Keeping Curriculum Relevant: Identifying Longitudinal Shifts in Computer Science Topics through Analysis of Q&A Communities

Habib Karbasian
Information Sciences and Technology
George Mason University
Fairfax, VA, USA
habib.karbasian@gmail.com
Aditya Johri
Information Sciences and Technology
George Mason University
Fairfax, VA, USA
johri@gmu.edu

Abstract—Keeping up with new knowledge being produced in computing related domains is a difficult task given the pace of change in the field. Specifically, in domains that are undergoing a lot of innovation, such as Data Science or Artificial Intelligence, updating curricula is not easy. Yet, there is a need to be cognizant of new topics in order to create and update curricula and keep it relevant. In this paper we present an innovative approach to help educators keep a better track of changes in a domain and be able to map their curricula objectives to emerging topics and technologies. We leverage Q&A sites, Reddit and StackExchange, which provide a useful online platform for sharing of information and thereby generate a valuable corpus of knowledge. We use Data Science as a case study for our work and through a longitudinal analysis of these sites we identify popular topics and how they have changed over time. We believe innovations such as these are essential for improving computer science education and for bridging the workplace-school divide in teaching of newer topics. Our unique and innovative approach can be applied to other CS topics as well.

Index Terms—online Q&A platforms, text mining, topic modeling, StackExchange, Reddit, curriculum development

I. Introduction

In computing, keeping up with advances in the field is difficult and changes in curriculum become an ongoing exercise. Although regular updates should not be made just to be topical and fashionable[1], a timely response to advances in the field is essential for preparing the future workforce [2]. The usual processes for curriculum development and update such as the use of committees is comprehensive but often slow [3] and there is a need and opportunity for instructors and programs to have a more dynamic view of changes in the field [4] to respond effectively [5]. In this paper we outline an innovative approach towards curricula development and updating that uses data mining and relies on drawing insights from online question and answering (Q&A) communities [6].Although Q&A sights are recognized as a useful resource of Computer Science (CS) knowledge, their application for curricula update

and verification has not been investigated so far. An inclusive approach to curricula development and enhancement is crucial to capture and expand the diversity of ideas in the field [7], and our work contributes to that effort.

II. Domain of Application - Data Science

Data Science is one of the most popular topics within CS at both the undergraduate and graduate levels and has emerged and been established a formal offering, through degrees in the topic, in recent years. As with any new domain, developing relevant curricula for the field has required significant efforts from the ACM Education Council, [8], the U.S. National Science Foundation (NSF) [9], the EDISON Data Science Project [10], which was launched in 2015 with the purpose of "accelerating the creation of the Data Science profession" [11], the Park City Math Institute, and the U.S. National Academies of Sciences, Engineering, and Medicine released an extensive report on Data Sciences for Undergraduates [12]. Yet, as the ACM task force has noted, a particular challenge of developing Data Science curricula is that keeping up with new topics particularly, with an eye towards balancing computing, statistics, and domain knowledge, and teaching important professional and ethical skills [13], [14]. Our work is motivated in part by recognizing that not only is there a need to keep current, it is necessary to do so swiftly given the advances in the field and it is important to learn not just from advances in research but also from practice i.e. from how the workforce is using data science. It is in this respect that Q&A communities provide a unique resource.

III. Prior Work: Q&A Communities

Q&A communities are used by thousands of professionals in the field as well as by newcomers to the topics creating a valuable resource of knowledge in the process. They provide easy access to experts [15] who can scaffold newcomers' learning [16]. They have high quality information [17], response rates are fast and they are up-to-date with new information[18], they contain useful examples of code [19], and they are largely publicly available. Overall, they have changed the landscape for information sharing and knowledge building [20] across a range of topics [21] and have been found to be useful for teaching [22] and for better understanding a topic [23]. Q&A communities are a useful resource for CS related topics as they are popular in that field and in addition to topics, they discuss tools, and techniques. For field such as Data Science, in addition to theoretical and mathematical issues, which form the core of the curricula, there is a need to address pragmatic

TABLE I: Submissions and comments for each subreddit

Ranking	Subreddit	Submissions+Comments (%)	
		Original	Preprocessed
1	DataIsBeautiful	2,975,912 (83.60)	1,781,973 (80.52)
2	MachineLearning	307,210 (8.63)	222,370 (10.05)
3	DataScience	154,149 (4.33)	116,328 (5.26)
4	LearnMachineLearning	40,642 (1.14)	31,088 (1.40)
5	Analytics	27,794 (0.78)	21,428 (0.97)
6	MLQuestions	20,946 (0.59)	17,432 (0.79)
7	BigData	18,435 (0.52)	11,843 (0.54)
8	DeepLearning	11,262 (0.32)	8,074 (0.36)
9	DataMining	3,352 (0.09)	2,669 (0.12)

issues such as tools [24] which requires following a functional approach [25] and being able to find and includes real world examples and scenarios [26], [27].

IV. Research Study

This study builds on prior work where data mining techniques were used to identify current topics of interest in data science [28]. In this study, we focus primarily on a longitudinal analysis of topics within Data Science and the guiding question for this research was: What are the increasing or decreasing trends in Stack Exchange (data science) and Reddit (data science subcommunities) over time? To answer this, we look at the trending topics as well as change in topics over time. We first provide an explanation of our datasets followed by description of the methods and analysis.

A. Datasets

1) Stack Exchange: Data Science: Stack Exchange is an online platform that hosts a variety of Q&A forums including one on data science. Stack Exchange makes its data publicly available in XML format under the Creative Commons license [29]. For this study we use posts.xml and comments.xml, which contain the actual text content of the posts and the comments, as well as the view count, favorite count, post type, creation date, and ID of the user who created each post and comment. The dataset spans for four years from January 2015 until January 2019.

2) Reddit: Data Science Subreddits: Reddit is a community-driven platform for submitting, commenting and rating links and text posts. Content entries, submissions, are organized by areas of interest or sub-communities called subreddits, such as politics, programming, science. We used the data dump provided here [30] under public licence which was collected originally from Reddit's official API [31] for submissions and comments. For the purpose of this work we decided to filter the entire dataset to these 9 data science related subreddits: 1-DataIsBeautiful, 2-MachineLearning, 3-DataScience, 4-LearnMachineLearning, 5-Analytics, 6-MLQuestions, 7-BigData, 8-DeepLearning and 9-DataMining. To put them in the same time line as StackExchange, we limited the data for the recent four years, January 2015 until January 2019, 48 months.

B. Methodology and Analysis

1) Preprocessing: Text was extracted from posts and comments by removing code snippets, HTML tags, URLs, hashtags, and common English-language stop words [32]. To increase the quality of text analysis 2-grams (equivalently, bi-grams) were used in the model [33]. Lemmatization was applied to identify intention in a part of speech and meaning of a word in a sentence. We used adjective, adverb, noun and verb as accepted parts of speech. To remove less frequent words from the sentences we set the minimum threshold as 10 for each word and 60% of the documents for maximum threshold. Finally, we removed sentences with less than 5 words to help with topic modeling (discussed later). After preprocessing, the

dataset contained 26,856 (52.3%) posts and 24,152 (47.7%) comments for Stack Exchange dataset and 137,060 (6.2%) submissions and 2,076,145 (93.8%) comments for Reddit dataset. The breakdown of the submissions and comments for subreddits is shown in table I under preprocessed column.

2) Topic Modeling: Following [34], we applied latent Dirichlet allocation (LDA) [35] to infer topics using the MALLET version 2.0.8 [36], an implementation of the Gibbs sampling algorithm [37]. The coherence score provides a rough estimate of the quality of the model and was used to decide the numbers of topics for each dataset [38], [39]. The result is (a) a set of topics, defined as distributions over the unique words in the dataset and (b) a set of topic membership vectors, one for each post, indicating the percentage of words in the post that came from each of the K topics. The highestprobable words in a topic are semantically related, which together reveal the nature, or concept, of the topic.

3) Metric: We first defined a threshold, _, to indicate whether a particular topics is "in" a document. Usually, a document will have between 1 and 5 dominant topics, each with memberships of 0.10 or higher [35]. In this study, we set _ to 0.10, which we found to remove noisy topic memberships while still allowing only the dominant topics to be present in each document. Then for each text, we normalized the weight of topics to be 1.

a) Weight Impact: We used the weight impact (WI) of a topic z_k in month m following the approach in [40]:

$$WI(z_k;m) = \sum_{d_i \in D(m)} \_(d_i; z_k) \quad (1)$$

where D(m) is the set of all posts in month m. The "weight impact" metric measures the texts for one given topic compared to the other topics in that particular month in terms of the topic weight. This approach helps to see which topic has been gaining or losing popularity over the course of four years.

b) Proportional Weight Impact: We used the proportional weight impact (PWI) of a topic z_k in month m inspired by [40] as the following:

$$PWI(z_k;m) = \frac{\sum_{d_i \in D(m)} \_(d_i; z_k)}{\sum_{j \in 1::k} \sum_{d_i \in D(m)} \_(d_i; z_j)} \quad (2)$$

where D(m) is the set of all posts in month m. The "weight impact" metric measures the relative proportion of posts/comments related to that topic compared to the other topics in that particular month in terms of the topic weight. This approach shows which topics have been being discussed

more or less frequently on a monthly basis in the last two years (2017-2018) compared to the first two years (2014-2015). Although “Temporal Weight Impact” measures the overall popularity of a topic in the course of action, “Temporal Proportional Weight Impact” allows us to find out which of those topics are becoming the center of the attention and discussion proportionately compared to other topics in each month. It also gives us a more detailed and more accurate understanding of which topics should be more focused in curriculum development.

c) Trend Difference: To find out which trend was increasing or decreasing in 2017-2018 compared to 2015-2016, we used the following equation to sort the trends for both “WI” and “PWI”:

$$\_Trend_{WI(z_k)} = 1 \; \frac{P_{m2 \in 2017 \text{–} 2018} WI(z_k;m)}{P_{m2 \in 2015 \text{–} 2016} WI(z_k;m)} \quad (3)$$

$$\_Trend_{PWI(z_k)} = 1 \; \frac{P_{m2 \in 2017 \text{–} 2018} PWI(z_k;m)}{P_{m2 \in 2015 \text{–} 2016} PWI(z_k;m)} \quad (4)$$

where it calculates WI and PWI for months from 2017 and 2018 over those for months from 2015 and 2016. It denotes the momentum of each topic per month for the last two years (2017-2018) against the two earlier years (2015-2016).

V. Results

We analyzed both datasets using a range of topics (from 2 to 100) and chose the highest coherence score as the basis for our optimal model for each dataset. The final model was trained for 1000 iterations. We uncovered 32 and 62 topics for Stack Exchange and Reddit respectively.

Topics on Reddit were of wider variety as compared to Stack Exchange and are probably an artifact of how the two platforms are moderated differently. The posting guidelines in Reddit are flexible but StackExchange enforces strict rules and off-topic postings are disallowed. The only non-data science topic among 32 topics in the StackExchange dataset was Q&A Guidelines which was discarded from our analysis. In Reddit, we removed several topics (such as “US election”, “entertainment industry”, “sports”, “climate change”) from our analysis as by analyzing few samples from each topic, we realized that those topics were not data science related and mostly personal opinion exchange with no data related substance. Therefore, we found only 19 topics relevant to data science out of 62 topics

Overall, the 31 data science related topics out of 32 in StackExchange (data science community) have 89.45% of total posts and comments. The 19 data science related topics out of 62 in the Reddit (9 data science subreddits) have 27.37% of total submissions and comments. The topics of each platform

were discussed and analyzed in more details in [removed due to blind review].

To calculate the trends for each topic if it is increasing or decreasing over time, we used the Cox Stuart trend test [41], to a statistically significant degree, using the standard 95% confidence level. Briefly, the Cox Stuart trend test compares the earlier data points against the later data points in a time series to determine whether its trend is increasing or decreasing, and uses the magnitudes of the differences to determine if the trend is significant.

Tables II and III provide the temporal trends for each topic in terms of "WI" and "PWI" metrics for both Stack Exchange and Reddit datasets respectively. On the left, trends based on "WI" are sorted in decreasing order and numbered from 1 and on the right are trends based on "PWI" metric using the same index number from "WI" table for the sake of comparison. Also the increasing trends for each datasets, Stack Exchange and Reddit, based on "WI" and "PWI" are visualized in figures 1, 2, 3 and 4 respectively. The trendlines are stacked on each other for more clarity and sorted from the least increasing to the most increasing order.

For Stack Exchange, 29 out of 31 topics have an increasing trend while the other two are constant (i.e., neither increasing nor decreasing to a significant degree) in terms of "WI". Deep Learning, Reinforcement Learning and Optimization are among the top increasing trends. But in terms of "PWI", there are 11 increasing trends, 14 constant trends and 6 decreasing trends. Deep Learning, Model Selection and Neural Networks are increasing among other topics. Although the few top increasing trends from "WI" stay popular in "PWI", some of them such as Visualization or Job/Education Advice either do not change or are discussed less compared to newly popular topics like Deep Learning.

On the other hand, for Reddit dataset, 14 out of 19 topics have an increasing trend and 5 are constant in terms of "WI". Job/Education Advice and Readings (Intro to DS) and Math Discussion in DS have the highest increase compared to other topics. But in terms of "PWI", there are 12 increasing trends, 6 constant trend and 1 decreasing trend. It is interesting that most of increasing trends in "WI" are also popular in terms of "PWI".

VI. Discussion

It is clear from the overall data we analyze that Data Science is a popular field that has gained a great deal of attention among technology workers and educators (as is evident from the temporal figures 1 and 3). Although online content generation and user participation have increased across the board due to easier access to the digital technologies, most of the topics have trended upwards. This can be seen in both figures 1 and 3 where "WI" metric captured the trend of content generation for each topic, 29 out of 31 topic in Stack Exchange, figure 1 , and 14 out of 19 topics in Reddit, figure 3 . There are some differences for sure. For instance, in the

Reddit dataset, in terms of both “WI” and PWI, Table III, Job / Education Advice is ranked first, although this topic is not as

TABLE II: Trends for 31 Topics in Stack Exchange Dataset Based on “Weight Impact” (left) and “Proportional Weight Impact” (right)

Id Topics (Sub-Topics) _Trend$_{WI}$(%)

1 Deep Learning (GAN-CNN) * 493:95%
2 Reinforcement Learning * 390:57%
3 Deep Learning (RNN-LSTM) * 381:57%
4 Optimization (Neural Network, SGD) * 346:02%
5 Code Debugging * 284:73%
6 Model Selection (Cross Validation) * 278:73%
7 Neural Network (Layer Structure, Activation Func) * 275:6%
8 Math Discussion in DS (Formula) * 249:65%
9 Model Selection (Performance Evaluation) * 225:96%
10 Preprocessing (Categorical Encoding, Missing Data) * 217:78%
11 Classification (Algorithm Selection) * 214:75%
12 Classification (Imbalanced- MultiClass) * 187:82%
13 Outlier in TimeSeries * 171:4%
14 Libraries Installation * 162:88%
15 Visualization (Plotting) * 142:54%
16 Preprocessing (Pandas, Data Manipulation) * 139:75%
17 Regression/Correlation * 130:27%
18 Feature Engineering (RF, DT) * 125:04%
19 Problem Formulation * 121:87%
20 Temporal Analysis (Prediction, TimeSeries) * 113:11%
21 Dimensionality Reduction * 104:19%
22 Statistical Tests * 91:69%
23 Social Network Modeling * 81:64%
24 NLP (BOW, Word2vec) * 80:38%
25 Readings (ML) * 73:72%
26 NLP (Text Extraction, Scraping) * 69:75%
27 Clustering * 68:08%
28 Generative Models (PGM-GAN-MLE) * 65:94%
29 Recommender System * 36:18%
30 Job/Education Advice -
31 Big-Data processing (Hadoop, Spark, NoSQL) -

Id Topics (Sub-Topics) _Trend$_{PWI}$(%)

3 Deep Learning (RNN-LSTM) * 276:26%
1 Deep Learning (GAN-CNN) * 177:41%
2 Reinforcement Learning * 100:28%
4 Optimization (Neural Network, SGD) * 92:56%
7 Neural Network (Layer Structure, Activation Func) * 77:55%
6 Model Selection (Cross Validation) * 66:44%
5 Code Debugging * 63:84%
10 Preprocessing (Categorical Encoding, Missing Data) * 42:43%
9 Model Selection (Performance Evaluation) * 41:79%
8 Math Discussion in DS (Formula) * 40:84%
11 Classification (Algorithm Selection) * 33:67%
13 Outlier in TimeSeries -
12 Classification (Imbalanced- MultiClass) -
14 Libraries Installation -
18 Feature Engineering (RF, DT) -
15 Visualization (Plotting) -
16 Preprocessing (Pandas, Data Manipulation) -
20 Temporal Analysis (Prediction, TimeSeries) -
17 Regression/Correlation -
21 Dimensionality Reduction -
22 Statistical Tests -
24 NLP (BOW, Word2vec) -
28 Generative Models (PGM-GAN-MLE) -
23 Social Network Modeling -
19 Problem Formulation + ⍰5:83%
27 Clustering + ⍰25:74%
25 Readings (ML) + ⍰28:83%
26 NLP (Text Extraction, Scraping) + ⍰32:83%
29 Recommender System + ⍰41:74%
30 Job/Education Advice + ⍰60:05%
31 Big-Data processing (Hadoop, Spark, NoSQL) + ⍰61:87%

TABLE III: Trends for 19 Topics in Reddit Dataset Based on “Weight Impact” (left) and “Proportional Weight Impact” (right)

Id Topics (Sub-Topics) _Trend$_{WI}$(%)

1 Job/Education Advice * 130:72%
2 Preprocessing (Pandas, Data Manipulation) * 126%
3 Logic in Game * 111:08%
4 Q/A in ML * 111:01%
5 Readings (Intro to DS) * 108:9%

6 Math Discussion in DS (Explanation) * 96:82%
7 Model Selection (Cross Validation) * 95%
8 Programming Languages * 83:81%
9 Readings (NN ,RL) * 65:29%
10 Visualization (Links) * 60:06%
11 Deep Learning (TensorFlow, Performance) * 56:95%
12 Visualization (Graph, Colors) * 56:37%
13 Readings (ML) * 37:77%
14 Statistical Analysis (Mean, Median, STD) * 27:24%
15 Statistical analysis (Correlation, Causation) -
16 Google Analytics -
17 Data (External Links) -
18 Deep Learning (CNN, GAN, LSTM) -
19 Readings (AI) -

Id Topics (Sub-Topics) _Trend$_{PWI}$(%)

1 Job/Education Advice * 80:1%
2 Preprocessing (Pandas, Data Manipulation) * 78:6%
4 Q/A in ML * 68:13%
5 Readings (Intro to DS) * 64:79%
6 Math Discussion in DS (Explanation) * 56:95%
7 Model Selection (Cross Validation) * 54:84%
18 Deep Learning (CNN, GAN, LSTM) * 48:34%
8 Programming Languages * 47:24%
9 Readings (NN ,RL) * 34:19%
11 Deep Learning (TensorFlow, Performance) * 28:26%
19 Readings (AI) * 26:82%
10 Visualization (Links) * 22:04%
12 Visualization (Graph, Colors) * 21:95%
3 Logic in Game -
13 Readings (ML) -
14 Statistical Analysis (Mean, Median, STD) -
15 Statistical analysis (Correlation, Causation) -
16 Google Analytics -
17 Data (External Links) + 20:49%

popular as other topics discussed before in Stack Exchange dataset (-60.05% decreasing trend in"PWI", see Table II). As Reddit is more suitable for newcomers and people from different background to familiarize with data science field, it shows that there are many opportunities available that have attracted broad range of users to be able to switch their existing career or education path to data science.

A. Optimization-Based Algorithms

As discussed in details about the topics of each platform in [removed due to blind review], Stack Exchange is more dedicated to technical community as opposed to Reddit has been freely used by the broad range of users, basic to advanced level. Therefore, to gain a more technical understanding of which trends are more popular than others, Stack Exchange is more informative. That being said, the top 5 increasing trends in that datasets are "Deep Learning (RNN-LSTM)", "Deep Learning (RNN-LSTM)", "Reinforcement Learning", "Optimization" and "Neural Network" are all topics that are primarily based on the foundation of the optimization topic. Although there are other topics such as Preprocessing and Visualization

Fig. 1: The Trendlines for Increasing Trends in Stack Exchange Dataset Based on "Weight Impact"

Fig. 2: The Trendlines for Increasing Trends in Stack Exchange Dataset Based on "Proportional Weight Impact"

in the rise, these optimization-based algorithms are gaining momentum and they should be given more attention and studied in more details in data science skill set development.

B. Classification vs. Clustering

In machine learning algorithms, two major parts being discussed is "supervised learning", classification, and "unsupervised

learning", clustering. In "classification", the deciding step to compare the models against each other in terms of evaluation metric is called Model Selection. The trend of this topic in both datasets in terms of "PWI" is upward, figures 2 and 4 respectively. Also the topic Classification (Algorithm Selection) in Stack Exchange is in rise in terms of "PWI" whereas Clustering is in decline, (-25.74% in "PWI", see Table II right section).

These positive trends for classification show that in research community and job market there are more problems dedicated toward classification rather than clustering suggesting the more focus on the former than latter in curriculum development.

VII. Conclusion

In this paper we present an innovative study that uses data from Q&A communities to identify Data Science topics that are relevant and maps them longitudinally to shed light on how those topics have trended over time. This information, we believe, is useful for improving curricula and ensuring

Fig. 3: The Trendlines for Increasing Trends in Reddit Dataset Based on "Weight Impact"

Fig. 4: The Trendlines for Increasing Trends in Reddit Dataset Based on "Proportional Weight Impact"

that topics that are taught are in congruence with the latest advances. This can assist in preparing students for the workforce and for also introducing them to useful topics and tools. Although we have applied our method to a single domain, this approach can easily be applied to other topics within CS. Our work is limited in scope as we rely on data from only two communities, the overall approach is still relevant and can be used with other datasets. In future work, we plan to map these topics with syllabi data collected from different programs to identify gaps and also to build a dashboard application that allows a quick comparison of topics and current syllabus. We also plan to pursue a more in-depth analysis of the various topics identified in our study to see if the topics can be segregated based on the level of course (undergraduate/graduate) it can be used in.

Acknowledgment

This work is supported in part by NSF Grants DUE-1712129 DUE-1707837. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] D. G. Kay, "Bandwagons considered harmful, or the past as prologue in curriculum change," SIGCSE Bull., vol. 28, no. 4, pp. 55–58, Dec. 1996. [Online]. Available: http://doi.acm.org/10.1145/242649.242666

[2] M. Milosz and E. Lukasik, "Reengineering of computer science curriculum according to technology changes and market needs," in 2015 IEEE Global Engineering Education Conference (EDUCON), March 2015, pp. 689–693.

[3] "Computer science curricula 2013: Curriculum guidelines for undergraduate degree programs in computer science," Jan 2013. [Online]. Available: http://dx.doi.org/10.1145/2534860

[4] G. Subrahmanyam, "A dynamic framework for software engineering education curriculum to reduce the gap between the software organizations and software educational institutions," in 2009 22nd Conference

on Software Engineering Education and Training, Feb 2009, pp. 248–254.
[5] J. Roberg´e and C. R. Carlson, “Broadening the computer science curriculum,” in Proceedings of the Twenty-eighth SIGCSE Technical Symposium on Computer Science Education, ser. SIGCSE ’97. New York, NY, USA: ACM, 1997, pp. 320–324. [Online]. Available: http://doi.acm.org/10.1145/268084.268206
[6] R. Ball, L. Duhadway, K. Feuz, J. Jensen, B. Rague, and D. Weidman, “Applying machine learning to improve curriculum design,” in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’19. New York, NY, USA: ACM, 2019, pp. 787–793. [Online]. Available: http://doi.acm.org/10.1145/3287324.3287430
[7] M. Sahami, A. Aiken, and J. Zelenski, “Expanding the frontiers of computer science: designing a curriculum to reflect a diverse field,” in Proceedings of the 41st ACM technical symposium on Computer science education. ACM, 2010, pp. 47–51.
[8] A. Danyluk, P. Leidig, L. Cassel, and C. Servin, “Acm task force on data science education: Draft report and opportunity for feedback,” in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’19. New York, NY, USA: ACM, 2019, pp. 496–497. [Online]. Available: http://doi.acm.org/10.1145/3287324.3287522
[9] B. Cassel and H. Topi, “Strengthening data science education through collaboration,” in Workshop on Data Science Education Workshop Report, vol. 7, 2015, p. 27.
[10] “The edison data science competence framework,” 09 2018. [Online]. Available: http://edison-project.eu/edison/edison-data-science-framework-edsf
[11] A. Manieri, S. Brewer, R. Riestra, Y. Demchenko, M. Hemmje, T. Wiktorski, T. Ferrari, and J. Frey, “Data science professional uncovered: How the edison project will contribute to a widely accepted profile for data scientists,” in 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom). IEEE, 2015, pp. 588–593.
[12] E. National Academies of Sciences, Medicine et al., Data science for undergraduates: Opportunities and options. National Academies Press, 2018.
[13] J. S. Saltz, N. I. Dewar, and R. Heckman, “Key concepts for a data science ethics curriculum,” in Proceedings of the 49th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’18. New York, NY, USA: ACM, 2018, pp. 952–957. [Online]. Available: http://doi.acm.org/10.1145/3159450.3159483
[14] C. B. Simmons and L. L. Simmons, “Gaps in the computer science curriculum: An exploratory study of industry professionals,” J. Comput. Sci. Coll., vol. 25, no. 5, pp. 60–65, May 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1747137.1747147
[15] X. Liu, G. A. Wang, A. Johri, M. Zhou, and W. Fan, “Harnessing global expertise: A comparative study of expertise profiling methods for online communities,” Information Systems Frontiers, vol. 16, no. 4, pp. 715–727, 2014.
[16] A. Johri and S. Yang, “Scaffolded help for learning: How experts collaboratively support newcomer participation in online communities,” in Proceedings of the 8th International Conference on Communities and Technologies. ACM, 2017, pp. 149–158.
[17] H. J. Teo and A. Johri, “Fast, functional, and fitting: expert response dynamics and response quality in an online newcomer help forum,” in Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014, pp. 332–341.
[18] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, “Design lessons from the fastest q&a site in the west,” in Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2011, pp. 2857–2866.
[19] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns, “What makes a good code example?: A study of programming q&a in stackoverflow,” in 2012 28th IEEE International Conference on Software Maintenance (ICSM).

IEEE, 2012, pp. 25–34.
[20] H. J. Teo, A. Johri, and V. Lohani, “Analytics and patterns of knowledge creation: Experts at work in an online engineering community,” Computers & Education, vol. 112, pp. 18–36, 2017.
[21] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov, “How social q&a sites are changing knowledge sharing in open source software communities,” in Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014, pp. 342–354.
[22] C. Hsing and V. Gennarelli, “Using github in the classroom predicts student learning outcomes and classroom experiences: Findings from a survey of students and teachers,” in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’19. New York, NY, USA: ACM, 2019, pp. 672–678. [Online]. Available: http://doi.acm.org/10.1145/3287324.3287460
[23] S. K. Moudgalya, K. M. Rich, A. Yadav, and M. J. Koehler, “Computer science educators stack exchange: Perceptions of equity and gender diversity in computer science,” in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’19. New York, NY, USA: ACM, 2019, pp. 1197–1203. [Online]. Available: http://doi.acm.org/10.1145/3287324.3287365
[24] A. C. Bart, D. Kafura, C. A. Shaffer, and E. Tilevich, “Reconciling the promise and pragmatics of enhancing computing pedagogy with data science,” in Proceedings of the 49th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’18. New York, NY, USA: ACM, 2018, pp. 1029–1034. [Online]. Available: http://doi.acm.org/10.1145/3159450.3159465
[25] S. Dahlby Albright, T. H. Klinge, and S. A. Rebelsky, “A functional approach to data science in cs1,” in Proceedings of the 49th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’18. New York, NY, USA: ACM, 2018, pp. 1035–1040. [Online]. Available: http://doi.acm.org/10.1145/3159450.3159550
[26] M. E. Hoffman, P. V. Anderson, and M. Gustafsson, “Workplace scenarios to integrate communication skills and content: A case study,” in Proceedings of the 45th ACM Technical Symposium on Computer Science Education, ser. SIGCSE ’14. New York, NY, USA: ACM, 2014, pp. 349–354. [Online]. Available: http://doi.acm.org/10.1145/2538862.2538916
[27] A. McGettrick, M. D. Theys, D. L. Soldan, and P. K. Srimani, “Computer engineering curriculum in the new millennium,” IEEE Transactions on Education, vol. 46, no. 4, pp. 456–462, Nov 2003.
[28] H. Karbasian and A. Johri, “Insights for curriculum development: Identifying emerging data science topics through analysis of qa communities,” in The 51st ACM Technical Symposium on Computer Science Education (SIGCSE ’20), 2020.
[29] (2019, April) Stackexchange data dump website. [Online]. Available: https://archive.org/download/stackexchange
[30] (2019) Reddit data dump website. [Online]. Available: http://files.pushshift.io/reddit/
[31] (2019) Reddit api. [Online]. Available: http://www.reddit.com/dev/api
[32] K. S. Jones, Readings in information retrieval. Morgan Kaufmann, 1997.
[33] C.-M. Tan, Y.-F. Wang, and C.-D. Lee, “The use of bigrams to enhance text categorization,” Information processing & management, vol. 38, no. 4, pp. 529–546, 2002.
[34] J. M. Rouly, H. Rangwala, and A. Johri, “What are we teaching?: Automated evaluation of cs curricula content using topic modeling,” in Proceedings of the Eleventh Annual International Conference on International Computing Education Research, ser. ICER ’15. ACM, 2015, pp. 189–197.
[35] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.
[36] A. K. McCallum, “Mallet: A machine learning for language toolkit,” 2002.
[37] S. Geman and D. Geman, “Stochastic relaxation, gibbs distributions,

and the bayesian restoration of images," in Readings in computer vision. Elsevier, 1987, pp. 564–584.
[38] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in Proceedings of the 26th annual international conference on machine learning. ACM, 2009, pp. 1105–1112.
[39] S. Grant and J. R. Cordy, "Estimating the optimal number of latent concepts in source code analysis," in 2010 10th IEEE Working Conference on Source Code Analysis and Manipulation. IEEE, 2010, pp. 65–74.
[40] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," Empirical Software Engineering, vol. 19, no. 3, pp. 619–654, 2014.
[41] D. R. Cox and A. Stuart, "Some quick sign tests for trend in location and dispersion," Biometrika, vol. 42, no. 1/2, pp. 80–95, 1955.