2019

Promoting Institutional Repositories via Visualizations - A Changepoint Study

Zhehan Jiang
University of Alabama - Tuscaloosa, jiangzhehan@gmail.com

Sarah Fitzgerald
University of Massachusetts Amherst, sfitzgerald@umass.edu

Recommended Citation

Jiang, Zhehan and Fitzgerald, Sarah, "Promoting Institutional Repositories via Visualizations - A Changepoint Study" (2019). *New Review of Academic Librarianship*. 100.
https://doi.org/10.1080/13614533.2018.1547775

Promoting Institutional Repositories via Visualizations: A Changepoint Study

Abstract

This paper examines whether implementing visualizations on an institutional repository webpage increases traffic on the site. Two methods for creating visualizations to attract faculty and student interest were employed. The first is a map displaying usage of institutional repository content from around the world. This map uses Tableau software to display Google Analytics data. The second method is a text mining tool allowing users to generate wordclouds from dissertation and thesis abstracts according to discipline and year of publication. The wordcloud uses R programming language, the Shiny software package, and a text mining package called tm. Change in the number of institutional repository website sessions was analyzed through changepoint analysis.

Introduction

Institutional repositories (IRs) make scholarship accessible in ways that surpass toll barricaded outlets. However, the value of this access depends on amassing quantity and quality content from faculty and students at the institution in question. Creating an institutional repository website that inspires interest from faculty and students is therefore vital. Since institutional repository content is expected to be accessed primarily through search engines such as Google Scholar, a repository's homepage is called upon to be more of a demonstration of the repository's value to faculty and students than a discovery tool. We set out to provide several visualizations for the University of Alabama's institutional repository page which would show its worth to faculty and staff content contributors. Open access has two primary benefits (Tennant, et al, 2016). First, it increases the impact of scholarly articles and second, it allows researchers to mine scholarly literature. Our goal was to highlight both benefits through our IR visualizations.

We began our IR visualization project by creating a map showing the use of the content from the University of Alabama's institutional repository across the world. This demonstrates to prospective contributors that work they deposit will be available to scholars worldwide. This will increase the potential impact of their research. It serves to advertise the capacity of the institutional repository to track downloads of scholarly work. The University of Alabama map shares similarities to one produced for Georgia State University's institutional repository, ScholarWorks (Georgia State University Library, 2017), which shows the locations of real time downloads from the repository. Our map also shares similarities with the readership map created by Zhang and Lopez (2017) for Oregon State University's institutional repository, ScholarsArchive@OSU. Unlike these two maps, ours visualizes cumulative use from around the world rather than the locations of downloads occurring at the time of viewing.

Next, we created a visual tool for exploring trends in repository content across time, disciplines, and scholarly attainment levels. We were inspired by Sterman and Borda's (2017) creation of a visual discovery interface for Montana State University's institutional repository. Sterman and Borda's visualization tool displays the distribution of dissertations and theses across colleges and departments and the change in productivity for the departments over time. This allows users to compare outputs between departments. Our wordcloud visualization allows users to see trending topics in a discipline, shifts in focus within a discipline across years, or differences in focus from the master's level to the doctoral level. We aimed to create a tool to allow users without experience in text mining to look for and discover trends in the research outputs available in the University of Alabama's repository. In contrast to Sterman and Borda, our goal was to illuminate trends within disciplines rather than between departments. We felt this would have more value to scholars, as they tend to identify more with their disciplines than with their institutions (Austin, 1991).

We measured the impact of these visualizations by running a changepoint analysis on the number of Institutional Repository website sessions over time.

Literature Review

Participation in institutional repositories has been slow growing (Haijem, Harnad, & Gringas, 2006). In fact, Gargouri, et al. (2012) found that only 24% of articles published between 2005 and 2010 were openly accessible. Kim (2010) found in a study of 17 doctoral granting institutions with institutional repositories that 66.7% of faculty who self-archived their publications used their personal webpages, 51.5% used research group websites, 41.7% used departmental websites, 28.7% used disciplinary repositories, and only 22.7% used institutional repositories. This is not desirable, as personal and departmental websites are not as good at providing persistent access to articles as institutional and subject repositories built to preserve access to scholarship. Another advantage that repositories offer over personal and departmental websites is that many monitor submissions for compliance with copyright agreements. Therefore, it is important for libraries to encourage participation in institutional repositories. To encourage such participation, librarians must draw attention to their institutional repositories and convince scholars of their value.

It is understandable that many scholars are unaware of the existence and benefits of institutional repositories. When scholars search for content, they are not usually interested in limiting their search to content from a single institution. This makes searching an institutional repository via its website instead of a search engine an infrequent practice. In fact, Wolff, Schonfeld, and Rob (2016) found that databases are the most common starting point for scholars (reported by 42% of scholars). Library webpages account for 11% of initial searching, Google Scholar accounts for 21% and Google accounts for 20%. At the University of Alabama, less than 1% of downloads from the institutional repository in its first six months were from UA's institutional repository website rather than an external site. Most scholars visit Google Scholar or their library's search page as their first stop in seeking information, not their institutional repository. The value of institutional repositories is largely hidden to scholars because they access the content from other sites. However, when scholars are looking for data about their own institution's specialties or the trending topics of a certain discipline, an institutional repository website can be a valuable tool. This makes the institutional repository more visible to scholars.

Depositing one's scholarship into an institutional repository can be advantageous for academics. Hajjem, Harnad, and Gingras (2006) showed that open access articles are cited more often than articles only available through toll access outlets (with varying levels of advantage by discipline). Open access increases dissemination of articles in popular venues as well as scholarly ones. Wang, Liu, Mao, and Fang (2015) showed that open access articles get more downloads and social media mentions than toll access only articles. This use reflects work which is highlighted by news media in circles beyond academe, in addition to its academic use as evidenced by citations. The map visualization we created highlights this advantage in dissemination by displaying the long reach of materials deposited into the institutional repository. Submission to an institutional repository can increase a scholar's prestige, by widely disseminating their work. It aligns with their professional value to educate. Our aim was to target these needs and values by making the benefits of institutional repository submission visible to our scholars.

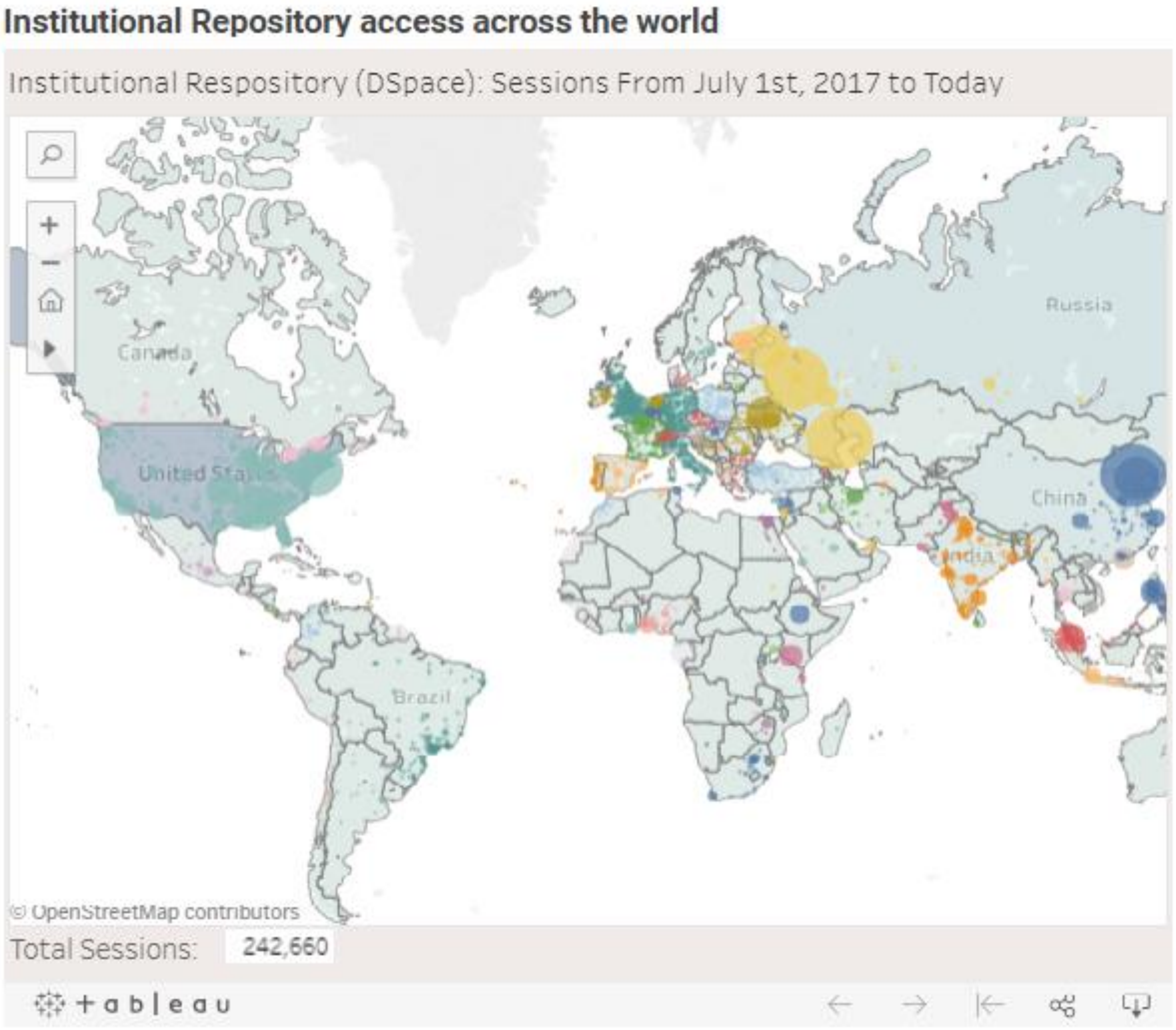
Visualizations help people understand data and patterns in ways not facilitated by text and tables alone. Visualizations have been used for institutional repositories in different ways. Several universities have employed maps to display downloads from their institutional repositories (Georgia State University Library, 2017; Zhang and Lopez, 2017). We employed this technique to highlight the advantages of dissemination granted by institutional repository use, but we also wanted to highlight the text mining possibilities offered by uploading work to an institutional repository. Polley (2016) created an interesting visualization of term co-occurrence in abstracts from the University of Indiana Purdue University Indianapolis's institutional repository, however, this visualization was intended to be of use mostly to librarians because it required knowledge of the repository's metadata structure to interpret. Our visualizations are intended to be of interest to patrons rather than (or in addition to) librarians. Sterman and Borda (2017) created a visual search tool in the shape of a sunburst graph to display the quantity of dissertations and thesis in departments at Montana State University over time. We appreciated the simplicity of their interface and sought to emulate it. Rather than focusing on the structure of the university colleges and departments in the way Sterman and Borda did, we chose to focus on the language used by authors to display content in the dissertations and theses in the repository. This visualization tool highlights the capacity for data mining research made possible by providing access to research through an institutional repository.

Institutional context

The institutional repository at the University of Alabama is a new endeavor started in June 2017. By May 2018, it contained 2,817 documents. These are primarily dissertations and theses, but the university hopes to expand the types of materials in the collection to include more postprints of articles published in peer reviewed journals in the future. The dissertations and theses in the repository date back to 2009, when the university first implemented electronic dissertation and thesis submission. The power of the wordcloud tool will be enhanced once the Libraries digitize dissertations completed before the instatement of electronic submission. The repository is powered by DSpace, an open source repository software package. Because the repository is so new, the University Libraries are actively working to increase interest in this service.

Map Visualization

The map of institutional repository usage displays differences in use between countries using a gradation of blues. The United States is the darkest country in the figure because usage has been dominated by users from the United States. Usage differences between cities are signified by the size of the dots over them. Figure one is the default view of the map seen by visitors to the repository website when they first arrive at the page. Hovering the sensor on a certain country/city will show the corresponding location label and the number of sessions from that location. In addition to the domestic display, sessions in the global scale indicate distribution across the world. Moscow and Saint Petersburg in Russia, Kiev in Ukraine, Mantsala in Finland, and Beijing in China had high volumes of hits.

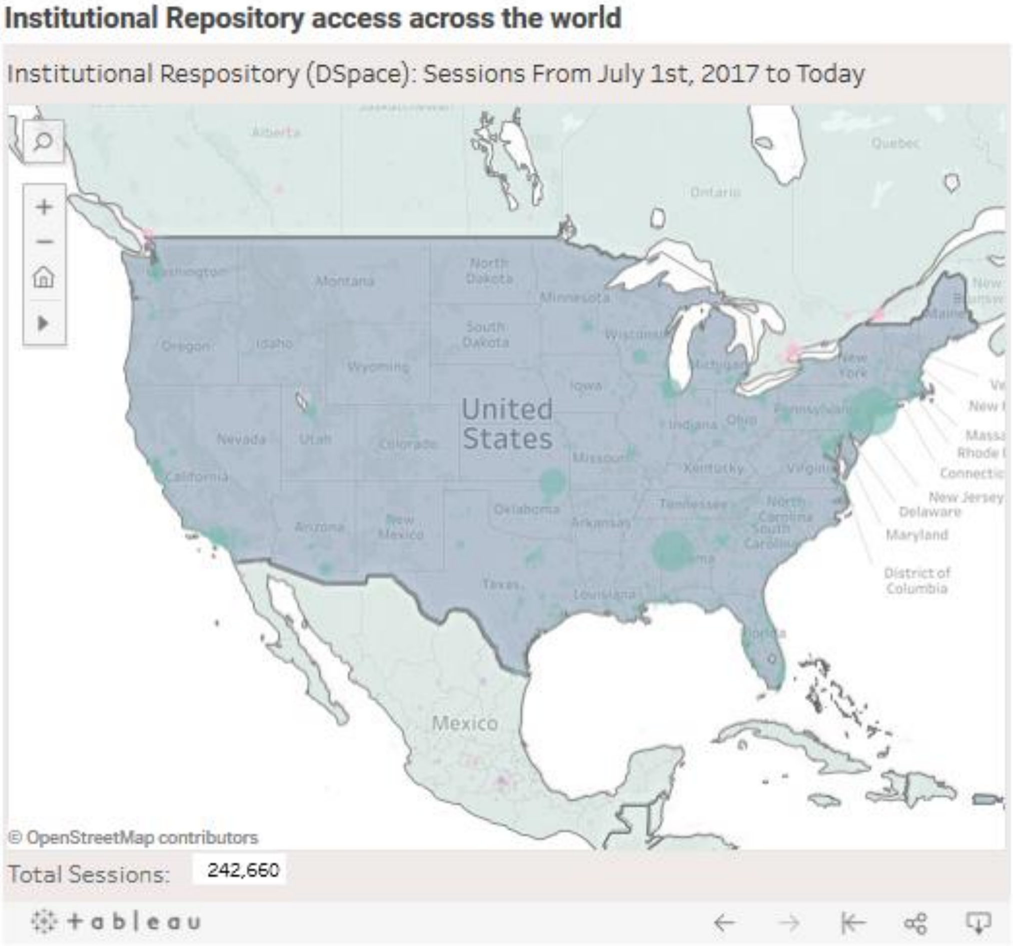


Institutional Repository access across the world

Institutional Respository (DSpace): Sessions From July 1st, 2017 to Today

Total Sessions: 242,660

The map allows users to zoom in to an area of interest or search for a place by name to see it closer, for instance, a user could search for United States, Alabama, or Tuscaloosa. Figure 2 displays the map zoomed in to view the contiguous United States. The large green dot over Tuscaloosa represents the large number of users it has from the University of Alabama. The capabilities of (1) dragging the sensor to select areas and (2) pressing the control and shift keys and clicking on the regions of interest to zoom in on an area makes Tableau a convenient medium for presentation.



Institutional Repository access across the world

Institutional Respository (DSpace): Sessions From July 1st, 2017 to Today

© OpenStreetMap contributors

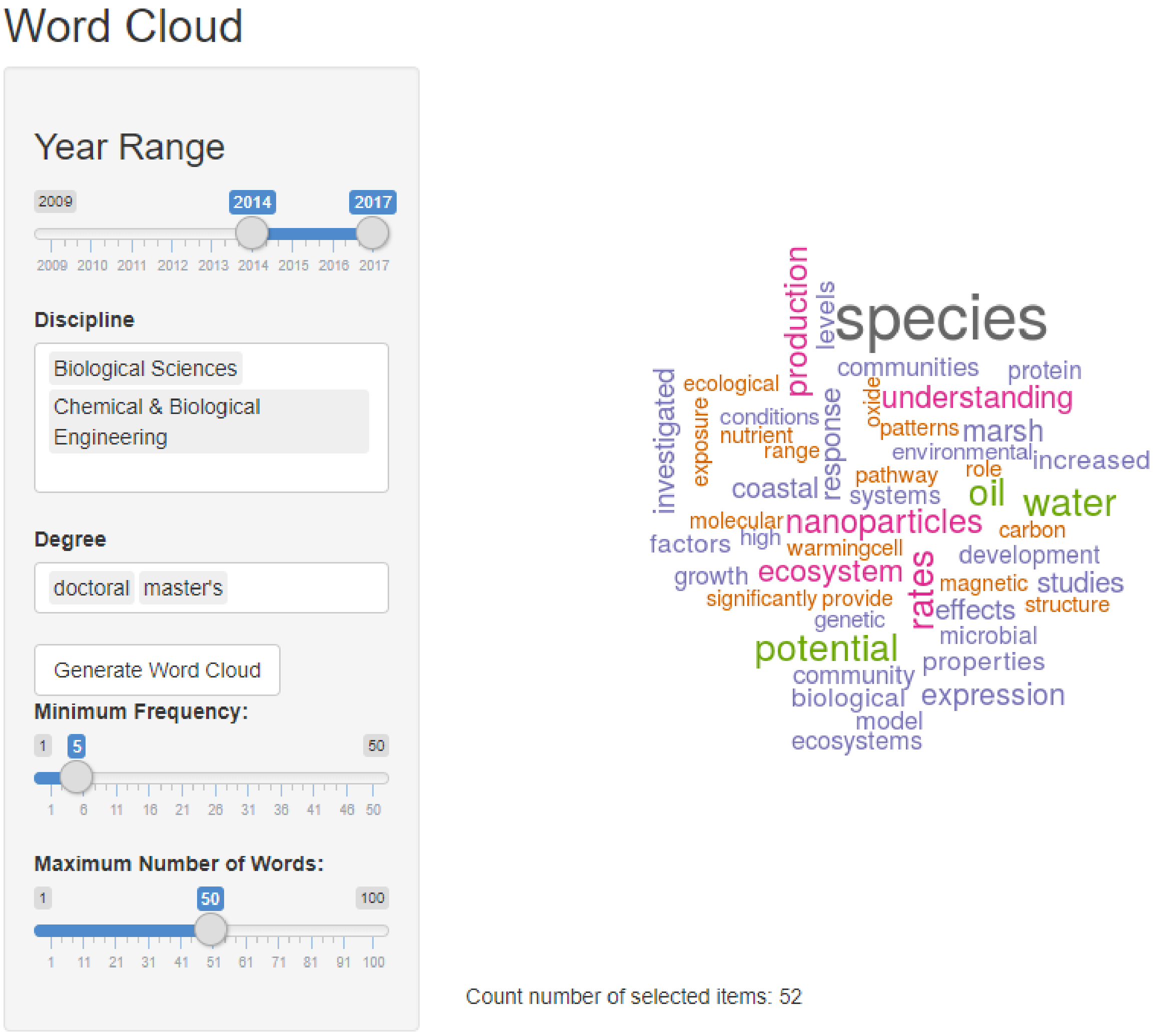
Total Sessions: 242,660

The IR map is built in Tableau, which makes use of filled areas, country and city, to show location. Given that (1) the location and (2) the number of sessions are two measures of interest, in the map they are represented by colors and sizes respectively. Other options for representing measures include shapes (e.g., square and circle) and tooltips (e.g., text label), but these options are not appropriate for our purposes. The built-in geographic codes search function of Tableau makes mapping easy. Datasets without latitude and longitude values can be plotted on the map simply. Tableau matches the location name information to its geographic databases automatically. In terms of data extraction, the IR web usage is recorded and archived in Google Analytics. Presumably, data from Google Analytics can be exported to Tableau directly, but Tableau seems to encounter data loss when connecting to Google Analytics. Unfortunately, this bug remained at the time this paper was written. We instead adopted the RGoogleAnalytics (Pearmain, Mihailowski, Prajapati, Shah, & Remy, 2014) package to continuously extract web metrics from Google Analytics, create a local database with the extracted web metrics, and finally form the map based upon the local web metrics database.

Facilitating data discovery and the analytic communication process, Tableau is designed to help users produce explanatory graphics and dynamic and interactive visualizations (Jones, 2014). Although essentially Tableau deploys VizQL, a programming language for describing tables, charts, graphs, maps, time series and tables of visualizations (Hanrahan, 2006), the Tableau interface is user-friendly, such that simple clicks of sensors can accomplish complex tasks. Tableau has great interactive functions and a mobile-friendly interface, but the lack of the capability to drag and move the map via a mouse is not complaisant with typical user experience. In addition, although not the most expensive option, the price of Tableau licenses is higher than many products of its kind; in fact, many open source libraries offer similar functions, but cost nothing. Similarly, a host fee is required for launching R applications to the Shiny server, if functions and/or services that are practically essential to library needs are selected. Like many other software programs, free versions of Tableau and Shiny are available, but they are not realistic choices for sites with large visitor numbers. Other well-known visualization options include Microsoft Power BI Domo, QlikView, and MicroStrategy Analytics Desktop (Boost Labs, 2015). General software applications using open source libraries include D3.js, Chart.js, and Leaflet (Sharma, 2015). Choosing the appropriate one (or more) of these options for an organization is undoubtedly difficult, as there are no definitive benchmarks for evaluating the benefits of adopting a certain tool with optional costs. Although discussing the pros and cons of these software programs and libraries is beyond the scope of the present paper, in the decision process one should consider (1) the skill sets of current and future employees, (2) the targeted audiences and the size of datasets fed into visualizations, and (3) the budget constraints of the organization. For example, if employers assigned to handle visualization tasks have little programming experience, R, D3.js, Chart.js, and Leaflet may not be viable options as the learning curves are relatively steep.

Wordcloud Visualization

The wordcloud allows users to choose years, disciplines, and levels of education of interest to them. Users can also manipulate the wordcloud by setting the minimum number of times a word must appear in the abstract of the document in order to appear in the wordcloud and by setting the maximum number of words the wordcloud will display. To give users an idea of how to use the tool and expose them to a variety of possible search results, we set the wordcloud's default starting page to randomly display different combinations of years and disciplines. This demonstrates to users that they are not limited to searching a single discipline, but can combine related disciplines to generate their cloud. Disciplines are not cleanly distinct from one another, but often overlap, so this functionality allows users to view trending topics from multiple interconnected disciplines at the same time. For instance, the following pairs of disciplines are recognized for sharing substantive similar research topics: (1) applied statistics and mathematics (Moore & Cobb, 2000), (2) educational measurement and cognitive psychology (Mislevy, 2006), and (3) electrical engineering and computer science (Deo, 2017). Figure 3 is a screen shot with the conditions that (1) time span is 2014 to 2017, (2) disciplines are Biological Sciences and Chemical & Biological Engineering, and (3) both doctoral and master degrees are included. Using the default setting in both 'Minimum Frequency' and 'Maximum Number of Words,' it can be seen that terms such as 'species' and 'nanoparticles' are prevalent among research works hosted in the institutional repository. As displayed below the wordcloud, the number of items comprising the cloud is 52 for the given condition.



Word Cloud

Year Range: 2014 – 2017

Discipline: Biological Sciences; Chemical & Biological Engineering

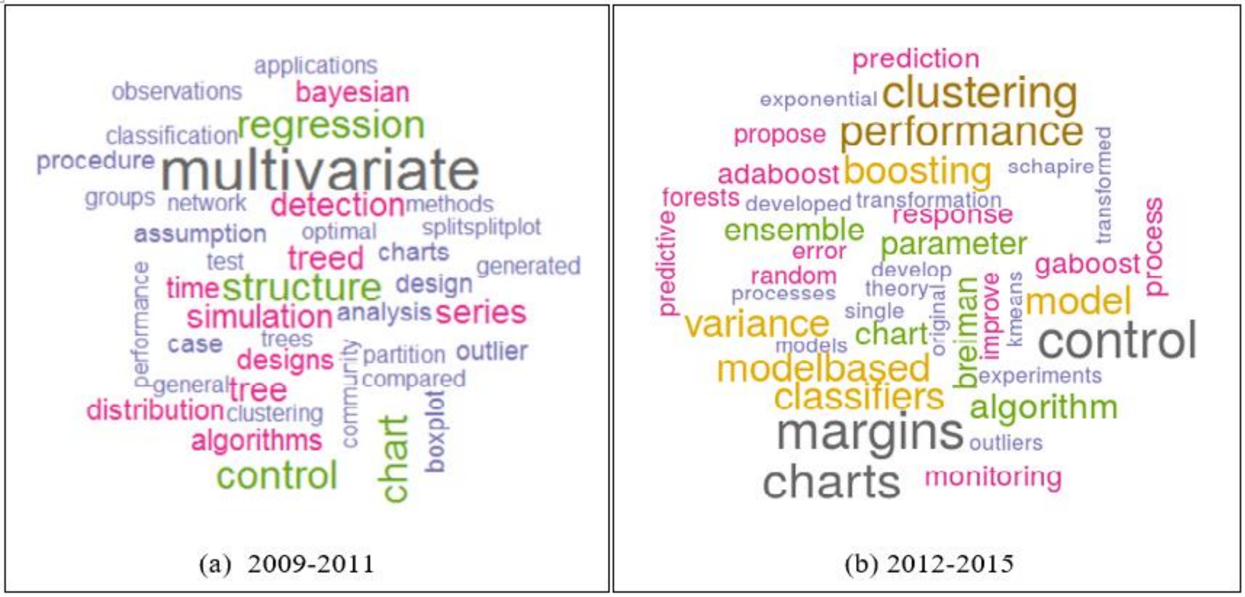
Degree: doctoral; master's

Minimum Frequency: 5

Maximum Number of Words: 50

Count number of selected items: 52

To further demonstrate the tool’s value, we now provide an example comparison of a discipline’s most used words over time. The discipline of applied statistics provides the example. Figure 4 shows a comparison between the discipline’s most used words in 2009-2011 and 2012-2015, provided both degrees (doctoral and master’s) are included and ‘Minimum Frequency’ and ‘Maximum Number of Words’ are set to 4 and 40 respectively. It can be seen that ‘multivariate’ and ‘regression’ are the mainstream research topics from 2009 to 2011, while the trending terms become ‘clustering,’ ‘boosting’ and other boost-related works in the following three years. These changes reflect the fact that machine-learning techniques became popular in the applied statistics field after 2010 (Alpaydin, 2014). Clustering, boost, and ensemble are powerful tools for recent big data problems, which were not an appreciable concept in the previous time period. The wordcloud output accurately reflects these topic shifts of the discipline.



(a) 2009-2011

(b) 2012-2015

The software we employed to create the wordcloud is R (R Core Team, 2017), which is one of the world's most popular programming languages due to its cost-free availability, flexible extensions, rapid package updates, and active community support (Roberts, Best, Dunn, Treml, & Halpin, 2010). Specifically, the wordcloud package (Fellows, 2015) is used to create graphs and the Shiny package (Chang, Cheng, Allaire, Xie, & McPherson, 2017) is implemented to wrap the graphs into web-based applications. Detailed instructions about combining R visualizations and Shiny can be found in Moon (2017).

The wordcloud retrieves dissertation and thesis abstracts from the IR database. The abstracts (technically text files) were transformed to a computer-readable format prior to the cloud generation. A text mining package, tm, by Feinerer, Hornik, and Meyer (2008) was loaded to (1) create a corpus from the abstracts, (2) decode the corpus by removing punctuation, symbols, whitespaces, and meaningless terms, and (3) convert the purified corpus into a structured dataset which is ready for generating wordclouds. In the creation of the stopword list, it is inevitable that both false positives (i.e., redundant words are labeled as necessary) and false negatives (i.e., essential words are filtered out) would occur during the text mining process. Therefore, making educated judgements is critical to produce meaningful wordclouds. That said, in addition to a set of stopword lists from the Snowball stemmer project (a small string processing language designed for creating stemming algorithms for use in Information Retrieval; http://snowball.tartarus.org/), we also include words such as “thesis,” “research,” “dissertation”, “chapter”, “study”, “data”, and similar terms that are content-free in the present application. After tm package execution, the structured datasets were further fed to the wordcloud package (Fellows, 2015) for producing clouds. The order of words in any given cloud is randomly generated, while the sizes of the words are directly proportional to the frequency of occurrence of the word in the abstracts after the text mining procedure. For added readability, color is also assigned to the words based on frequency.

One limitation of this project is the data handling problem we encountered. Like any other data cleaning tasks, the institutional repository has missing/unclassified and misclassified data in both applications (i.e., the IR map and the wordcloud). Google Analytics, where the IR usage data are stored, is unable to trace all user locations. This produces incomplete records occasionally. That is, some data points do not have location records and therefore cannot be projected on the map. Like other missing data problems, if there are systematic patterns hidden behind these missing records (i.e., all missing records came from the same region), the information presented by the IR map will be unrepresentative of real usage and therefore inappropriate (see Ender, 2010 for missing data problems). In addition, blank cells and incorrect labels were found during the investigation of the wordcloud database. For instance, some records did not have the labels for the 'discipline' column, while other records in the same column were mistakenly filled with the names of academic departments. When the scale of these data problems is small, an educated guess can sufficiently handle these problems by editing the records according to their content. However, if the problems occur on thousands of records, manually categorizing may become laborious and unrealistic. Another limitation of the wordcloud is processing text mining of abstracts (i.e., metadata) instead of full-text files. Choosing abstracts or full-text files for text mining has been controversial (Westergaard, Stærfeldt, Tønsberg, Jensen, & Brunak, 2017) because using abstracts may exclude and/or underrepresent data, while full-text files are always word-redundant and time-consuming for processing. To process text mining faster and therefore provide interactive features instantly, we choose abstracts instead of full-text files. Finally, the selection of stopwords can cause problems too. Even with careful filtering, the stopwords can create biases. For example, the term 'factor' may contain little actual meaning in some narrative works, but can reflect more valuable information in quantitative works as it may represent 'factor analysis,' a modern statistical method used to describe variability among observed and unobserved variables.

Research Method

To investigate the promotion effect rendered by both the map and the word cloud additions to the University of Alabama IR website, a changepoint analysis was adopted to detect the change in number of website sessions within a given period. With changepoint detection, one can identify the location of an event at a given longitudinal scale; it has been adopted in many pre- and post- tests and/or single subject studies in various fields. For example, Beem (1995) conducted a changepoint analysis for a mental-rotation experiment to investigate the intervention time, Zhao and Chu (2010) used it to model typhoons and heavy rainfall, and Luna, Garver, Urban, Lazar, and Sweeney (2004) implemented the technique in predicting the maturation of cognitive processes from late childhood to adulthood. Prior to describing the actual investigation, a review of changepoint analysis is provided in the following section.

According to Killick and Eckely (2014), changepoint analysis is a means of "estimating the point at which the statistical properties of a sequence of observations change." Detecting such changes is important, as it provides information about the effectiveness of a certain event or intervention. A straightforward example is a pharmacodynamics study where researchers are examining how fast a medicine of interest is taking effect. More formally, assume a vector of sequence data, $\mathrm{y}_{1:\mathrm{n}}$= ($\mathrm{y}_1$, y_2, …, y_n) is presented where *n* is the time label. A changepoint occurs at time τ ∈ {1, …, *n*} (i.e., between y_1 and y_n), such that the statistical properties of {y_1, … , y_τ} and {$\mathrm{y}_{\tau+1}$, . . . , y_n} are significantly different. To define the difference, the likelihood-based framework, a prevalent choice of statistical kernel, is selected to model the data. Like many other well-known parametric statistical techniques such as *Student's t*-test and χ^2 test, the changepoint detection has a null hypothesis, H_0, corresponding to a model with no changepoint. Therefore, the alternative hypothesis, H_1, is a changepoint exists. Statistical details can be found in Hinkley (1970), Gupta and Tang (1987), and Silva and Teixeira (2008). Naturally, this idea can be extended to multiple changepoints scenarios.

There are several variants of the changepoint analysis. For instance, the selection of objective functions (i.e., the estimation criteria), the components that are defined as change/constant, the number of changepoints, the penalty approaches on the objective functions, etc. In this paper, a Normal distribution and a Poisson distribution were adopted in the analysis and model comparisons. A Poisson distribution is for modeling count data, which is appropriate for the visit counts. However, it is not necessary to fit count data to a Poisson distribution; a Normal distribution can approximate a Poisson distribution and produce better model fits in some situations. In addition, both the mean and the variance of the data were assumed to be changing; this provides more flexibility than the practice of only assuming one is changing while the other is fixed. The aim of the present paper is detecting the promoting effect of the launched products; In order to achieve that, we selected a multiple change analysis such that the number of the change events could be automatically detected such that their means and variance would be available.

The hypothesis test is based upon the likelihood ratio test, meaning that if a changepoint doesn't exist, the model with and without the changepoint should not be statistically different. This approach requires the calculation of the maximum log-likelihood under both null and alternative hypotheses. The log-likelihood function for the model without the changepoint (i.e., the null hypothesis) is $\log p(\mathrm{y}_{1:\mathrm{n}} | \widehat{\boldsymbol{\theta}})$, where $p(\cdot)$ is the probability density function for the distribution of data and $\widehat{\boldsymbol{\theta}}$ contains estimated parameters of the model. Typically, researchers use a Normal distribution to model $p(\cdot)$ and therefore the mean and the variance are the $\widehat{\boldsymbol{\theta}}$, while with a Poisson distribution, there is only one parameter-$\boldsymbol{\lambda}$ For the alternative hypothesis where a changepoint τ is assumed to exist, let the subscript 0 and 1 represent before and after the changepoint, the log-likelihood function can be defined as:

$$\mathrm{LogL}\ (\tau) = \log p(\mathrm{y}_{1:\tau} | \widehat{\boldsymbol{\theta_0}}) + \log p(\mathrm{y}_{1+\tau:\mathrm{n}} | \widehat{\boldsymbol{\theta_1}}).$$

Let the log-likelihood for the null hypothesis LogL (null) equal $p(\mathrm{y}_{1:\mathrm{n}} | \widehat{\boldsymbol{\theta}})$. The log-likelihood difference between the null and the alternative hypotheses- can be expressed as:

$$L = \mathrm{LogL}\ (\tau) - \mathrm{LogL}\ (\mathrm{null}).$$

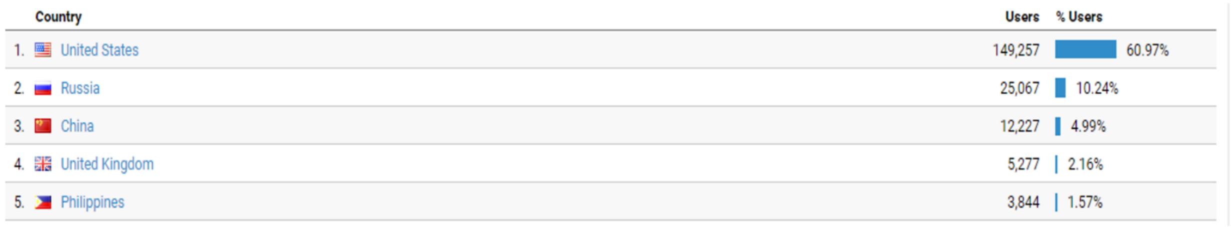
Given a L, one can use the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) to conduct model selection. AIC and BIC are defined as:

$$\mathrm{AIC} = 2\mathrm{k} - 2L,$$

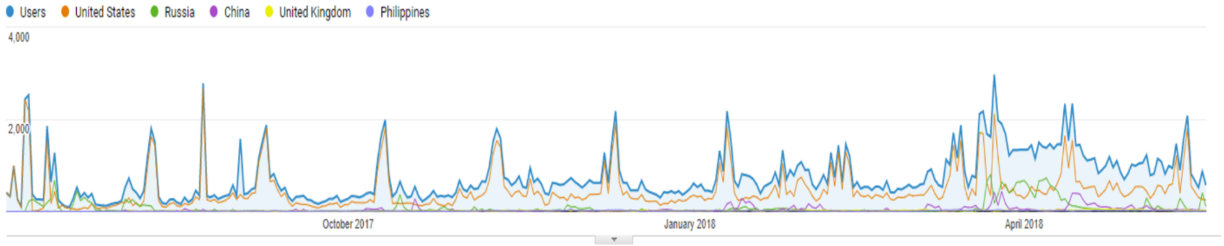
$$\mathrm{BIC} = \ln(\mathrm{n})\,\mathrm{k} - 2L,$$

where k is the number of parameters estimated by the model. In the changepoint situation, if AIC is used as a criterion and there is least one τ such that $\mathrm{AIC}(\tau)$ is smaller than $\mathrm{AIC}(\mathrm{null})$, the τ is obtained by minimizing AIC. A similar approach is applied to BIC. As the AIC has been shown to not provide a consistent estimate of the number of changepoints, we use BIC in this paper. The analysis was implemented via the changepoint package by (Killick & Eckley, 2014).

The daily session count for the IR webpage was extracted from July 1st, 2017 to May 20th, 2018; each data point indicates the session count of a day. As Figure 5 shows, there were fluctuations that can be attributed to the seasonal patterns of a school year. For example, the session count will drop down in December when classes end and bounce back in January when classes begin again. During the selected time span, over 242,000 sessions were captured by Google Analytics. The upper panel in Figure 5 shows the top five countries contributing to the session count dataset: as expected, the United States took the largest proportion of the share (nearly 61%), while countries such as Russia, China, and the Philippines ranking high was not as easily explained. Although population size is a potential reason for this phenomenon, further scientific investigations on the user country distribution are needed. Corresponding to the country rank, the lower panel in Figure 5 shows the longitudinal records for the world (labeled as ‘Users’ for the cumulative number) and all five top countries. Given the United States played a dominant role in the user shares, it is not surprising that the trends of the blue and the orange lines were consistent.



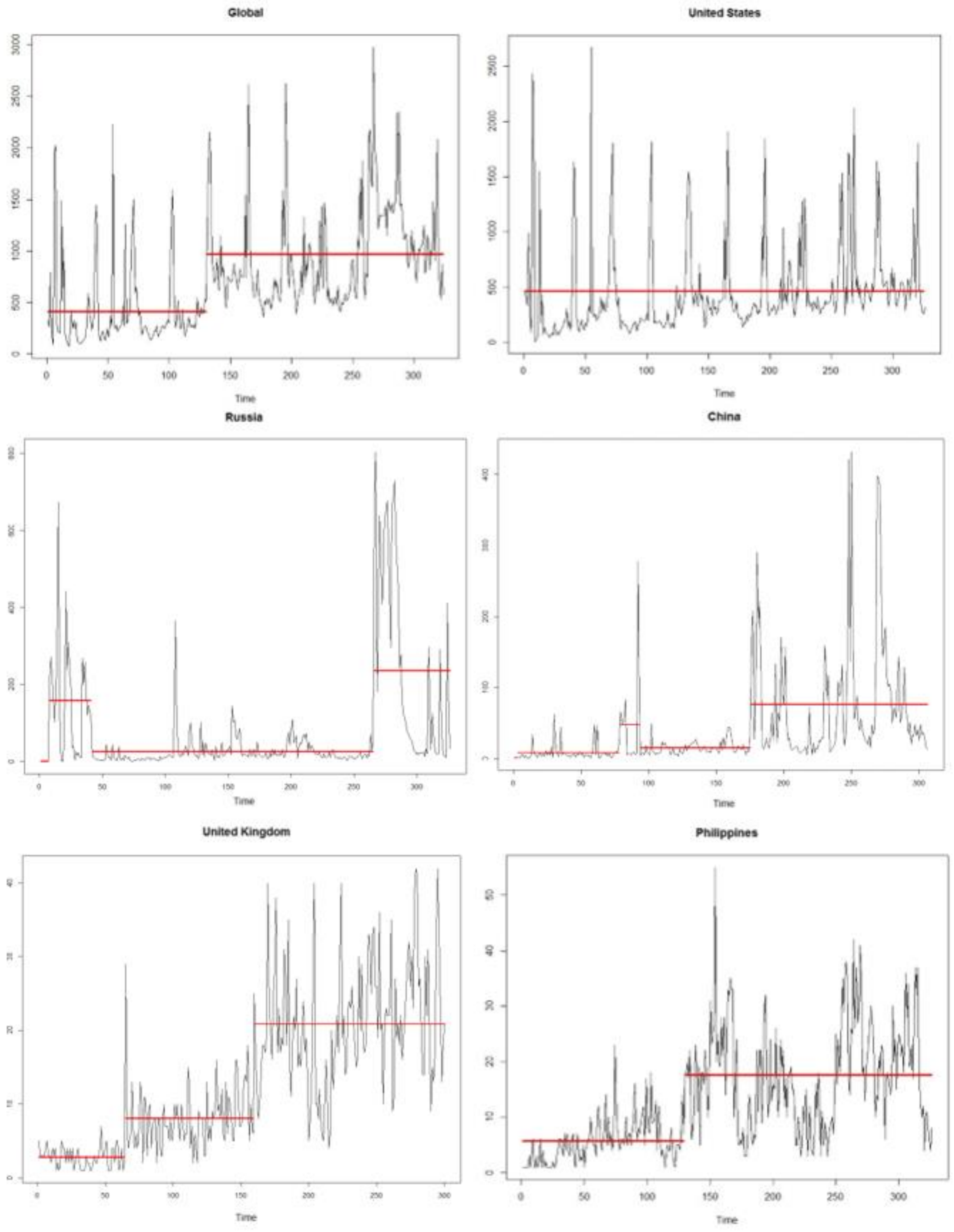
	Country	Users	% Users
1.	United States	149,257	60.97%
2.	Russia	25,067	10.24%
3.	China	12,227	4.99%
4.	United Kingdom	5,277	2.16%
5.	Philippines	3,844	1.57%



Findings

We started with a single change analysis. Our goals here were to determine (1) if changes happened during the observation, (2) the number of the changepoints, and (3)whether there were changepoints after the date the visualizations were added to the website, October 30, 2017. We started by feeding data at the global level to the algorithm. The changepoint analysis result is listed in Figure 6; the straight horizontal lines represent the means of the partitioned sub-time series. In all situations, the BIC values of using a Poisson distribution were lower than those of a Normal distribution. The model supported the decision of splitting the series into two sets at day 150, which was November 27, 2017. The BIC indicated that the model fit: the difference in the statistics between the null model and the τ was 11.6. For all global users, the mean and the standard deviation for the series before November 27, 2017 were 417 and 400, where the same statistics for the second period were 970 and 488. The United States, on the other hand, did not show significant segment patterns in the analysis. The BIC indicated that the null model fitted the United States better. Therefore, no changepoints were detected, meaning that the newly available website visualization tools did not influence the users in the United States significantly.

Following the same method, the influences on users from the four other top countries aside from the U.S. were investigated. These countries all showed changepoints and therefore were the driving force of the global changepoint. Russia showed four changepoints: days 7, 41, 106, and 265. The initial launching did not immediately increase the visits: the last changepoint occurred with a lag far after the addition of the new visualizations, and therefore, it is not clear what factors drove the change. The changepoints of China indicated three times of change: overall the trend is similar to the global one, where between days 78 and 93 it encountered a spike. The larger change took place on day 175, which is a post-launching day. The UK showed an increasing trend where two changepoints were discovered: days 64 and 159 separated the wave into three parts which, again, showed that after the launching day the visit number had a significant increase. Last but not least, the Philippines's trend was close to the global one: there was only one change point that happened within the 50-day-window after the web site enhancements.


Global
0
500
1000
1500
2000
2500
3000
0
50
100
150
200
250
300
Time
United States
0
500
1000
1500
2000
2500
Time
Russia
0
200
400
600
800
Time
China
0
100
200
300
400
Time
United Kingdom
0
10
20
30
40
Time
Philippines
0
10
20
30
40
50
Time

Discussion

Our goal with these visualizations was to appeal to scholars both as authors of scholarly content looking for audiences, as well as consumers of scholarly content looking for easily accessible and minable research. In the digital age, scholarly works can be both a secondary source to provide context for new scholarship and a primary source which is machine searchable for the discovery of trends. By demonstrating the twofold benefits of making scholarship open access, we hoped to attract greater participation in institutional repositories. The lack of increase in the rate of growth of traffic to the IR website from within the U.S. suggests faculty and students at the University of Alabama were not influenced by the addition of these visualizations. This may be due to a lack of marketing of the visualizations rather than a lack of influence on those who view them. The uptick in the rate of increase of traffic to the IR website from countries outside the U.S. was an unexpected finding for which we currently have no explanation.

We hope our project inspires other libraries to create visualizations and discovery systems for their institutional repositories that encourage their growth. Increasing the amount of scholarship which is available open access is important to disseminating research to scholars in developing countries, scholars at institutions with limited subscription budgets, policy makers who rely on scholarly research for their decision making, educators at the K-12 level, and practitioners in fields that depend on the ever-changing body of knowledge produced by university research. Creating analysis tools like the wordcloud at other institutions could allow for illuminating comparisons between universities. We plan to continue tracking traffic in our institutional repository with these visualizations in place to verify that interest in the repository is growing. This study has highlighted the need for disseminating information about the tools we offer. In terms of next steps for raising awareness of the institutional repository at the University of Alabama, the authors plan to distribute a survey about open access options to university faculty and graduate students. We also plan to increase our efforts at outreach to faculty on the topics of data storage and preservation.

In addition to the value of visualizations for faculty, they can be a tool for prospective students. The wordcloud tool can be of use to prospective graduate students in identifying the strengths of the graduate programs offered by the institution. Furthermore, creating a positive impression of institutional repositories early in a scholar’s formative study increases the likelihood that they will make informed choices about open access publication as they grow into publishing scholars.

References

Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge, MA: MIT press.

Austin, A. E. (1991). Faculty cultures, faculty values. *New Directions for Institutional Research, 1990*, 68, 61-74. Doi: https://doi.org/10.1002/ir.37019906807

Beem, A. L. (1995). A program for fitting two-phase segmented-curve models with an unknown change point, with an application to the analysis of strategy shifts in a cognitive task. *Behavior Research Methods, Instruments, & Computers*, *27*(3), 392-399. Doi: https://doi.org/10.3758/BF03200435

Boost Labs. (2015). *4 great data visualization tools for data analysis and insights.* Retrieved from http://www.boostlabs.com/4-great-data-visualization-tools-data-analysis-insights/

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2017). *Shiny: Web Application Framework for R.* Version 1.0.5. [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/shiny/shiny.pdf

Deo, N. (2017). Graph theory: With applications to engineering and computer science. Mineola, NY: Courier Dover Publications.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of statistical software*, *25*(5), 1-54. Doi: 10.18637/jss.v025.i05

Fellows, I. (2015). *Wordcloud: Word Clouds R package version 2.5*. [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

Gargouri, Y., Larivière, V., Gingras, Y., Carr, L., & Harnad, S. (2012). *Green and gold open access percentages and growth, by discipline*. Retrieved from https://arxiv.org/abs/1206.3664

Georgia State University Library. (2017). *ScholarWorks @ Georgia State University*. Retrieved from https://scholarworks.gsu.edu/

Gupta, A. K., & Tang, J. (1987). On testing homogeneity of variances for Gaussian models. *Journal of Statistical Computation and Simulation*, *27*(2), 155-173. Doi: https://doi.org/10.1080/00949658708810988

Hajjem, C., Harnad, S., & Gingras, Y. (2006). *Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact*. Retrieved from https://arxiv.org/ftp/cs/papers/0606/0606079.pdf

Hanrahan, P. (2006). Vizql: a language for query, analysis and visualization. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (pp. 721-721). ACM. Doi: 10.1145/1142473.1142560

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 1-17. Retrieved from http://www.jstor.org/stable/2334932

Jones, B. (2014). Communicating data with tableau: designing, developing, and delivering data visualization. Sebastopol, CA: O'Reilly Media.

Killick, R., & Eckley, I. (2014). Changepoint: An R package for changepoint analysis. *Journal of statistical software*, *58*(3), 1-19. Doi: 10.18637/jss.v058.i03

Kim, J. (2010). Faculty self-archiving: Motivations and barriers. *Journal of the American Society for Information Science and Technology, 61,* 9, 1909-1922. DOI: 10.1002/asi.21336

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child development*, *75*(5), 1357-1372. DOI: 10.1111/j.1467-8624.2004.00745.x

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement,* 4, (pp. 257-305). Westport, CT: Praeger Publishers.

Moon, K. W. (2017). *Learn ggplot2 Using Shiny App*. Springer.

Moore, D. S., & Cobb, G. W. (2000). Statistics and mathematics: Tension and cooperation. *The American Mathematical Monthly*, 107(7), 615-630. Doi: 10.2307/2589117

Pearmain, M., Mihailowski, N., Prajapati, V., Shah, K., & Remy, N. (2014). *RGoogleAnalytics: R Wrapper for the Google Analytics API. R package version 4.0.* [Computer software manual]. Retrieved from http://code.markedmondson.me/googleAnalyticsR/

Polley, D.E. (2016). Visualizing the topical coverage of an institutional repository using VOSviewer. In L. Magnuson (Ed.), *Data visualization: A guide to visual storytelling for librarians*. Lanham, MD: Rowman & Littlefield.

R Core Team. (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Roberts, J. J., Best, B. D., Dunn, D. C., Treml, E. A., & Halpin, P. N. (2010). Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environmental Modelling & Software*, *25*(10), 1197-1207. Doi: 10.1016/j.envsoft.2010.03.029

Sharma, N. (2015). *The 14 best data visualization tools*. Retrieved from https://thenextweb.com/dd/2015/04/21/the-14-best-data-visualization-tools/

Silva, E. G., & Teixeira, A. A. (2008). Surveying structural change: Seminal contributions and a bibliometric account. *Structural Change and Economic Dynamics*, *19*(4), 273-300. Doi: https://doi.org/10.1016/j.strueco.2008.02.001

Sterman, L. B., & Borda, S. (2017). Making visualization work for institutional repositories: Information visualization as a means to browse electronic theses and dissertations. *Journal of Librarianship and Scholarly Communication*, *5*(1), eP2140. DOI: http://doi.org/10.7710/2162-3309.2140

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000research, 5*. Retrieved from https://f1000research.com/articles/5-632/v3

Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics, 103,* 2, 555-564. Retrieved from https://f1000research.com/articles/5-632/v3

Westergaard, D., Stærfeldt, H. H., Tønsberg, C., Jensen, L. J., & Brunak, S. (2017). Text mining of 15 million full-text scientific articles. bioRxiv, 162099. Retrieved from https://www.biorxiv.org/content/biorxiv/early/2017/07/11/162099.full.pdf

Wolff, C., Schonfeld, R. C., Rod, A. B., & Ithaka S + R. (2016). *Ithaka S+R US library survey 2015*. Retrieved from http://www.sr.ithaka.org/wp-content/uploads/2016/03/SR_Report_US_Faculty_Survey_2015040416.pdf

Zhang, H., & Lopez, C. (2017). An interactive map for showcasing repository impacts. *The Code4lib Journal*, 36. Retrieved from http://journal.code4lib.org/articles/12349

Zhao, X., & Chu, P. S. (2010). Bayesian changepoint analysis for extreme events (typhoons, heavy rainfall, and heat waves): An RJMCMC approach. *Journal of Climate*, *23*(5), 1034-1046. Doi: https://doi.org/10.1175/2009JCLI2597.1