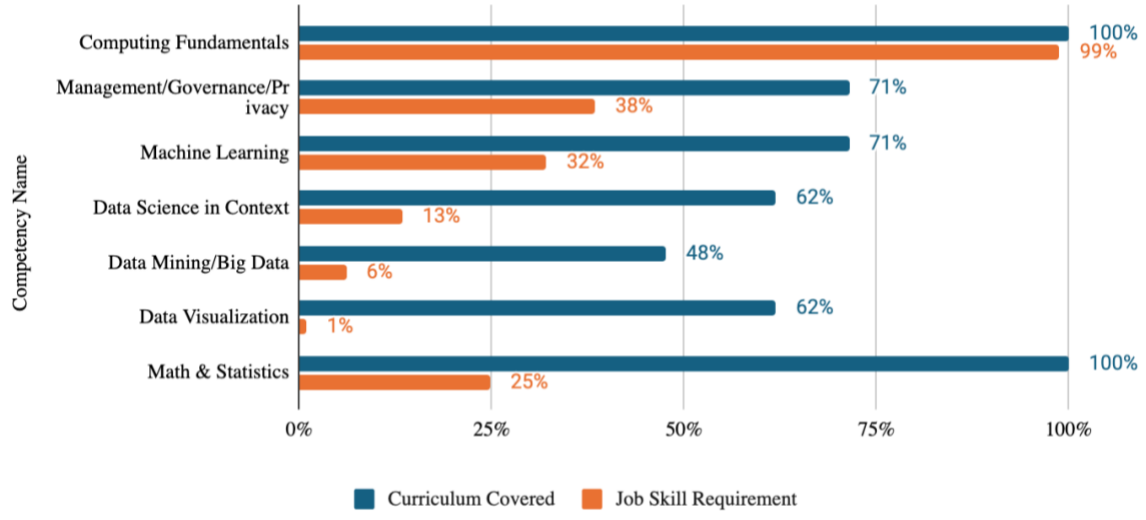

Figure 4: Alignment of Competencies in Degree Program Curriculum Coverage within Job Requirements (West Region)

4.4.3 Region 3: Midwest Region (Chicago)

The Midwest region data consisted of job data for Chicago. As shown in Figure 5, the competency percentage in job postings in Chicago was compared with that in degree program curriculum data for the Midwest region, as outlined in [12]. The findings are similar to the overall alignment results above except for the Management competency, which is required by more job postings (50%) than in overall job postings (34%).

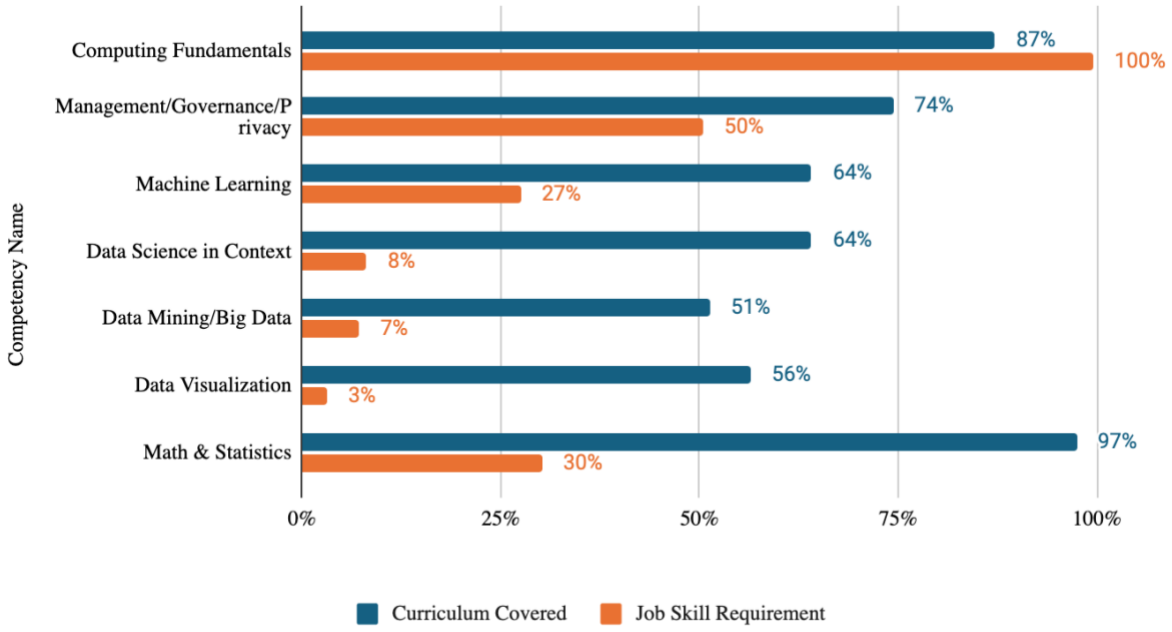

Figure 5: Region 3 Alignment of Competencies in Degree Program Curriculum Coverage within Job Requirements (Midwest Region)

4.4.4 Region 4: Southwest Region (Austin)

The Southwest region data consisted of job data for Austin. As shown in Figure 6, the competency percentage in job postings in Austin was compared with that in degree program curriculum data for the Southwest region, as outlined in [12]. Similar to the results for the West region, the Management competency is required in more job postings (47%) than in overall job postings (34%). In addition, this region shows unique findings related to the Machine Learning competency, in which more jobs (33%) are required than degree curriculum percentage (29%). In recent years, many high-tech companies (such as Tesla and Apple) have built their regional headquarters or factories in the Austin area. Current data science degree programs may need to update their curriculum to respond to the increasing job requirement for Machine Learning skills.

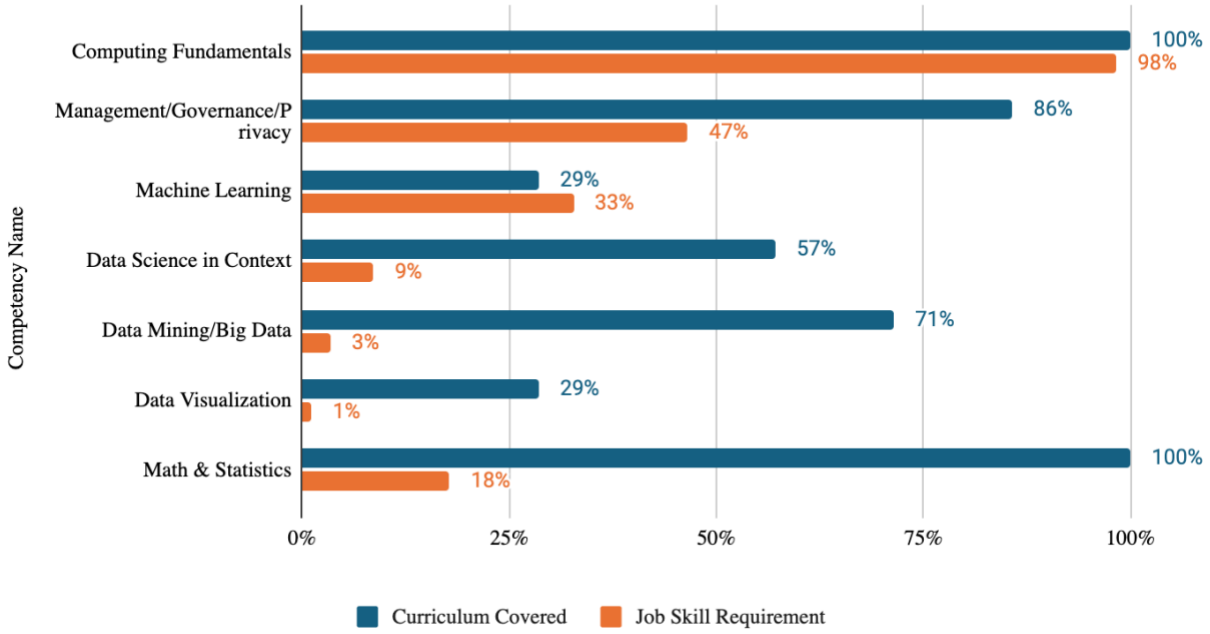

Figure 6: Region 4 Alignment of Competencies in Degree Program Curriculum Coverage within Job Requirements (Southwest Region)

5. Conclusion

In summary, our study revealed that the overall competencies covered in the data science degree program curriculum are more than those competencies required in job skills, suggesting that most current undergraduate data science programs effectively prepare students to meet the minimum job requirements, regardless of the geographic regions.

Each region also demonstrated some unique regional characteristics. Management competency is increasingly sought after in job postings in the Midwest and Southwest regions, whereas in the West region, the emphasis is more on the Computing Fundamentals competency. Furthermore, employers in the Southwest region prioritize proficiency in Machine Learning competency. Here, the percentage of Machine Learning competency in job postings exceeds the percentage in the degree program curriculum.

Three data science competency groups are in alignment (data science degree program curriculum with data science job requirements): high immediate demand competency group, low demand competency group, and long-term foundational competency group. A follow-up study in the next 5-10 years can be conducted to assess whether such alignment evolves with the advancement of new technologies in the data science field.

One of the limitations of this study is that it only examined job requirement data from representative metropolitan cities in each region instead of analyzing the job information for the whole region. Additionally, further analysis of job requirements in different industries could offer in-depth insights into the alignment of data science in education and job requirements.

Future studies may reveal whether jobs requiring specific domain expertise require students to pursue advanced studies or degrees to meet these specific requirements effectively.

6. Work Cited

[1]. S. Gottipati, K. J. Shim, and S. Sahoo, "Glassdoor Job Description Analytics–Analyzing Data Science Professional Roles and Skills," in *2021 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1329-1336, IEEE, 2021.

[2]. B. A. Quismorio, M. A. D. Pasquin, and C. S. Tayco, "Assessing the alignment of Philippine higher education with the emerging demands for data science and analytics workforce," *PIDS Discussion Paper Series*, 2019, no. 2019-34.

[3]. M. Almgerbi, A. De Mauro, A. Kahlawi, and V. Poggioni, "A systematic review of data analytics job requirements and online-courses," *Journal of Computer Information Systems*, vol. 62, no. 2, pp. 422-434, 2022.

[4]. E. Milonas, Q. Zhang, and D. Li, "Do Undergraduate Data Science Program Competencies Vary by College Rankings?" In Proceedings of the 2022 ASEE Annual Conference & Exposition, Minneapolis, MN, August 2022. [Online]. Available: https://peer.asee.org/41153

[5]. M. A. Halwani, S. Y. Amirkiaee, N. Evangelopoulos, and V. Prybutok, "Job qualifications study for data science and big data professions," Information Technology & People, vol. 35, no. 2, pp. 510-525, 2022.

[6]. A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, "Human resources for Big Data professions: A systematic classification of job roles and required skill sets," Information Processing & Management, vol. 54, no. 5, pp. 807-817, 2018.

[7]. Z. Radovilsky, V. Hegde, A. Acharya, and U. Uma, "Skills requirements of business data analytics and data science jobs: A comparative analysis," Journal of Supply Chain and Operations Management, vol. 16, no. 1, pp. 82-101, 2018.

[8]. A. Gardiner, C. Aasheim, P. Rutner, and S. Williams, "Skill requirements in big data: A content analysis of job advertisements," Journal of Computer Information Systems, vol. 58, no. 4, pp. 374-384, 2018.

[9]. A. Persaud, "Key competencies for big data analytics professions: A multimethod study," Information Technology & People, vol. 34, no. 1, pp. 178-203, 2021.

[10]. Y. Moghaddam, S. Kwan, L. Freund, and M. G. Russell, "A Proposed Roadmap to Close the Gap Between Undergraduate Education and STEM Employment Across Industry Sectors," in International Conference on Applied Human Factors and Ergonomics, pp. 363-373, Cham: Springer International Publishing, 2021.

[11]. D. McGunagle and L. Zizka, "Meeting real-world demands of the global economy: an employer's perspective," Journal of Aviation/Aerospace Education & Research, vol. 27, no. 2, pp. 59-76, 2018.

[12]. Milonas, Elizabeth, Duo Li, Qiping Zhang. “Content Analysis of Two-year & Four-year Data Science Programs in the United States.” *In Proceedings of the 128th ASEE Annual Conference, July 26-29, 2021*. [Online]. Available:

file:///Users/emilonas/Downloads/content-analysis-of- two-year-and-four-year-data-science-programs-in-the-united-states.pdf.