ISSN 2161-3974

JeSLIB 2017; 6(1): e1105
doi:10.7191/jeslib.2017.1105

Journal of eScience Librarianship
putting the pieces together: theory and practice

Video Article

Data Science Programs in U.S. Higher Education: An Interview with the Authors

Rong Tang & Watinee Sae-Lim

Simmons College, Boston, MA

Abstract

Rong Tang, Associate Professor, and Watinee Sae-Lim, Doctoral Student, from the School of Library and Information Science, Simmons College, share research presented in their article "Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus" published in the journal of *Education for Information*.

Their exploratory content analysis of 30 randomly selected Data Science (DS) programs from eight disciplines revealed significant gaps in current DS education in the United States. These findings have implications for improving DS education in iSchools and across other disciplines.

Disclosure

The original published research article can be found:

Tang, Rong and Watinee Sae-Lim. 2016. "Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus." *Education for Information* 32(3): 269-290. http://dx.doi.org/10.3233/EFI-160977

Correspondence: Rong Tang: rong.tang@simmons.edu

Keywords: data science, program description, curriculum structure, course focus, iSchools

Ms. Julie Goldman: I'm Julie Goldman, Managing Editor for the *Journal of e-Science Librarianship*, published by the Lamar Soutter Library, University of Massachusetts in Worcester, Massachusetts.

I am joined today by Associate Professor Rong Tang and Doctoral Student Watinee Sae-Lim from the Simmons College School of Library and Information Science in Boston to talk about their recent research project analyzing data science programs in United States higher education. Thank you both for sharing your work with the e-Science community.

Ms. Rong Tang: You're welcome.

Ms. Watinee Sae-Lim: You're welcome.

Ms. Goldman: So, first, your study looked at data science programs in many disciplines, including the arts and sciences, business, computer science, engineering, independent data centers, math and statistics, professional studies, and iSchools. Why do you think it is important for iSchools in particular to provide programs and training in data science?

Ms. Tang: Well, that's a great question. I don't know if there's an easy answer to it. “Data Science,” as it first started, was a phrase initially coined by Jeff Wu from Georgia Institute of Technology. In a keynote speech he made at the University of Michigan, he spoke about the change of statistics — the discipline of statistics — into data science. He believes that statistical work is a trilogy of data collection, data modeling, and analysis and decision making.

So, he popularized this notion of data science and advocated for statistics to be renamed as data science and a statistician as data scientist. For us, what's intriguing about data science is we don't think that data is purely statistics. There's some kind of conception or theoretical foundation, which ties very closely to information science.

So, when we talk about iSchools, we’re talking about the information school. And there are a lot of theories. Information science is the discipline that deals with the creation, origination, retrieval, storage, and sharing — the entire process of information. There is a conceptive affinity between information science and this newly emerged data science.

I believe information science is a matter of discipline, has a lot to contribute to data science, and probably the data science itself will also enrich the curriculum of information science as a better discipline.

Ms. Goldman: You kind of alluded to why you're interested in this.

Ms. Tang: Yes.

Ms. Goldman: Wa, why are you interested in this area of librarianship?

Ms. Sae-Lim: My interest goes back several years ago. I first did a pilot study about how researchers manage their research data. I found out that their data management behaviors did not differ from one another that much. And I also did a small-scale study about how future librarians manage their own data.

And interestingly, I found out that the pattern of their behaviors was not that much different. To me, I think in order to support our users, I think we should do better than that. And today, research is more and more data intensive. One skill set or one discipline is not sufficient to handle that massive amount of data anymore.

So, we looked at the data science program, which is the blending of disciplines. And also it is not only our schools that host this kind of program. There are various schools that offer DS programs. We were interested in their focuses, similarities, differences, and how they position themselves, especially iSchools. So, that's why we began this study.

Ms. Goldman: Your research paper is really an exploratory content analysis of these data science programs, their curriculums, and the courses they offer. What were your main research questions?

Ms. Tang: We had four research questions. This is the more simplified version of the questions we asked. The first question was "What are the linguistic characteristics of the DS program descriptions?" So, we looked at the websites of these data science programs, and we looked at how they describe the program. The second question asked "What kind of curriculum requirements do they have?" And the third question was "What proportion of the DS courses was all-domain knowledge or covering-domain knowledge?" "What proportion of the DS courses was covering analytical skills?" This is actually based on the business intelligence framework we borrowed from the Chen et al. publication[1].

They divided analytical skills into three levels. So, we wondered what percentage of the DS courses currently address levels of analytical skills and domain knowledge. And finally, we looked at the DS courses in terms of whether they're focusing on communication skills, information skills, visualization skills, and math and statistics.

Ms. Goldman: So, really blending, as you said, the different disciplines together.

Ms. Sae-Lim: That's correct, because we have learned from our literature review that a data scientist must not only have analytical skills, but they should be able to communicate their research [results] to people who are not data scientists.

Ms. Goldman: That's great. So, you sampled 30 data science programs throughout the United States, and how did you identify these programs? And what was your methodology for analyzing each program?

Ms. Sae-Lim: We discovered three sources for —

Ms. Tang: — Three websites —

Ms. Sae-Lim: — Three websites for our sample framework. But, those websites are not grouped by disciplines. So, we needed to go through each and group those. First, we needed to separate the schools, like whether or not they are in the US, because they aren't all sorted

[1] Chen, Hsinchun, Roger H. L. Chiang and Veda C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36(4): 1165-1188. http://www.misq.org/skin/frontend/default/misq/pdf/V36I4/SI_ChenIntroduction.pdf

alphabetically. America and Canada have mixed sources. Then we extracted those to be our sample framework. And then we grouped by discipline and then randomly selected four of each discipline.

Ms. Tang: We had eight disciplines, and there were two disciplines that didn't have more than three. So, we ended up having 30 in total. It's because of these two disciplines that we do not necessarily have four for.

Ms. Goldman: What were your roles in the research project?

Ms. Tang: Well, Wa did mainly the data collection, extraction processing, and we did coding together. The first task was looking at the program description in terms of the length, in terms of what are the common terms they use? What are the unique terms they use? And then we coded together for that. And then we looked at each course included on the website of these DS programs.

And we coded by the analytical skills and domain knowledge. Then we did a second round of coding, which is looking at whether they have communication skills, whether they cover communication skills, data visualization skills, information skills, and math and statistics.

Ms. Goldman: Your analysis really focused on those program descriptions, what the curriculum structure was, and what each course was focused on?

Ms. Tang: Right.

Ms. Goldman: What were your findings based on the different disciplines?

Ms. Tang: Well, there are a good set of quantitative results. So, in terms of the program description, we looked at the average length of each discipline's program description (Figure 1). So, the highest was independent data science centers. They have an average of 303 words in their program description. Then we had iSchools and math as second and third. The shortest average description is from the business schools, and arts and sciences.

Ms. Sae-Lim: [For] the unique words, we found that independent data science centers have the most unique words used, iSchools ranked number two, and then engineering (Figure 2).

Ms. Tang: So, it's sort of similar to the average length.

Ms. Sae-Lim: To the average length of program description. [For] frequently used words, we looked at similarities, and for those program descriptions among eight disciplines (Figure 3). And we found that, of course, data is the most frequently used term, and then business, program, analysis, and skills. And the least frequently used words were computer science.

Ms. Tang: We looked at, as Wa said, some of the common terms, either by the calculation of a cross discipline (Table 1). How many terms are covered by all eight disciplines and how many terms are covered by all 30 programs?

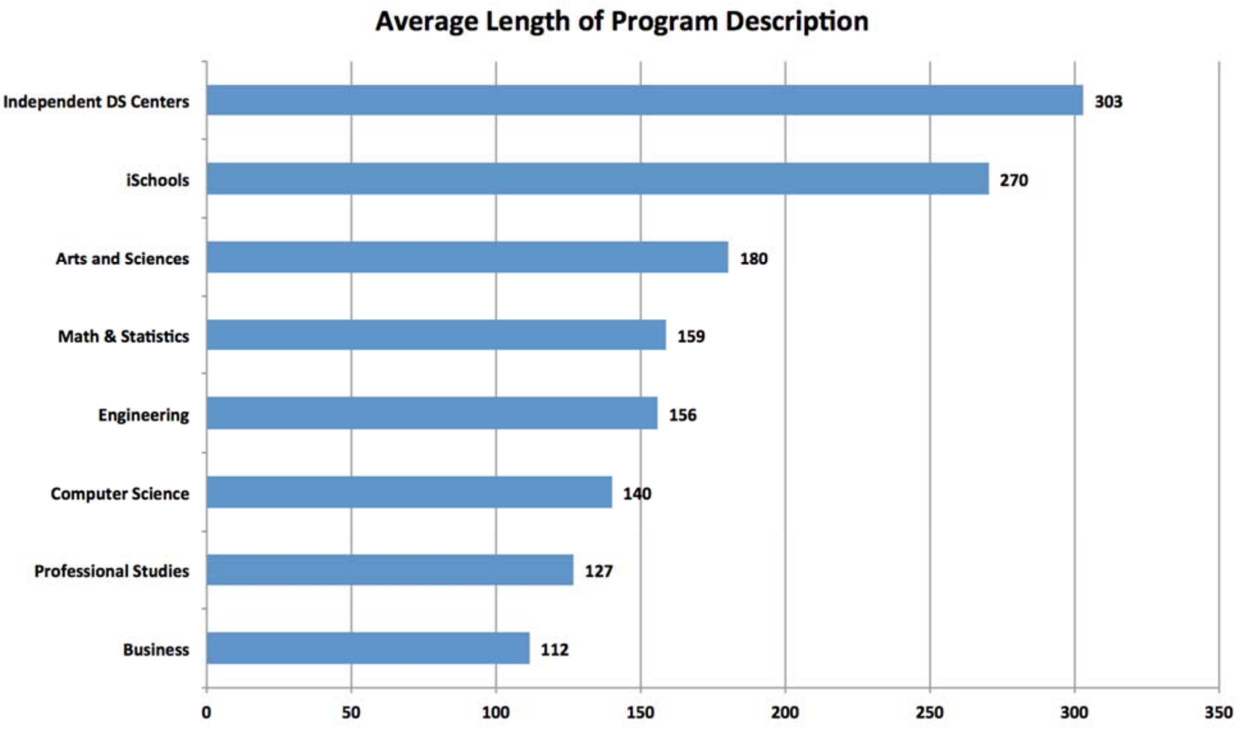

Figure 1: Average Length of Program Description by Discipline.

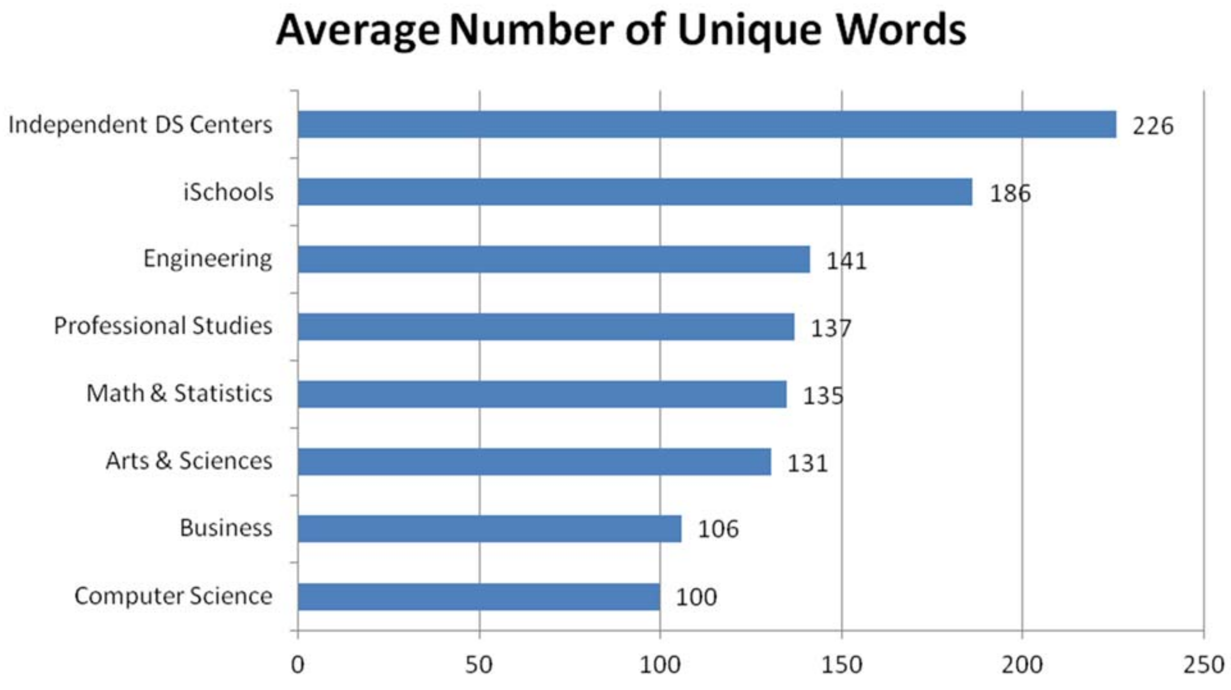

Figure 2: Average Number of Unique Words in Program Description by Discipline.

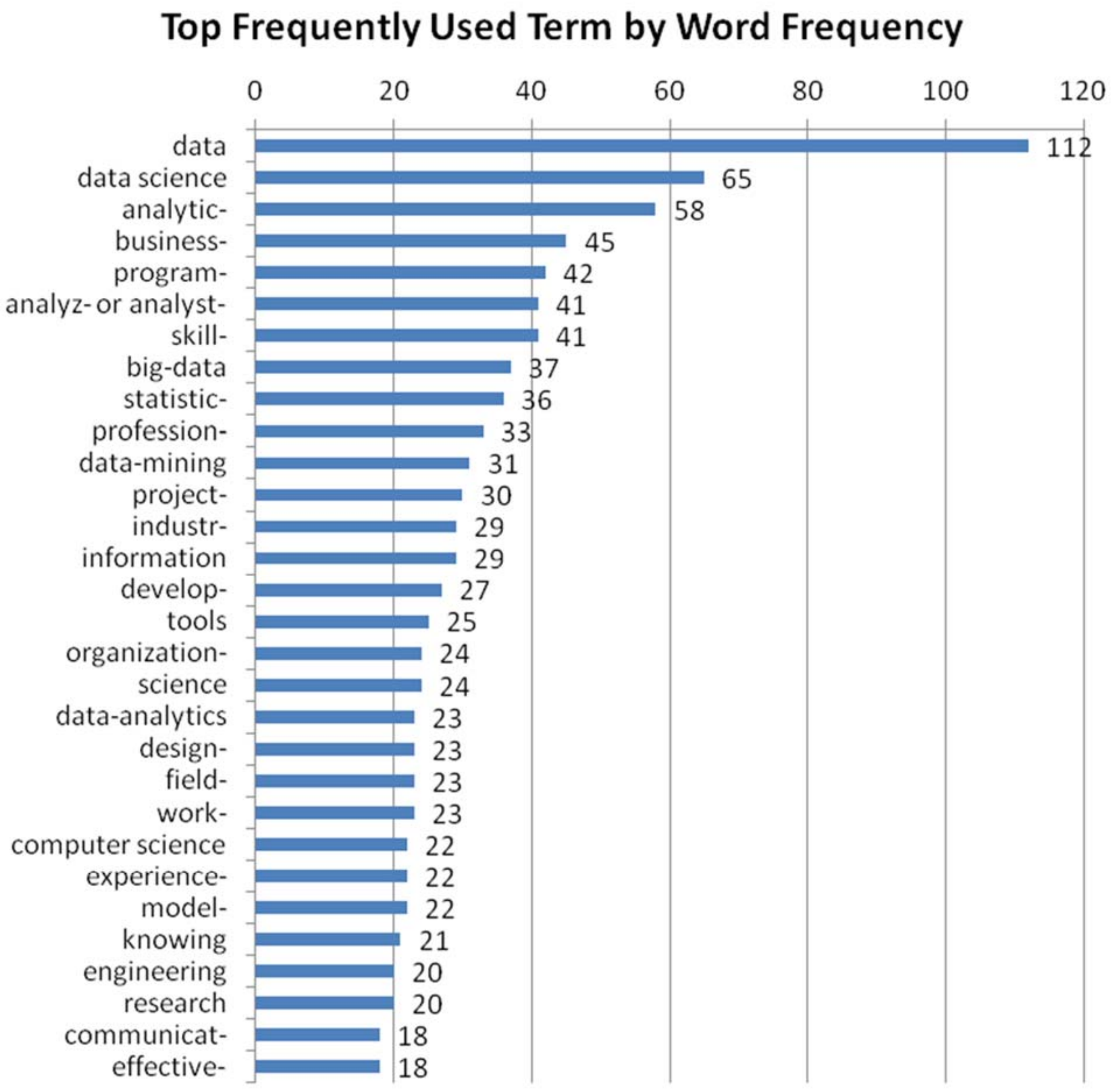

Figure 3: Top Most Frequently Used Terms in Program Description.

In terms of the highest, the most common term, as Wa said, is ‘data,’ which all eight disciplines had the program description using the term ‘data,’ and then 26 programs out of 30 used the term ‘data.’ But, then there are other terms, ‘business industry,’ ‘model,’ ‘profession,’ ‘tools,’ ‘fields,’ ‘techniques,’ and so forth. Those are the common terms.

Ms. Sae-Lim: But, the term ‘big data,’ interestingly, is the least frequent term to appear

Table 1: Common Terms Occurring Across at Least 12 Programs.

Common Terms	**Across Disciplines (n=8)** (Count/Percentage)	**Across Programs (n=30)** (Count/Percentage)
data	8 (100%)	26 (87%)
business-	8 (100%)	21 (70%)
industr-	8 (100%)	20 (67%)
skill-	8 (100%)	20 (67%)
analy- or analyst	8 (100%)	19 (63%)
analytic-	8 (100%)	16 (53%)
statistic-	8 (100%)	16 (53%)
model-	7 (88%)	15 (50%)
information	8 (100%)	14 (47%)
organization-	8 (100%)	14 (47%)
profession-	8 (100%)	14 (47%)
data science	7 (88%)	14 (47%)
know-	7 (88%)	14 (47%)
tools	7 (88%)	14 (47%)
computer science	7 (88%)	13 (43%)
develop-	7 (88%)	13 (43%)
field-	7 (88%)	13 (43%)
leader-	8 (100%)	12 (40%)
techniques	8 (100%)	12 (40%)
science	7 (88%)	12 (40%)
big-data	6 (75%)	12 (40%)

across eight disciplines.

Ms. Tang: [‘Big data’ appeared across six disciplines, and across 12 program descriptions.]

The next set of results is related to curriculum structure. We looked at how many courses were required, how many credit hours were required courses, and how many credit hours were devoted to electives (Table 2). And we also looked at whether they have a capstone or a practicum because one of the things that's frequently mentioned in their program description is the practiced knowledge, applied knowledge. So, for that, out of 30 programs, the average of credit hours required is 40 hours. But, what's interesting is there are a large proportion of required courses, and only a small portion of electives.

Table 2: Summary of Average Number of Credit Hours.

Discipline	Instruction Level		
	Total Credits	**Core**	**Electives**
Arts & Sciences	33.67	24.67	9
Business	35	28.25	6.75
Computer Science	48.75	42.25	6.5
Engineering	34	22	12
Independent DS Centers	32	17.67	13.33
iSchools	**71**	**53**	**18**
Mathematics/Statistics	33	26	7
Professional Studies	34	26	8

So, we're talking about the average of 40 required credit hours with 30 in core courses, core credit hours, and then only 10 in electives. They're focusing more on the required, just a little bit on electives.

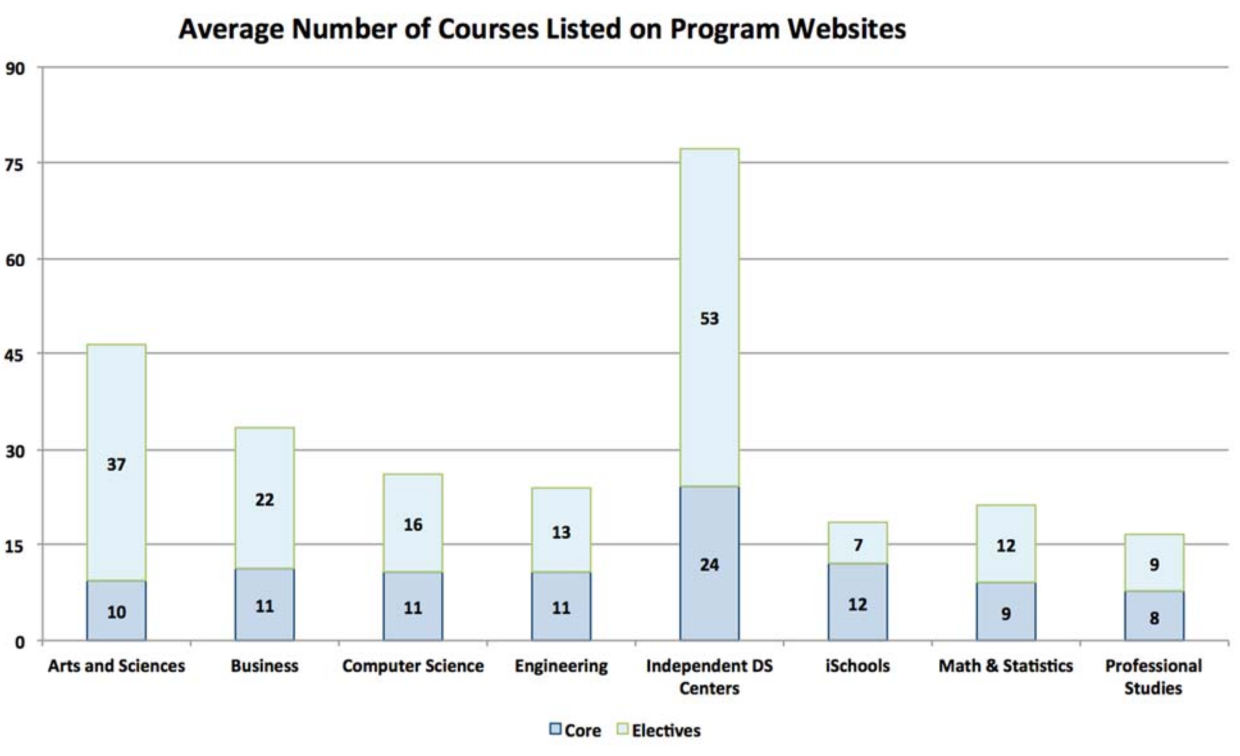

Figure 4: Average Number of Courses Listed on Program Website.

iSchools, interestingly, have the highest credit hours required for both core and elective. These have 71 total credit hours required, with 53 devoted for core, and then 18 devoted for electives. But, this is biased by Carnegie Mellon University, where they have 180 total credit hours required.

Then in terms of the number of courses listed in their website (Figure 4), independent data science centers has the highest, and then professional studies has the lowest in the core. They only have eight courses listed. iSchools has the lowest number of classes listed as electives with seven.

Then we looked at the capstone, as I mentioned before, capstone and practicum. What's interesting is [that] out of 30 programs, we found only 10 have a requirement of practicum or special projects. And only six require a capstone, even though we find the frequently occurring term ‘practice,’ applied knowledge in their program description. We are talking about 33 percent offering practicum or requiring practicum, and 20 percent requiring a capstone.

Seemingly, there's a gap there. For instance, the University of Virginia’s data science institute describes their capstone as a research project proposal. They're working with faculty to address a particular important data science challenge, and then the faculty from different disciplines help them to work together to develop the proposal. So, that's a capstone example from UVA.

Ms. Sae-Lim: We also looked at their core course focus (Table 3). So, for the core course focus, we coded the courses by looking at the domain knowledge and analytical skills. In analytical skills, we divided it into three levels, skill one, two, and three, and then domain knowledge, and looked at these eight disciplines. iSchools has the most domain knowledge, with 49.29 percent. For skill one, mathematics and statistics ranked number one with 77.78 percent. And iSchools is the only discipline that offers courses for skill three levels.

Table 3: Percentage of Analytical Skills and Domain Knowledge of Core Courses.

Disciplines	Skill 1.0	Skill 2.0	Skill 3.0	Domain Knowledge
Arts & Sciences	74.84%	6.30%	0.00%	18.88%
Business	50.08%	18.07%	0.00%	31.86%
Computer Science	55.35%	12.68%	0.00%	31.97%
Engineering	64.58%	*5.36%*	0.00%	30.06%
Independent DS Centers	69.77%	8.40%	0.00%	21.82%
iSchools	*26.74%*	**19.39%**	**4.58%**	**49.29%**
Mathematics/Statistics	**77.78%**	9.72%	0.00%	12.50%
Professional Studies	63.19%	15.28%	0.00%	21.53%

Ms. Tang: Just to add some explanation of the skill one, two, three: the differences between them is skill one is addressing more structured data. Skill two is unstructured. And skill three is mobile.
So, it's sort of ubiquitous data, mobile enabled. If you're looking at skill one, emphasis there will be just really much more structured data, and then upper level more unstructured.

Ms. Sae-Lim: Then for electives, we also coded the same as the core courses or full courses (Table 4). For skill one about the structured data, mathematics and statistics is ranked, again, number one with 64.42 percent. And engineering is ranked number one in skill two with 27.56 percent. For domain knowledge in electives, business schools are number one with 86.67 percent. And for the core courses, the focus is on the communication skills, mathematics and statistics, information skills, and visualization skills.

Table 4: Percentage of Analytical Skills and Domain Knowledge of Elective Courses.

Disciplines	Skill 1.0	Skill 2.0	Skill 3.0	Domain Knowledge
Arts & Sciences	49.06%	5.40%	5.40%	40.13%
Business	*11.11%*	*2.22%*	0.00%	**86.67%**
Computer Science	39.81%	17.14%	2.22%	40.82%
Engineering	48.56%	**27.56%**	0.00%	*23.89%*
Independent DS Centers	54.23%	15.07%	1.11%	29.59%
iSchools	31.21%	17.81%	0.00%	50.98%
Mathematics/Statistics	**64.42%**	4.17%	0.00%	31.41%
Professional Studies	30.48%	19.70%	0.00%	49.82%

Ms. Tang: What's interesting is in terms of core courses, information skills, the highest percentage coverage in the core curriculum is from professional studies, not from iSchools (Table 5). You would imagine that iSchools would emphasize a lot more in their core the information skills. But, they are embarrassingly listed as average at number four. So, we are talking about eight different disciplines, and number four is just right in the middle.

And in terms of the communication skills, all disciplines cover a very limited proportion. The highest is from computer science with 13.04 percent coverage. Math and statistics, no surprise there, because math and statistics tend to offer more on their own subject matter, which is 38, almost 39 percent. Visualization also didn't get a high coverage either. The highest would be from computer science with 12.50 percent. Next is from engineering, which is 10.9 percent.

So, it was sort of surprising that, one, information skills were not covered the highest by iSchools. Secondly, visualization seemed to be very lacking, and communication skills seems to be very lacking in core.

Table 5: Percentage of Communication, Math & Statistics, Information, and Visualization Skills in Core Courses.

Disciplines	Core			
	Communication Skills	**Mathematics & Statistics**	**Information Skills**	**Visualization Skills**
Arts & Sciences	3.17%	23.33%	17.78%	8.25%
Business	0.00%	20.08%	5.40%	3.41%
Computer Science	**13.04%**	8.33%	32.13%	**12.50%**
Engineering	0.00%	25.37%	12.18%	10.90%
Independent DS Centers	0.75%	38.60%	19.37%	0.50%
iSchools	3.13%	1.56%	20.73%	7.08%
Mathematics/Statistics	0.00%	**38.89%**	20.83%	0.00%
Professional Studies	0.00%	32.50%	**46.88%**	0.00%

Ms. Goldman: Do you think the visualization is because it's kind of a newer area?

Ms. Tang: Yes. We don't know why that's happening. That's covered more from computer science and engineering. Probably more from the programming point of view.

We also looked at communication, mathematics, information skills, and visualization [for electives] (Table 6). For electives, the communication skills [are also nearly covered]; no disciplines actually covered communication skills. The only discipline that covered communication skills is arts and sciences, with 0.57 percent. And then we have math and statistics again covering the highest proportion of math and statistics knowledge, with 46.70 percent.

Table 6: Percentage of Communication, Math & Statistics, Information, and Visualization Skills in Elective Courses.

Disciplines	Electives			
	Communication Skills	**Mathematics & Statistics**	**Information Skills**	**Visualization Skills**
Arts & Sciences	**0.57%**	26.52%	13.59%	2.95%
Business	0.00%	0.00%	6.91%	0.00%
Computer Science	0.00%	28.18%	10.12%	0.00%
Engineering	0.00%	28.00%	4.00%	**12.00%**
Independent DS Centers	0.00%	20.06%	0.61%	0.00%
iSchools	0.00%	10.29%	**27.29%**	0.00%
Mathematics/Statistics	0.00%	**46.70%**	16.85%	0.00%
Professional Studies	0.00%	10.37%	15.93%	10.37%

Information skills are covered in electives by iSchools, the highest proportion. So, we're talking about 27.29 percent. Visualization is again not being fully emphasized by any discipline very much, but engineering has the highest coverage, which is 12.00 percent. So, that's sort of interesting to us, specifically related to iSchools.
I think you asked a question about iSchools. Why are we interested in iSchools? Because we are in the LIS field. So, there are some gaps we found from studying this sort of curriculum requirement.

In the DS programming in iSchools, the core coverage on skill 1.0 is the lowest, [and this is] their core coverage. Their core coverage on math and statistics is the lowest. Their core coverage on information skills is not the highest because they're ranked at number four. And they have very low core coverage on communication skills and none in electives. They have very low core coverage on visualization skills and none in electives.

So, there are a lot of gaps that we identified for iSchools, of course, for other areas as well. But, in terms of the results, these are the key things.

Ms. Goldman: Your results show those low percentages associated with the data science programs at iSchools.

Ms. Tang: Yes.

Ms. Goldman: Addressing those programs specifically, how do you think these findings really influence library students, educators, and just professional practice in general?

Ms. Tang: In terms of how our finding actually influenced library students, I think for the future of library, LIS, in terms of curriculum development, data science is a very important skill set. It's sort of a newly emerged field, but as the field keeps evolving, the traditional LIS curriculum needs to be adjusted to enable our graduates to go out there, to be able to function in libraries that actually handle large amounts of research data.

As LIS students look for data science degree programs to enhance their education and to enhance their marketability, or seek a new career direction, they need to keep in mind in terms of current DS offerings, that really different disciplines have different directions.

So, they need to see which one actually matches with their background, which one actually can help them to expand their horizon. So, an LIS student can become a very good data scientist. It's not just exclusively for students who can do quantitative work.

Ms. Sae-Lim: I want to add to Rong's point that when we did the literature review from the industry's point of view and from the educator's point of view, they both point out that domain knowledge and analytical skills both are equally important. As you can see from our results, many disciplines focus on only analytical skills and leave domain knowledge mostly to electives, which could mean that some students may not or may choose those kinds of courses for their domain knowledge. So, this is the other thing we found out that students should keep in mind if they want to be a data scientist.

Ms. Tang: In terms of professional practice, we did find there are gaps in current higher levels of analytical skills, as we mentioned before. So students are only able to handle more structured data, [they are] not able to handle more large quantity, unstructured, new forms of data, and there is also a lack of training in visualization skills and communication skills.

So, I think for libraries that have positions open for data science librarians, they need to keep in mind there are some gaps in terms of [what students are] learning. So, when they actually go to work in the field, they need to learn more at work [about] these more realistic, new forms of data, how to handle them and enhance their knowledge.

Also for me, as a LIS faculty member, my observation is actually there's a trend in terms of developing data science programs. A lot of schools, iSchools, or just regular library schools, they want to develop this program as a directive from upstairs. So, people, let's say the President or the Dean believe there's an importance to developing such a program.

What they do is they change the current existing courses, change the title to make it more suitable for data science. I think what's lacking is more in-depth thinking about what is the intellectual core for data science? Is it the same as statistics? Is it the same as computer science? Can we just change the current existing computer science courses and give them new titles? [For example], it used to be, let's say, Introduction to Programming. Do you call it Introduction to Data Management or something like that? Can that work?

What I observe is there's a lack of in-depth thinking of what defines a data science program as an independent academic discipline. One of the writers reviewing our literature review said data science curriculum should have 20 percent devoted to theory development. So, what is the theory for data science?

And I think that's sort of lacking from what we find from the result of the data sets, but also from my experience in looking at how new data sciences programs, some of the data science programs, emerge without really some kind of logistical thinking of why do we need this and what's the difference between data science and statistics and computer science and engineering?

We actually tried to look at independent data science centers because they're independent. So, we're thinking, they must have a lot of new courses developed on their own to sort of establish their disciplinary boundary, even though we know that data science is multidisciplinary.

But, there must be something that says this is data science’s own course. But, we didn't find a lot of new courses. They also borrow from other disciplines. So, you still have computer science courses there. You still have business courses there, which is fine. But, we were hoping to find something that sort of links all these courses together. So, you have some kind of structure in this sort of data science program. That's yet to be seen. We haven't actually found something that's really unique.

Ms. Sae-Lim: What we found is in other disciplines, it wasn't that surprising. But, when we looked closely at independent data centers, [there] were some interesting [things].

Ms. Tang: Yes.

Ms. Goldman: You mentioned information professionals have that theoretical background and are in a prime position to help with the managing and curating and even visualizing data. And your findings suggest that data science in the librarian information schools needs to improve because there are gaps.

Ms. Tang: Right.

Ms. Goldman: What do you really envision for the future of library school curriculum and education surrounding data science programs?

Ms. Tang: Well, I mean library information science has always been evolving. It is a dynamic field because, let's say 10 years ago, what students learned from the [curriculum] in terms of cataloguing and classification or some kind of traditional LIS skills now have been evolved into metadata. But, we have new things emerging all the time. I believe that data science gives us new innovative skills that we need to incorporate into library and information science education.

We, as an information profession, need to be educating students who are forward-looking and who can actually handle whatever will emerge in the future. Right now we're actually looking at a completely new information environment than, let's say, 10 years ago.

Now we have information in different kinds of formats, not just textual information or printed. We have everything in electronic formats. We also have access anywhere, you can access information in different formats. So, we need to be able to provide curriculum that can adjust and can adopt new ways of handling and managing data. And this is absolutely necessary for the students who learn about data science.

It will also give them, as I mentioned before, new marketability and a new sort of profession. In the recent *LJ*, *Library Journal*'s salary survey, placement salary survey[2], their top fields earned positions in library school, the first one is software engineer.

But, the second one very close to software engineer is the UX [usability] researcher and the third one is the data scientist and data analyst. And there are currently not a lot. I believe out of the LJ survey, only five people responded for software positions, 35 for UX, but only five [for data scientists][3]. So, I think there's a need for data scientists in libraries or other information organizations. We just need to be able to provide high-quality education for the future data scientists who are working in the library or information profession.

Ms. Goldman: So, as you said, it's really looking at the curriculum as a whole, the program as a whole, not just renaming a course.

Ms. Tang: Yes. Exactly.

[2] Library Journal, Placements & Salaries 2015: Salary by Library Type. http://lj.libraryjournal.com/2015/10/placements-and-salaries/2015-survey/salary-by-library-type

[3] Correction: 10 “Software Engineer/Developer”; 37 “UX Designer/Researcher”; 7 “Data Analyst/Scientist”

Ms. Goldman: It's to fulfill a need and doing more overhaul to the curriculum.

Ms. Tang: Yes.

Ms. Goldman: Great! Well, thank you so much, Rong and Wa, for joining me today. Your research is of great interest to information education and the role librarians can have in many areas of research. Thank you.

Please read Rong and Wa's full research paper by following the link in the corresponding transcript article.

Disclosure

The original published research article can be found:

Tang, Rong and Watinee Sae-Lim. 2016. “Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus.” *Education for Information* 32(3): 269-290. http://dx.doi.org/10.3233/EFI-160977