Developing Models to Visualize & Analyze User Interaction for Financial Technology Websites

Project Team

Amanda Ezeobiejesi	Computer Science
Guy Katz	Industrial Engineering
Alissa Ostapenko	Computer Science and Mathematics

Project Advisor

Professor Michael Ginzberg
Foisie Business School

Project Co-Advisors

Professor Rodica Neamtu
Department of Computer Science

Professor Sara Saberi
Foisie Business School

Professor Jon Abraham
Department of Mathematical Sciences

Abstract

Vestigo Ventures manually processes website traffic data to analyze the business performance of financial technology companies. By analyzing how people navigate through company websites, Vestigo aims to understand different customer activity patterns. Our team designed and implemented a tool that automatically processes clickstream data to visualize different customer activity within a website and compute statistics about user activity. This tool will provide Vestigo insight on the effectiveness of their clients' website structures and help them make recommendations to their clients.

Acknowledgements

Our team would like to extend our sincere gratitude to the individuals mentioned below, from Vestigo Ventures, Cogo Labs, and Link Ventures, as well as our advisors for their support and encouragement throughout the duration of this project:

Vestigo Ventures
Ian Sheridan, a managing director and co-founder of Vestigo Ventures, who made this project possible. He established the initial connection with WPI and checked in on us from time to time.

Frazer Anderson, a data-driven strategist, an investment analyst at Vestigo Ventures, and our project sponsor. He provided us with the resources needed and support, enabling our project's success.

Cogo Labs
Rob Fisher, the newly appointed Chief Executive Officer at Cogo Labs, was the mastermind behind our project. His thoughts, guidance, and consistent feedback were vital to the success of this project.

Daniel Brady, an analytics manager at Cogo Labs, helped us to fully understand the competitive intel database and user interface that held the company's various databases.

Link Ventures
Todd Federman and D'Mitri Joseph, technical fellow and software engineer, respectively, at Link Ventures helped us integrate our API into the existing business analysis tool, making our project a fully functioning application that anyone with proper access to the tool can use.

Our Advisors
Professor Michael Ginzberg, the advisor for the Fintech/Wall Street Project Center, ensured that the project was conducted smoothly by giving us access to the resources we needed prior to and during the project.

We thank our major project advisors, Professors Rodica Neamtu (Computer Science), Sara Saberi (Industrial Engineering), and Jon Abraham (Mathematical Science) who supported our team throughout the entire duration of the project. Their enthusiasm, kind attention, and diligent guidance increased our passion for the project and allowed the MQP to be as impactful as possible.

Executive Summary

Introduction

Vestigo Ventures, a venture capital firm investing in financial technology companies, aims to understand the different ways that people interact with their clients' websites and how many visitors make a purchase. However, Vestigo's clients have diverse website structures and varying definitions of what a purchase is, making it difficult for the firm to use a generalized visualization technique. Our goal was to create an easy-to-use tool that automatically processes internet traffic data to provide Vestigo, and similar companies interested in website performance analysis, insight into the effectiveness of a company website. We developed the *Website Private Investigator (WPI)*, an Application Programming Interface (API) that builds an interactive graph illustrating how people navigate through a given website. Moreover, our tool calculates statistics about customer interaction, such as the percentage of visits that start or end at a certain page within the company's website, allowing Vestigo analysts to understand customer activity patterns in depth.

In the following subsections, we discuss our project management approach, highlight the architecture, features, and performance analysis of our tool. In addition, we provide recommendations and takeaways from our project experience.

Methodology

We organized our work in four sprints, each two weeks long. Each sprint comprised of multiple meetings within the team and with our faculty advisors. Moreover, we consulted our company sponsor, and other knowledgeable employees from Vestigo's partnering company, Cogo Labs. Cogo is an incubator of internet companies and provided the internet traffic data we worked with to develop the *WPI*. We used the feedback from our meetings to guide our project work and to iteratively develop our tool in four phases:

(1) understand the dataset and experiment with visualization techniques using Python libraries,
(2) develop the API using Python and Github for version control,
(3) document our API, and
(4) deploy the API into production with a user interface (UI).

At the conclusion of our project, we produced a command-line interface (API) version of the *WPI*, as well as a deployed version complete with a graphical UI. To better visualize the components of our tool, we present our work in its deployed form.

Website Private Investigator - Architecture and Features

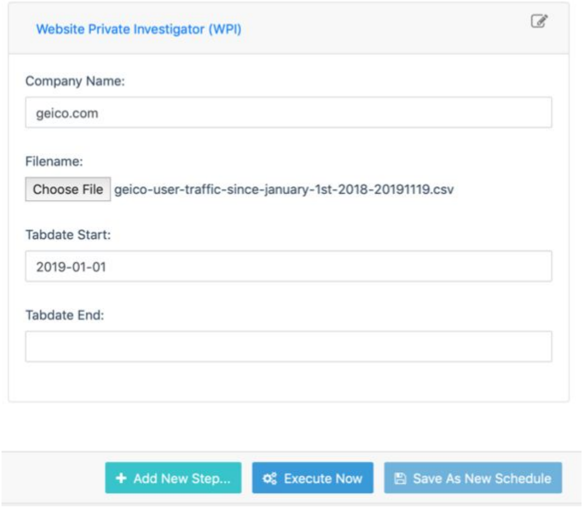
Phase 1: Gathering Data

To use our tool, an analyst must first query Cogo's internet traffic database with a company website and start date of interest. We provided a query template that Vestigo can follow to easily gather the data in the format that the *WPI* expects. After running the query, an analyst needs to download the results to a file, producing a dataset of Uniform Resource Locators (URLs) people visited while browsing a company website during an input date range.

Phase 2: Building the Graph

Starting a Job

As illustrated in **Figure 1**, The *Website Private Investigator* expects an input data file, company website name, and a data range within the dataset for processing. The start date must be specified, however, the end date is optional; by default, *WPI* will process all of the data present in the data file from the start date. Clicking "Execute Now" starts a job to build an interactive graph.



Website Private Investigator (WPI)

Company Name: geico.com

Filename: geico-user-traffic-since-january-1st-2018-20191119.csv

Tabdate Start: 2019-01-01

Tabdate End:

Figure 1: Specifying input arguments to the Website Private Investigator.

Processing the Data

Cleaning the Data

To remove user-specific information while retaining general activity patterns, we reduce each URL to only the company domain name and the website path. For example, if given the URL *https://www.wpi.edu/admissions/graduate/how-to-apply?itemId=item-27*, we simplified this to *www.wpi.edu/admissions/graduate/how-to-apply*.

Building User Paths

Our tool uses *pandas* to separate the dataset by unique visits to build user flow paths. A visit is defined by a unique combination of user ID and tabdate, and each flow path is a list of URLs. To better understand which pages people start and end their browsing activity, we group URLs into start pages, intermediate pages, and exit pages.

Clustering Similar URLs and Computing Statistics

To further summarize the different browsing information within the dataset, we use *difflib* to cluster start, intermediate, and exit URLs by string similarity. Finally, using Python packages *scipy* and *math*, we calculate the percentage of visits that landed on each start, intermediate, and end page.

Interacting with the Graph

After several minutes, an interactive graph will load in the browser, as shown in **Figure 2**. We developed the graph using Plotly (Python 3 and JavaScript) and NetworkX (Python 3), open-source graph visualization packages. Each cluster in the start, intermediate, and exit groups is a node in the graph, and edges connect nodes according to the paths built during the preprocessing step. Start web pages are colored green and placed at the top of the graph, exit web pages are red and placed at the bottom, and intermediate webpages are blue, placed between start and exit pages, to better visualize the different components of the different user flow paths.

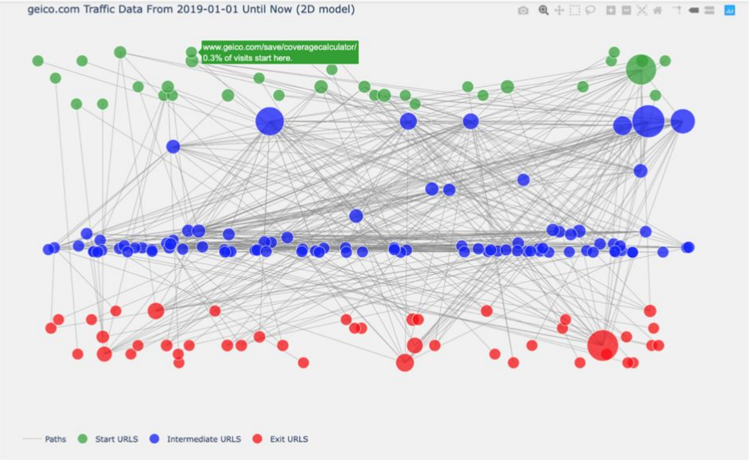

geico.com Traffic Data From 2019-01-01 Until Now (2D model)
www.geico.com/save/coveragecalculator/
0.3% of visits start here.
Paths
Start URLS
Intermediate URLS
Exit URLS

Figure 2: Interactive graph showing user traffic flow for geico.com

Hovering

As illustrated in **Figure 2**, hovering over any node will show the webpages represented by the node, as well as the percentage of visits that pass through these webpages.

Website Private Investigator Special Features

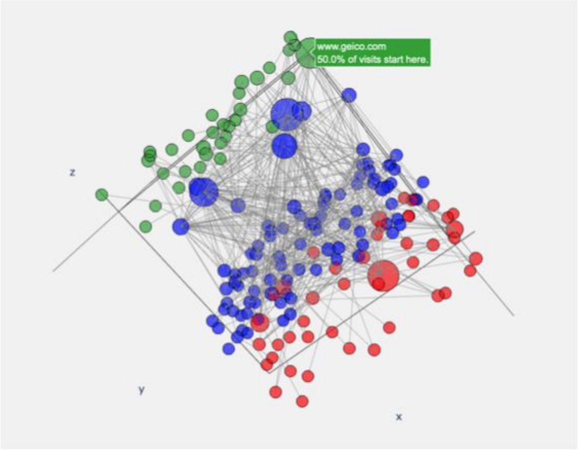


Figure 3: User Traffic Graph in Three Dimensions

Displaying the graph in Three Dimensions

The two-dimensional graph often has many overlapping nodes. For an analyst to view the graph from different angles, we added a 3D toggle which visualizes the data in three dimensions, as illustrated in **Figure 3**.

Highlighting Adjacent Nodes

An analyst may want to see which pages a user went to immediately before or after a particular page, especially if this is an exit page (such as a purchase page) or an entry page (such as a login page). The highlighting features allows an analyst to click on a node of interest to see the nodes immediately connected to it. **Figure 4** illustrates the possible pages of people went to immediately after checking their account pages on *Geico.com*.

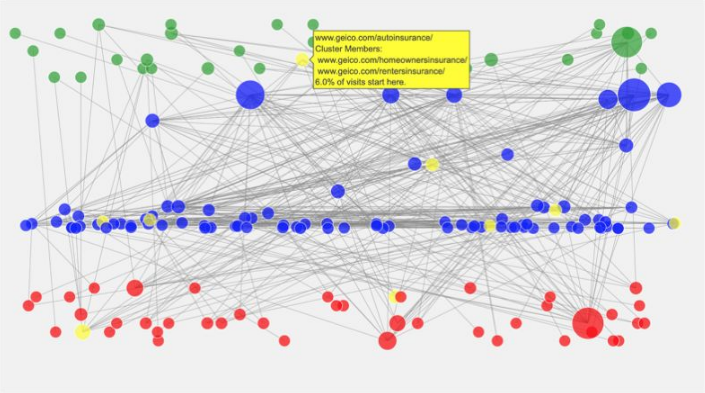

www.geico.com/autoinsurance/
Cluster Members:
www.geico.com/homeownersinsurance/
www.geico.com/rentersinsurance/
6.0% of visits start here.

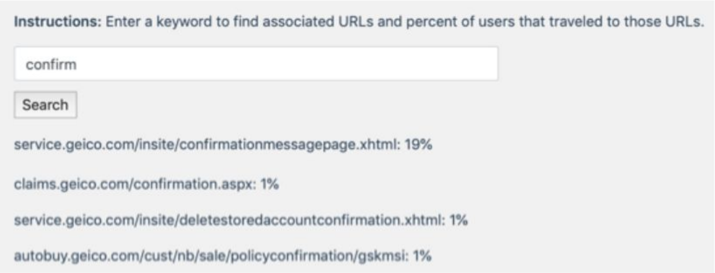
Figure 4: Highlighting pages immediately adjacent to insurance-related pages

Highlighting User Paths

Moreover, it may be useful to see all pages that people viewed if they passed through a particular webpage. As with *Highlighting Adjacent Nodes*, an analyst can click on a node to highlight all potential pages on a path containing this page. This way, the analyst can identify possible entry, intermediate, and exit points that people could have traveled to before, for example, confirming a purchase.

Keyword Search

Clustering is not a precise technique, and a webpage of interest may be clustered together with less relevant pages. To isolate visit statistics for particular pages of interest, an analyst can use the *WPI's* keyword search functionality to search for webpages containing a particular term, such as *confirm*. **Figure 5** illustrates the keyword search functionality.



Instructions: Enter a keyword to find associated URLs and percent of users that traveled to those URLs.

confirm

service.geico.com/insite/confirmationmessagepage.xhtml: 19%

claims.geico.com/confirmation.aspx: 1%

service.geico.com/insite/deletestoredaccountconfirmation.xhtml: 1%

autobuy.geico.com/cust/nb/sale/policyconfirmation/gskmsi: 1%

Figure 5: Keyword search of geico.com for *'confirm'*

Experimental Analysis

To evaluate our clustering methods, as well as explore financial technology website statistics more in-depth, we used techniques from mathematics and industrial engineering.

Clustering Evaluation

There are many techniques for grouping text data, however, we focus our analysis on Agglomerative Clustering and Gestalt Pattern Matching. Table 1 details our approaches to these techniques.

Agglomerative Clustering	Gestalt Pattern Matching
To capture local character-to-character differences between URLs, we compute the edit distances between each pair of URLs in a dataset. We used scikit-learn's implementation of agglomerative clustering to group URLs.	We use *difflib* to compute the Gestalt ratio between URL pairs and grouped together URLs above a threshold ratio. The Gestalt ratio ranges from 0 to 1, where 1.0 indicates a perfect match. We experimented with different ratios between 0.6 and 1.0. The Gestalt ratio reflects sequence-level comparisons between URLs.

Table 1: Clustering techniques

To evaluate our methods, we manually labeled URLs collected from four websites, one of which is *geico.com*. We compared the outputs of the agglomerative clustering and gestalt pattern matching to our labels by computing *Adjusted Rand Index (RI)* and *V-measure*, which are defined in **Table 2**.

Adjusted Rand Index	**V-measure**
Rand Index (RI) $\frac{a+d}{a+b+c+d}$ a: same cluster, same label b: same cluster, different label c: different cluster, same label d: different cluster and label *Adjusted Rand Index* (ARI): $\frac{RI - Expected(RI)}{Max(RI) - Expected\ (RI)}$	$V_\beta = (1+\beta)\frac{h*c}{\beta * h + c}$ *h*: homogeneity (a cluster should have only members of the same class) *c:* completeness (all class samples should be in the same clusters) β: *beta,* harmonic mean weight of *h* and *c*

Table 2: Clustering evaluation metrics

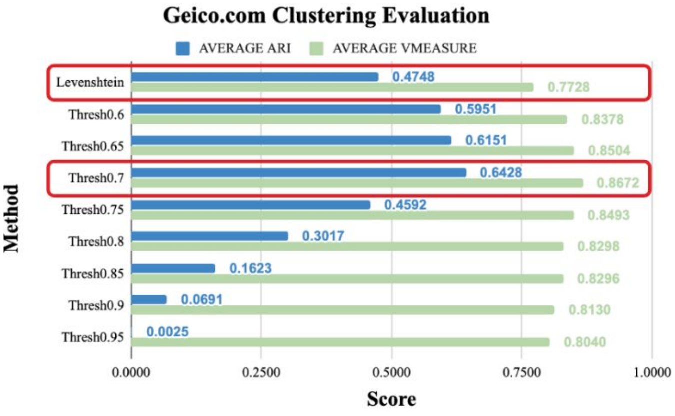


Figure 6: Clustering Evaluation Comparison

In general, Gestalt pattern matching with a threshold around 0.70 and 0.75 outperforms agglomerative clustering for URL grouping in financial technology websites. We programmed the tool to use a threshold of 0.75 based on our experimental results. **Figure 6** illustrates the ARI and V-measure of clustering URLs in *Geico.com* user traffic data.

Linear Regression

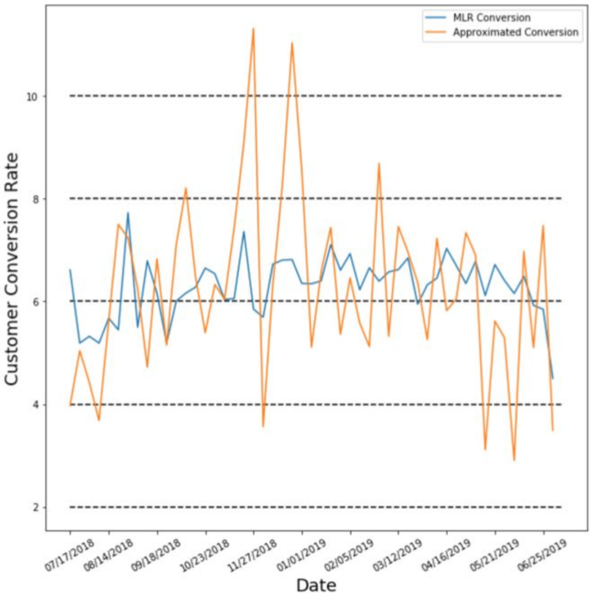


Figure 7: Time series linear regression conversion vs. approximated conversion

Vestigo requested to find ways to calculate the customer conversion rate of a website, or the percentage of visits that end in a customer purchase. Although it is difficult to compute exact conversion rates, we used the WPI tool to approximate the statistics. Using the financial company Wells Fargo as an example, we created a multiple linear regression of the conversion rate. The data we used had been collected over a period of 50 week. The statistical information for each week was collected from the *WPI* along with an approximated conversion rate, which was the percentage of visits that traveled through a web page with the keyword ‘billpay.’ Plotted in **Figure 7** is the multiple linear regression conversion rate compared to the approximated conversion. We can solve for Wells Fargo’s conversion rate using the following equation:

$$\textbf{\textit{Conversion Rate}} = -0.06*(\textit{unique visits}) + 0.09*(\textit{unique users}) + 0.15*(\textit{percent return users}) - 0.66*(\textit{average pages in a visit}) + 4.49$$

The equation is not accurate, with an r squared value of only 0.12. However, the general trend of the week to week customer conversion rate is similar to that of the approximated conversion.

Recommendations & Conclusions

Our tool works best for small data sizes. We recommend that users of the tool use at most 50 MB of data to ensure graph creation under one hour. On average, *WPI* will process datasets of approximately 10,000 rows in under 60 seconds. Furthermore, having at least 500 rows of data will ensure a good quality of the graph. While this might limit collecting data for time series regression models, it will ensure that the graph has enough data to provide useful information. Overall, the project allowed the team to apply interdisciplinary knowledge gained from classroom study to real-world data. The team learned how to work win a business setting, with diverse groups of people, and how to create business focused applications which can be used by analysts now and in the future.

The rest of the paper is removed by request from the sponsoring company.

References

"About-Cogo Labs." *Cogo Labs*, https://www.cogolabs.com/about.

Albert, Bill, and Donna Tedesco. "Clickstream Data." *Clickstream Data - an Overview | ScienceDirect Topics*, https://www.sciencedirect.com/topics/computer-science/clickstream-data.

Allahyari, Mehdi, et al. "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques." *ArXiv.org*, Cornell University, 28 July 2017, https://arxiv.org/abs/1707.02919v2.

Anderson, Frazer. Personal Interview. 9 Sept. 2019.

Black, Paul E. "Ratcliff/Obershelp pattern recognition." Dictionary of algorithms and data structures 17, 2004.

Bonaccorso, Giuseppe. Mastering Machine Learning Algorithms. Packt Publishing Limited, 2018.

Bonnin, Rodolfo, and Claudio Delrieux. Machine Learning for Developers: Uplift Your Regular Applications with the Power of Statistics, Analytics, and Machine Learning. Packt, 2017.

Brady, Daniel. Personal Interview. 22 October 2019.

Chen, James. "Venture Capital Definition." *Investopedia*, Investopedia, 29 Sept. 2019, https://www.investopedia.com/terms/v/venturecapital.asp.

"Conversion Funnel." Wikipedia, Wikimedia Foundation, 11 Oct. 2019, en.wikipedia.org/wiki/Conversion_funnel.

"Data Analysis." *Data Analysis - Pearson's Correlation Coefficient,* University of the West of England, Bristol, 2019, http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442.

Desai, Falguni. "The Evolution Of Fintech." *Forbes*, Forbes Magazine, 9 Feb. 2016, https://www.forbes.com/sites/falgunidesai/2015/12/13/the-evolution-of-fintech/.

Fisher, Robert. Personal Interview. 2 Oct. 2019.

"*GEICO At A Glance*." GEICO, https://www.geico.com/about/corporate/at-a-glance/.

Gilleland, Michael. "Levenshtein Distance, in Three Flavors." *Levenshtein Distance, in Three Favors*, Merriam Park Software, https://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein Distance.htm.

Han, Jiawei, et al. *Data Mining: Concepts and Techniques*. Elsevier, 2012.

"Home-Vestigo." *Vestigo Ventures*, https://www.vestigoventures.com/.

Hunter, John, et al. *"Installation."* Matplotlib, 2019, https://matplotlib.org/.

"Hypertext Transfer Protocol (HTTP) Status Code Registry." Hypertext Transfer Protocol (HTTP) Status Code Registry, www.iana.org/assignments/http-status-codes/http-status-codes.xhtml.

Jinka, Preetam. *"Slow Queries? Move Fast to Fix Them."* Database Monitoring Tools from VividCortex, www.vividcortex.com/blog/slow-queries-move-fast-fix-them.

Kaushik, Avinash. "Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity." *O'Reilly* | *Safari*, Sybex, https://learning.oreilly.com/library/view/web-analytics-20/9780470529393/ch08.html.

Keohane, Dennis. *"David Blundin's Cogo Labs' Formula for Startup Success."* VentureFizz, 13 July 2018, https://venturefizz.com/stories/boston/david-blundins-cogo-labs-formula-startup-success.

Markus, Justas. *"What Is Conversion Funnel? - Learn How to Optimize Your Conversions."* Oberlo, Oberlo Dropshipping App., 30 Oct. 2019, www.oberlo.com/ecommerce-wiki/conversion-funnel.

"NetworkX." NetworkX, 2019, https://networkx.github.io/.

"Mission of NumFOCUS." NumFOCUS, 2019, https://numfocus.org/community/mission.

"NumPy." NumPy, 2019, https://numpy.org/.

"Python Data Analysis Library." Pandas, 2019, https://pandas.pydata.org/.

“Getting Started with Plotly.” Getting Started with Plotly | Python | Plotly, https://plot.ly/python/getting-started/.

Robinson, Edward, and Julie Verhage. *“Quicktake: Fintech.” Bloomberg.com, Bloomberg, https://www.bloomberg.com/quicktake/financial-technology-companies-disrupt-comfy-banks-quicktake.*

“Learn.” Scikit, https://scikit-learn.org/stable/.

Sheil, Humphrey, et al. *“Predicting Purchasing Intent: Automatic Feature Learning Using Recurrent Neural Networks.”* 21 July 2018, doi:arXiv.1807.08207.

“Sklearn.metrics.v_measure_score.” Scikit, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.v_measure_score.html.

Sraders, Anne. *“What Is Fintech? Uses and Examples in 2019.”* TheStreet, 8 Mar. 2019, https://www.thestreet.com/technology/what-is-fintech-14885154.

“Assumptions of Linear Regression.” Statistics Solutions, Statistics Solutions, 2019, https://www.statisticssolutions.com/assumptions-of-linear-regression/.

“Techniques for Improving the Performance of SQL Queries under Workspaces in the Data Service Layer.” IBM Knowledge Center, www.ibm.com/support/knowledgecenter/en/SSZLC2_9.0.0/com.ibm.commerce.developer.doc/refs/rsdperformanceworkspaces.htm.

“The Consumer Decision Journey.” McKinsey & Company, www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-consumer-decision-journey.

“URL Components Explained.” URL Components Explained - Tealium Learning Community, 18 Oct. 2016, community.tealiumiq.com/t5/iQ-Tag-Management/URL-Components-Explained/ta-p/5573.

“VC Vestigo Ventures.” Massinvestor Venture Capital and Private Equity Database, https://massinvestordatabase.com/publicfirm.php?name=Vestigo+Ventures.

“Vestigo Ventures Closes $58.9 Million Funding Round.” PR Newswire: Press Release Distribution, Targeting, Monitoring and Marketing, 23 Aug. 2018, https://www.prnewswire.com/news-releases/vestigo-ventures-closes-58-9-million-funding-round-300701332.html.

"Vestigo Ventures Investments." CB Insights, https://www.cbinsights.com/investor/vestigo-ventures-investments.

Vieira, Armando. *"Predicting Online User Behaviour Using Deep Learning Algorithms."* 27 May 2016, doi:arXiv.1511.06247.

Wang, Gang, et al. *"Unsupervised Clickstream Clustering for User Behavior Analysis."* Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI 16, 2016, doi:10.1145/2858036.2858107.

"What Is SEM? PPC & Paid Search Marketing Explained." Search Engine Land, https://searchengineland.com/guide/what-is-paid-search.

"What Is SEO / Search Engine Optimization?" Search Engine Land, 2019, https://searchengineland.com/guide/what-is-seo.

Wooldridge, Jeffrey M. *Introductory Econometrics: a Modern Approach.* Cengage Learning, 2016.

Yeung, Ka Yee and Walter L. Ruzzo. *"Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper"* An empirical study on Principal Component Analysis for clustering gene expression data " (to appear in Bioinformatics)." 2001.

Zider, Bob. "How Venture Capital Works." *Harvard Business Review*, 1 Aug. 2014, https://hbr.org/1998/11/how-venture-capital-works.