Exploring Technology Trends in Data Journalism

A Content Analysis of CAR Conference Session Descriptions

Erica Yee
Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA
yee.er@husky.neu.edu

ABSTRACT

The evolving field of data journalism has been the subject of some scholarship, but no systematic analysis of specific technologies used. This study builds on previous surveys and case studies of newsrooms by discovering trends in software and programming languages used by practitioners. Keyword discovery is performed to assess the popularity of such tools in the data journalism domain, under the assumption that topics discussed at the major conference in the field reflect technologies used by practitioners. The collected and categorized data enables time trend analyses and recommendations for practitioners and students based on the current technology landscape. While certain technologies can fall out of favor, results suggest the software Microsoft Excel and programming languages Python, SQL, R and JavaScript have been used consistently by data journalists over the last five years.

KEYWORDS

data journalism; computer-assisted reporting; data visualization; journalism; trends analysis; content analysis; journalism tools

I INTRODUCTION

Contemporary journalism is becoming increasingly quantitatively oriented as more newsrooms incorporate the gathering, analyzing, and presenting of data into their reporting [1]. As the tools for such work grow more sophisticated and numerous, journalists of all expertise levels strive to adapt with the pace of technology. While there has been prior research on the rise of data journalism, these case studies are mostly deep dives into a handful of newsrooms in particular countries [2,3,4]. Though useful in understanding the current state of the field, these limited-scope interviews cannot provide a higher-level overview. There is a need for further research exploring how data journalism practices have changed over time. This study aims to expand on prior scholarship by gleaning technologies and skills practiced by data journalists. By examining trends in tools discussed at a major conference in the field, the study provides an in-depth look at where data and digital journalism has been and is going in terms of specific programming languages, software and tools.

2 RELATED WORK

Understanding a technology landscape: Previous research has endeavored to understand the overall technology landscape for computer programmers. Multiple studies have mined data from the community question answering site Stack Overflow, with results showing that a mined technology landscape can provide an aggregated view of a wide range of relationships and trends [5], especially relative popularity of languages [6]. Other researchers have explored the popularity of programming languages based on data from GitHub, a popular social, distributed version control system. One study analyzing the state of GitHub spanning 2007-2014 and 16 million repositories found JavaScript to be the most popular language by a far margin, followed by Ruby and Python [7].

Information extraction methods: Information extraction (IE) is used to locate specific pieces of data from a corpus of natural-language texts [8], often for the purpose of improving other information search tasks. Dictionary-based methods can be useful for conducting visibility analyses, such as how often a topic is discussed in a series of documents over time [9]. These methods usually require manually-constructed dictionaries limited to a specific domain, the creation of which is labor-intensive and time-consuming. There is not much data for most domains and languages. Therefore, results from manual information extraction can be useful to help automate future text mining in larger and different data sources.

3 RESEARCH QUESTIONS

This goal of this project is to implement keyword discovery to discover trends of how the popularity of particular technologies have changed over time in the domain of data journalism. The hypothesis assumes that topics and technologies discussed at the major conference in the field reflect the state of where the industry is going and tools what practitioners are learning how to use around the time of each conference. The study aims to answer the following research questions:

- **RQ1:** What technologies, languages and libraries are used in the practice of data journalism?
- **RQ2:** How have these technologies of data journalism changed over time?

4 METHDOLOGY

4.1 Data Collection

4.1.1 Building a Corpus. A corpus is a collection of written or spoken text that was produced in a natural communicative setting

and is machine-readable [10]. The ideal corpus is a representative and balanced subset of a chosen language, created by sampling existing texts of a language.

The analyzed corpus comes from the annual Computer-Assisted Reporting (CAR) conference, which has run for 25 years. The CAR conference is organized by the nonprofit Investigative Reporters and Editors (IRE) and its program the National Institute for Computer Assisted Reporting (NICAR). NICAR has been and continues to be an important organization for those who practice data-driven journalism, providing resources and connecting like-minded professionals and students [11]. Though focused on data journalism, some conference topics also branch into other digital skills such as web development. Speakers are mainly industry professionals discussing current technologies used in their newsrooms. Thus, conference session descriptions seem to be a representative and useful source of current and upcoming technologies in the field. In this case, the corpus is not a sample but rather includes all available sessions.

4.1.2 Preprocessing. The conference session data were preprocessed by removing noise and normalization to limit the number of words to sift through. Noise removed were punctuation and stop words. The stop words filtered out came from the default list in the Python nltk library, which includes common words such as prepositions and pronouns. The text was also converted to lowercase letters and non-alphanumeric characters were removed before words were stemmed to combine inflectional forms.

4.2 Keyword extraction

After preprocessing text, word counts were retrieved for each unique word and only counted once per session description. The lists of over 2000 unique words for each conference were manually analyzed and cross-referenced to extract technology keywords. For simplicity, only unigrams were extracted.

Included		**Excluded**	
Type	*Example*	*Type*	*Example*
Programming languages	Python	Concepts	spreadsheet
Libraries	matploblib	Databases	Zillow
Frameworks	Django	File types	PDF
Software	Excel, Tableau	Operating system	Ubuntu
		Platform sites	Twitter

Table 1: Overview of included and excluded keyword types.

Much deliberation was given to how to count occurrences of keywords. The first naïve iteration included all occurrences of a given word no matter how many times it was repeated in a single description. However, it was decided that this method would inflate frequency counts because of the varied ways the descriptions were written. Thus, the method of counting each keyword only once per description was chosen. This method has drawbacks, mainly that a keyword central to the overall message of a description is given equal weight to one that may have just been mentioned once as a comparison or example. Still, this method was chosen because even if keywords that are relatively unimportant are counted in the results, these keywords are what data journalism practitioners are familiar with and discussing, so they are relevant.

The extracted keywords were then categorized, mostly based on programming language and type of tool. The limited scope of this study warranted solely manual categorization. The categories are inherently subjective because they were determined by one reviewer and because each keyword was exclusive to a single category. For example, the keyword "d3" was categorized as only "javascript" even though it could also fit under "dataviz". When the reviewer was unfamiliar with a technology keyword, context and information about the keyword was gathered using search engines. Each specified category includes at least three unique keywords. If a keyword had fewer related keywords, it was labeled as the "other" category.

5 RESULTS

Five years of conference talks were analyzed, from 2015-2019. Each conference consisted of more than 200 sessions. Some session descriptions mention multiple technology keywords, while some none. The registration and sales sessions were filtered out. The lists of unique words for each conference year excludes stop words and includes numbers and partial words that split from joining punctuation. Out of an average of 2622 unique words per conference, 117 keywords were extracted and labeled into 15 categories: Python, JavaScript, R, web, SQL, datamgmt (data management/ wrangling), geo (geospatial/GIS-related), dataviz (GUI data visualization tools), other, Ruby, stats (statistics-related), bot, docs (document management tools), graph, text.

5.1 Category results

The top categories by number of unique keywords were: Python, JavaScript, R, web, and SQL. The results can be visualized to effectively show changes in popularity using a combination of stacked area charts and ranking visualizations [12]. The stacked area chart show changes in mentions of major programming languages and related keywords, such as libraries.

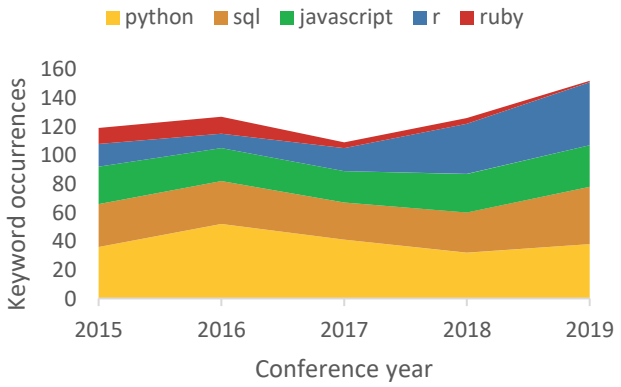

Figure 1: Mentions of major programming language keywords by category.

Out of the most frequently mentioned programming language categories, Python had the most keywords overall. SQL and JavaScript were consistent over the 5 years. The number of R keywords jumped in 2018 and stayed prominent in 2019. The relatively small number of Ruby keywords declined over the same time period.

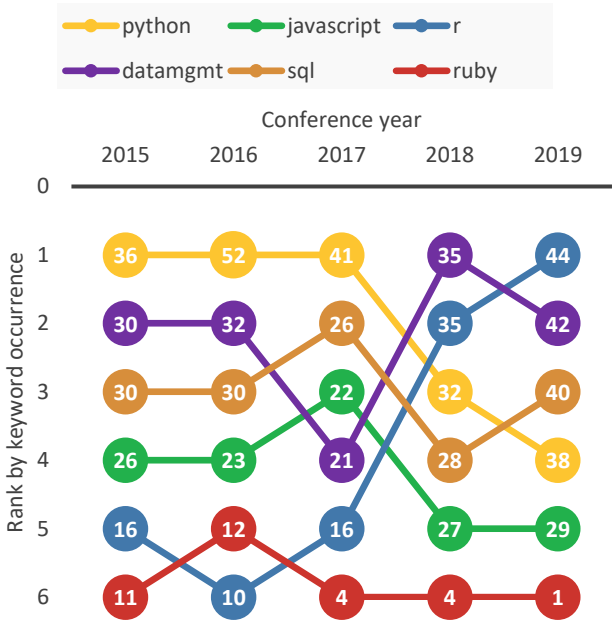

Figure 2: Ranking of categories per keyword occurrences.

The same categories are shown in the ranking chart as the stacked area chart, with the addition of the "datamgmt" category, which includes data management and wrangling tools like Microsoft Excel. The data labels are the occurrence of each keyword. The R category's jump in 2019 can be clearly seen.

5.2 Keyword results

Figure 3 shows the rankings of specific keywords, some of which are also category names. While the top 5 keywords moved around within that range, the sixth ranked keyword changed every year.

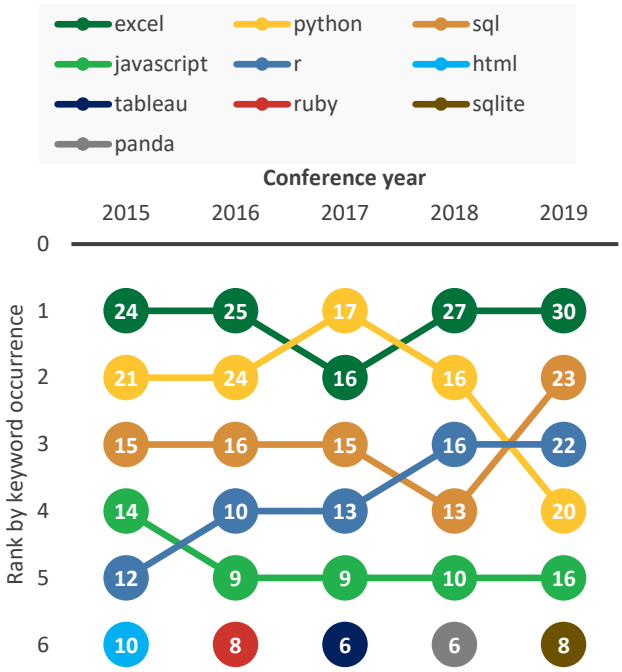

Figure 3: Top keywords by occurrence.

Besides ranking categories and keywords relative to each other, it also may useful to compare keywords within categories.

5.2.1 Python. The keyword "python", referring to the programming language, and its category were consistently ranked with some of the most mentions for every year data was analyzed. Out of the Python-related keywords, the data analysis library "pandas" was mentioned in the most descriptions every year except 2016, when it tied with "jupyter" and "django" for the most occurrences. The keyword "jupyter", referring to the interactive computational environment Jupyter Notebook rose in number of occurrences after not appearing in 2015. This increase appears in conjunction with the decrease of mentions of "ipython", the former name of Jupyter Notebooks. The web framework "django" and data analysis library "agate" also dropped in popularity over the 5 years.

5.2.2 JavaScript. Out of the JavaScript-related keywords, the data visualization library "d3" was mentioned in the most descriptions every year except 2016, when the runtime environment "node" occurred more times. The number of "d3" mentions increased from 3 in 2015 and 2016 to 6 in 2017 and 2018, then 7 in 2019. JavaScript tools "npm", a package manager, and "grunt", a task automator, as well as the popular "jquery" library often used for web page traversal and manipulation were consistently mentioned in 2015-2018, then not at all in 2019. The "react" library, used to build user interfaces, did not appear before 2018 and then appeared twice in 2019.

5.2.3 R. While multiple Python- and JavaScript-related keywords appeared intermittently over the course of the five years of conferences, the trend for R-related keywords looks markedly different. No r-related keyword beside the name of the language appeared in more than two descriptions from 2015-2017. Additionally, there were no more than four unique R-related keywords for those three conferences. The number of R-related keywords, both unique keywords and occurrences of each, then jumped for 2018 and 2019. Two notable related keywords were "tidyverse", a collection of R packages for data science, and data visualization package "ggplot2," both of which appeared in at least 3 descriptions in the two most recent conferences.

5.2.4 Data wrangling and management tools. The keywords under the "datamgmt" category included a mix of commonly used tools as well as newer technologies that were just getting introduced. The keyword "excel", referring to Microsoft Excel was by far the most common technology keyword in this category and overall. The keyword "fusion", referring to the Fusion Tables web service for data management that Google is retiring in December 2019, appeared 3 times in 2015 but decreased to 0 in 2019. The keyword "openrefine", referring to the desktop application OpenRefine used for data wrangling, was mentioned at least once every year. Finally, "pivot", referring to summary pivot tables for data analysis in Google Sheets, Microsoft Excel, and similar software was mentioned at least 2 times every year up to a peak of 7 times in 2019.

6 DISCUSSION

Though only limited and preliminary conclusions can be drawn from analyzing these five years of data, a few themes and practical takeaways emerge. The 15 categories of keywords show the breadth technologies discussed at the NICAR conference and that may be used by practitioners of data and other forms of digital journalism. Five of the categories were programming languages used in the field: Python, JavaScript, R, SQL, and Ruby. Each of these languages were themselves keywords as well as categories that included keywords such as libraries that extend capabilities and software to develop with the languages. The diversity of not only languages discussed, but also libraries and tools for each language indicate that there are many paths for similar end goals in data journalism.

The findings also indicate that certain technologies may fall out of favor and use over time. This has been shown in previous research on development trends. According to a 2019 IEEE conference paper analyzing GitHub repositories from 2009-2017, "JavaScript" has been a key front-end web development language over the last decade [12]. This research found the popularity of the "jQuery" library has shrunk since 2011. Around the same time, the "react" library that provides techniques delivered by "jQuery" has grown explosively. Interests in "jQuery" and "html" have declined as "react" has grown in popularity since 2012. The analysis of technologies mentioned at the CAR conference indicate journalism web development trends may follow these overall industry trends.

Still, some technologies seem to remain the prevailing standard, namely Excel. Microsoft's spreadsheet software appeared in the most descriptions every conference except in 2017, when Python had one additional occurrence. Excel's consistent top ranking may be explained by the software's relative longevity and accessibility to journalists who may have never programmed before but want to store and analyze data. Mentions of pivot tables, a key feature in Excel for summarizing data, have increased in the most recent conferences.

While the number of Python-, SQL-, and JavaScript-related keywords stayed relatively consistent over the years, the number R-related keywords increased in 2018 and 2019, surpassing the other two languages in keyword occurrences. During these two conferences, mentions of the data science package collection "tidyverse" especially increase, possibly speaking to its popularity among data journalism practitioners.

The fewer mentions of JavaScript, a programming language often used for web development, compared to Python, R, and SQL make be due to the conference's focus on tools for data analysis. Yet JavaScript's appearance among the top five keywords every year may point to a desire of journalists to learn skills for developing user-facing web pages and applications.

Finally, it is worthwhile to look at the most recent top words to assess the current landscape based on 2019 conference data. Keywords were counted as top here if they occurred in at least three unique descriptions in 2019. Keywords that are the name of their category (most significantly, names of programming languages) were not included here. These findings are suggestions about which tools and libraries are popular in the current and upcoming landscape, which could be useful for students and professionals to exploring technologies for journalism.

Category	***Key technologies***
Python	Jupyter Notebook, matplotlib, pandas
JavaScript	D3
R	ggplot2, RStudio, tidyverse
web	CSS, HTML
SQL	MySQL, SQLite
datamgmt	Excel, pivot tables
geo	QGIS
other	Github
dataviz	Tableau
stats	PSPP, SPSS

Table 2: Top keywords in major categories for 2019.

7 LIMITATIONS

Data source: The conclusions drawn from this analysis are based on the hypothesis that technologies discussed at the major conference in the field of data journalism reflect the state of the industry and technologies used by practitioners at the time of the conference. There are some possible drawbacks to solely analyzing this data source. First, any extrapolation of findings from conference data to the wider field assumes that the session topics and attendees are representative of practitioners in the industry Conference talks can also be opportunities for

introducing new and novel tools and platforms. As such, these technologies may not be widely used yet, if they ever will be. Thus, just because a technology is mentioned at a conference does not mean it is prevalent in the industry. Newer technologies may become more standard eventually, but this analysis may speak more to upcoming technologies than current ones. Additionally, talks at conferences are sometimes sponsored by companies or organizations who want to increase awareness and familiarity of their products by potential users and customers, but just because a company hosts a session does not mean the featured product is widely used by practitioners outside the conference.

Keyword extraction methodology: The manual keyword recognition and extraction performed in this analysis is limited by the single reviewer's knowledge, familiarity, and human error. In standard natural language annotation processes, a corpus should be annotated by at least people before it can be used for machine learning [10]. Though this study did not involve annotation, it follows that keyword extraction by at least two reviewers would lend credence to the analysis results. Extraction by a sole reviewer increases the likelihood that newer and/or more obscure technology names might not be recognized, so it was necessary to reference possible keywords of interest within their original context in the session description

Potential technology keywords that are common English words also caused some confusion. An especially challenging case was the word "access", which could refer to the Microsoft database management software but also showed up frequently in the data. Though Microsoft Access is a known tool used for data journalism, this potential keyword was not counted in the final analysis because there were too many occurrences to reasonably separate by a single reviewer under time and resource constraints. A possible improvement to this method that could catch some of these challenging cases would be to search for keywords that are bigrams and/or trigrams. Some technology names observed during manual analysis that consisted of more than one word were: "command line", "google sheets" and "amazon mechanical turk". It could be worthwhile to search for such keywords in a future analysis but searching for bigrams and trigrams also adds more complications. Technology names that include a stop word would be split apart in the current preprocessing algorithm. Additionally, full names of technologies are not always used colloquially, such as in the case of Microsoft Excel. Related problems are name references inconsistencies in the source material, which affects keyword counts and thus the data analysis. For example, the package R Markdown was sometimes referred to as "r markdown" and other times "rmarkdown".

8 FUTURE WORK

Other corpora: Another valuable record of how thinking and practices around data journalism have changed is the NICAR-L email listserv. NICAR-L serves primarily as a forum for the discussion of subjects related to getting and analyzing electronic information. As a text genre, an email listserv is useful for determining a technology landscape because its content is what students and professionals in the field are discussing. The NICAR-L archive is available online in a database hosted by the University of Missouri. Archives for the listserv go back to 2002, so it seems possible to glean trends that data journalism professionals have been discussing over time, though the scope of a future study may not allow a comprehensive analysis of all available data. To create a corpus for analysis, emails from a certain time range could be queried for their metadata (date) and text. The listserv was the original corpus intended for this study, but technical difficulties with the archive platform prevented access in time to conduct an analysis. Email text tends to be challenging to analyze, however, due to threading and lack of standardized formatting. Thus, analyzing conference descriptions was probably a more expedient for this initial study.

Future analyses could also examine job postings over time in the fields of data and other types of digital journalism to answer similar research questions about what skills are in demand and how they have changed. The methodology would be similar to this analysis of conference descriptions because the corpus would also be created from unstructured text with keywords of interest.

ACKNOWLEDGMENTS

This research was completed for IS 4900 Information Science Senior Project with Professor Martin Schedlbauer.

REFERENCES

[1] Coddington, M. (2015). Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital Journalism*, 331-348.

[2] Appelgren, E., & Nygren, G. (2014). Data Journalism in Sweden: Introducing new methods and genres of journalism into "old" organizations. *Digital Journalism*, 394-405.

[3] Hermida, A., & Young, M. L. (2017). Finding the Data Unicorn: A hierarchy of hybridity in data and computational journalism. *Digital Journalism*, 159-176.

[4] Wright, S., & Doyle, K. (2019). The Evolution of Data Journalism: A Case Study of Australia. *Journalism Studies*, 1811-1827.

[5] Chen, C., & Xing, Z. (2016). Mining technology landscape from stack overflow. *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (p. 14). ACM.

[6] Singh, A. K., Sena, D., Nagwani, N. K., & Pandey, S. (2016). Folksonomy based trend analysis on community question answering sites: A perspective on software technologies. *IEEE Access 4*, 5223-5233.

[7] Sanatinia, A., & Noubir, G. (2016). On GitHub's Programming Languages. *arXiv preprint*.

[8] Nahm, U. Y., & Mooney, R. J. (2002). Text mining with information extraction. *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, (pp. 60-67). Stanford CA.

[9] Boumans, J. W., & Trilling, D. (2015). Taking Stock of the Toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, 8-23.

[10] Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* O'Reilly Media, Inc.

[11] Fink, K., & Anderson, W. (2015). Data Journalism in the United States. *Journalism Studies*, 467-481.

[12] Park, C., Do, S. L., Jang, H., Jung, S., Han, H., & Lee, K. (2019). GitViz: An Interactive Visualization System for Analyzing Development Trends in the Open-Source Software Community. *2019 IEEE Pacific Visualization Symposium (PacificVis)* (pp. 179-183). IEEE.