Online Video Data Analytics

Pierce Vollucci
Benjamin Le
Jefferson Lai
Wenxuan Cai
Yaohui Ye
George Necula, Ed.
Don Wroblewski, Ed.

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-71
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-71.html

May 13, 2015

Acknowledgement

We would like to thank George Necula for advising us throughout the entirety of this Capstone project. We would also like to thank Jibin Zhan from Conviva for introducing the problem space to us and Pat McDonough from Databricks for providing extensive technical support throughout our use of Databricks.

University of California, Berkeley College of Engineering

MASTER OF ENGINEERING - SPRING 2015

Electrical Engineering and Computer Sciences

Data Science and Systems

Online Video Data Analytics

Pierce Vollucci

This **Masters Project Paper** fulfills the Master of Engineering degree requirement.

Approved by:

1. Capstone Project Advisor:

Signature: ____________________________ Date _____________

Print Name/Department: George Necula/Electrical Engineering and Computer Sciences

2. Faculty Committee Member #2:

Signature: ____________________________ Date _____________

Print Name/Department: Don Wroblewski/Fung Institute for Engineering Leadership

Abstract

This capstone project report covers the research and development of Smart Anomaly Detection and Subscriber Analysis in the domain of Online Video Data Analytics. In the co-written portions of this document, we discuss the projected commercialization success of our products by analyzing worldwide trends in online video, presenting a competitive business strategy, and describing several approaches towards the management of our intellectual property. In the individually written portion of this document, we discuss our investigation into smart anomaly detection techniques as well as pursue experiments using nonseasonal data to evaluate the effectiveness of the most viable techniques.

Contents

* Co-written with Wenxuan Cai, Benjamin Le, Jefferson Lai, and Yaohui Ye

I. Introduction

This report documents the Online Video Data Analytics capstone project completed in the course of the Data Science and Systems focus of the Master of Engineering degree at UC Berkeley. Through the collective efforts of Benjamin Le, Jefferson Lai, Pierce Vollucci, Wenxuan Cai, and Yaohui Ye, our team has not only characterized the need for effective data analysis tools in the domain of online video data, but has also developed analysis tools which attempt to address this need. As we will describe in detail in our Individual Technical Contributions, our work has produced many important findings and we have made significant strides towards a complete implementation of these tools. However, at the time of the writing of this report, additional work is required before our tools can be considered complete. That being said, our substantial progress has allowed us to form a very clear vision of what our finished tools will look like and how they will perform. Our vision leads us to believe that, once finished, our tools can be of great potential value to entities within the online data analytics industry. In order to understand how best to cultivate this value, we have extended our vision to depict tools to marketable products and we have evaluated the potential for our team to establish a business offering these products. In doing so, we have performed extensive research of the current market and industry which our potential business would be entering. The remainder of this report presents our findings and is divided into seven sections. First, we introduce our industry partner, Conviva, in the Our Partner section. Second, we present the objective of our work and the motivation behind the resulting products in the Products and Value section. Third, we introduce and describe the dataset leveraged by our products in the Our Dataset section. Fourth, our team characterizes our industry as well as our competitive strategy in the Trends, Market, and Industry section. Fifth, in our Intellectual Property section, we describe how we plan to protect the value of our work. Sixth, the Individual Technical Contributions section of this report details our specific contributions toward the goals of our project. Finally, the Concluding Reflections section contains a retrospective analysis of the significance of this work and provides an outlook on the potential for continuation of our work in future endeavors.

II. Our Partner

This project is sponsored by Conviva, a leading online video quality analytics provider. Conviva works with video content providers, device manufacturers, and developers of video player libraries to gather video quality metrics from content consumers. Through our partnership with Conviva, we have access to an anonymized portion of their online video quality metric dataset for the development of our products. We also have access to Conviva engineers for collaboration purposes who provide domain knowledge and on site support. For the purpose of the business analysis forthcoming, the entity, "we", will refer to our capstone team as a separate entity from Conviva. Furthermore, we consider Conviva to be a close partner to our capstone team on whom we can rely for continuous access to their dataset.

III. Products and Value

A vast and painfully prevalent gap exists between the amount of data being generated around the world and the global tech industry's ability to utilize it. According to IBM, "every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone" ("Bringing Big Data"). While the already monumental quantity of data continues to grow, scientists and engineers alike are just beginning to tap into the power of this data. This is not to say that data does not already pervade nearly every imaginable aspect of life today; it does. Large amounts of data crunching and predictive analysis go on behind the scenes of numerous activities, from returning search queries, to recommending movies or restaurants, to predicting when and where the next earthquake will occur. However, there remains a massive body of questions and problems in both academia and industry that researchers have been unable to use data to answer. One domain in which better utilization of data could yield tremendous benefit is that of online media. Our team aims to serve this niche by building tools that address two critical challenges of online video data analysis: accurate real-time anomaly detection on large scale data and subscriber churn analysis.

Online video providers struggle to consistently serve a TV-like experience with high quality video free of buffering interruptions. Many factors within the "delivery ecosystem" affect the throughput of a video stream and, ultimately, the end user's viewing experience (Ganjam et.al 8). These factors include "multiple encoder formats and profiles, CDNs, ISPs, devices, and a plethora of streaming protocols and video players" (Ganjam et al. 8). An automatic anomaly detection and alert system is necessary in order to both inform a video content provider when their customers are experiencing low quality and, among the many possible factors, diagnose the primary cause of the problem. For example, if all customers experiencing frequent buffering belong to a certain ISP, then the alert system should flag that ISP as the root of the problem. The challenge that plagues many current solutions, however, is related to the aforementioned growth in the amount of collected data. While it is easy to detect when and why predictable measurements misbehave at small scale, it is hard to do so with high accuracy at large scale, across a range of system environments. To meet this challenge, our team has developed our Smart Anomaly Detection system to detect when and why truly anomalous and interesting behavior occurs in measured data. Such a system would greatly help content providers improve both their operational performance and efficiency. This value will be passed down to the viewers who benefit from higher service quality.

A second problem for subscription-based online video content providers is the ability to retain their subscribers. While the problem of diagnosing and eventually reducing subscriber churn has existed as long as the subscription service model, only recently has the tech industry developed the capacity and means to use big data to do so (Keaveny). Furthermore, largely due to the fact that online video hosting and distribution is a relatively new service, nearly all previous works in the area have focused on other domains such as telecommunications or television service subscriptions (Keaveny; Verbeke 2357-2358). Our team's Subscriber Analysis toolset aims to fulfill this unmet need by developing predictive models of viewer engagement and churn based on viewing activity and service quality data. Being able to predict churners and identify characteristic predictors from the data allows companies to focus on addressing the problems most critical to their viewers, thereby reducing churn rates. As proven by Zeithaml, there is a real, high cost associated with subscriber churn (Zeithaml). Thus by aiding in the reduction of churn, our Subscriber Analysis product can both help content providers increase revenues and result in higher overall satisfaction for those who purchase online video subscriptions.

For the reasons described above, our team is confident that our Smart Anomaly Detection and Subscriber Analysis products are important and valuable to both content providers and their customers, the viewers.

IV. Our Dataset

Conviva provided 4.5 months of session summary data from a single anonymous content provider for our research and development. 73,368,052 rows of session summaries are in this dataset. Each session summary represents a single instance of a viewer requesting a video object. In addition to service quality data, the type of device used by the viewer, the approximate location of the viewer, and metadata about the video content being accessed are collected into 45 columns. Subscription and demographic information about a viewer beyond their location are not available within this dataset. Fields that might otherwise help identify the anonymous content provider such as video content metadata were also anonymized by Conviva prior to data transfer to protect their customer.

Although the data was preformatted by Conviva before being transferred to our capstone team, we identified two important challenges implicitly encoded within this data through exploratory data analysis and follow-up communication with Conviva engineers. First, several of the fields in the session summaries are not as reliable as we initially believed. For example, fields such as `season` and `episodeName` are often empty. Second, our initial dataset included data generated by an artificial "viewer" that was used by Conviva for testing purposes and exhibited very strange, abnormal behavior. This was very important to keep in mind as we developed and evaluated our tools based on this data.

For our Smart Anomaly Detection product, Conviva informed us of the two most important metrics in assessing QoS that they wished to detect anomalies for. First is the number of attempts in watching online video over a time period. Low number of attempts indicates that users may be unable to access the content due to a datacenter failure. A high number of attempts signals the presence of a viral video. Second is the video start failure (VSF) rate. The VSF rate is the percentage of attempts that have failed to begin properly. VSFs may be caused by bugs in the video player software or by improper encoding/decoding of video content. Unlike attempts, low VSF rate is not a concern for video content providers. However, high VSF rate indicates major issues in the content delivery pipeline. To determine if an attempt has ended in VSF, we look at the `joinTimeMs` and `nrerrorsbeforejoin` columns in the data. The table below provides description about these 2 columns. An attempt ends in VSF if `joinTimeMs < 0 AND nrerrorsbeforejoin > 0`.

Column Name	**DataType**	**Description**
joinTimeMs	int	How long this attempt spent joined with the video stream. If this attempt has not yet joined, then this value will default to -1.
nrerrorsbeforejoin	int	How many fatal errors occurred before video join

For Subscriber Analysis, on the other hand, the nature of the problem is that we cannot know beforehand which fields within the session summary are useful in distinguishing viewers who are likely to churn. At the same time, as a consequence of the first of the challenges mentioned above, the Subscriber Analysis product should not indiscriminately use all fields of the session summary, including both reliable and unreliable fields. Thus, a central component of the work in Subscriber Analysis revolves around selecting a subset of these fields to use to form “features” to be used by the product.

V. Trends and Strategy

Having defined our team’s product and established both how they generate value and for whom they are valuable, we can focus on how we plan to bring these products to market from the standpoint of a new business. Amidst an era of rapid information and especially within the technology-abundant Silicon Valley, bringing such innovations to market requires understanding the market and having a well-formed competitive strategy. In this section, we describe the social and technological trends relevant to our product as well as the market and industry our business would be entering. We then describe the strategy we have developed that would allow our business to be successful in this competitive environment.

Why Us, Why Now

In the past five years, the number of broadband internet connections in the United States has grown from 124 million in 2009 to 306 million in 2014, leading to a compound annual growth rate of 19.8% per year (“Num. of Broadband Conns.”). This growth is indicative of the ever-growing role the Internet plays in daily life. Along with the growth of the Internet, as both a cause and effect, comes the spread of online services. In her article for Forbes, Erika Trautman, CEO of Rapt Media, states that “each year, more and more people are ditching cable and are opting for online services like Netflix and Hulu.”

The emergence of online video services has been so disruptive a shift in video distribution, that it incited a 2012 public hearing concerning public policies from the Senate Committee on Commerce, Science, and Transportation. In the hearing, leaders from technology juggernauts and state senators alike echoed the same viewpoint: online video services are the future of video distribution. Susan D. White, the Vice Chair for Nielsen, a leading global information and measurement company, reported that "the use of video on PCs continues to increase—up 80 percent in the last 4 years…Consumers are saying, unequivocally, that online video will continue to play an increasing role in their media choices" (U.S. Sen. Comm. on Commerce, Sci. & Trans. 9).

Of course, similar to other industries, a business seeking to enter today's online video industry must meet a myriad of both business and engineering challenges. Unlike many of these industries, however, our industry is well-positioned to easily collect and analyze vast amounts of data to meet these challenges. Out of these conditions, the online video analytics (OVA) industry emerged, helping to translate and transform this data into useful insights that can be directly used by online video providers. A report by Frost & Sullivan summarized the rapid growth in the market:

> Still a largely nascent market, online video analytics (OVA) earned $174.7 million in revenue in 2013. It is projected to reach $472 million in 2020 as it observes a compound annual growth rate (CAGR) of 15.3%....The growth of OVA is largely attributed to the high demand for advanced analytics from online video consumption (Jasani).

Spurred by the massive opportunity in this market, our team has worked with Conviva to identify two of the most significant technical challenges faced by content providers: real-time detection of anomalies in a rapidly changing, unpredictable environment and efficiently reducing subscriber churn.

The challenge of retaining subscribers has existed as long as the subscription-based business model itself. As the competitive landscape of the online video market continues to evolve, the ability to diagnose and mitigate subscriber churn is a crucial component for business success. Sanford C. Bernstein estimated Netflix's average annual churn rate at 40-50%, which translates to 24-30 million subscribers

(Gottfried). Reducing this churn rate by even a small fraction and keeping the business of these subscribers could mean significant increases in revenue. Just as critical for success is the ability to detect and respond to anomalies or important changes in metrics such as network usage and resource utilization. On July 24, 2007, 18 hours of Netflix downtime corresponded with a 7% plummet in the company's stock (Associated Press).

As previously described, our team provides solutions to these challenges through our Subscriber Analysis and Smart Anomaly Detection products. We believe that while these solutions, which use a combination of statistical and machine learning techniques, are powerful, our primary value and competitive advantage lies in our use of the unique dataset available to us through our partnership with Conviva. In the following sections, we discuss in detail how we plan to establish ourselves within the industry. In particular, we describe how we will position ourselves towards our buyers and suppliers as well as how we will respond to potential new entrants and existing competitors to the market.

Buyers and Suppliers

One of the most important components of a successful business strategy is a deep and accurate understanding the different players involved in the industry. In particular, an effective strategy must define the industry's buyers, to whom businesses sell their product, and its suppliers, from whom businesses purchase resources. In this section, we provide an overview of important entities related to our industry and present an analysis of our buyers and suppliers.

Potential customers for the global online video analytics market include content providers, who own video content, and service providers, who facilitate the sharing of user-generated video content (Jasani; Smith). Among the content providers are companies such as HBO, CCTV, and Disney, who all bring a variety of original video content to market every year. These businesses serve a huge user base and are able to accumulate large amounts of subscriber data. HBO alone was reported to have over 30 million users at the beginning of 2014 (Lawler). This abundance of data presents massive potential for improving these companies' product quality and, correspondingly, market share. Our Subscriber Analysis product can realize some of this potential by helping understand the experience and behavior of their users. Furthermore, with our Smart Anomaly Detection product, content providers can be made aware when significant changes occur in viewer behavior, system performance, or both. These tools can lead to a more valuable product, as seen from the content provider's viewers. While service providers such as Twitch or Vimeo differ from content providers in that they tend to offer free services, the success of these companies is still highly dependent on retaining a large number of active users. Thus, we target service providers in much the same way as we target content providers. Overall, we find that content and service providers, as buyers, are at an advantage in terms of business leverage over us, as sellers. This is primarily due to low switching costs, which arise from the fact that other businesses such as Akamai and Ooyala offer products for processing video data similar to ours (Roettgers). Because buyers ultimately make the choice choosing where to send their data on which both Smart Anomaly Detection and Subscriber Analysis depend, it can be difficult to deter customers from switching to our competitors. However, as we describe later in this paper, our unique approach towards churn analysis may differentiate us from our competitors and decrease buyer leverage over us.

On the other end of the supply chain, we also must consider who our suppliers will be and what type of business relationship we will have with them. Because our product exists exclusively as software, we require computing power and data storage capacity. Both of these can be obtained through the purchase of cloud services. Fortunately, the current trends indicate that cloud services are becoming commoditized, with many vendors such as Amazon, IBM, Google and Microsoft offering very similar products (F. Hanley). Though our buyers benefitted from low switching costs between us and our competitors, we face even lower switching costs between our suppliers. This is because while there is a considerable amount of effort involved with integrating a monitoring or analytical system with a new set of data, migrating the services between the machines which host them is almost trivial, involving only a transfer the data and minor machine configuration. In addition to cloud services, to a certain extent, we are dependent on device manufacturers and developers of video player libraries. We require them to provide an Application Program Interface (API) which we can use to gather online video analytics data from users. Fortunately, prior relationships with these device manufacturers and developers have been established through our partner Conviva. Conviva can help us open APIs for new devices and video players to maintain the flow of data required for our products.

As Porter argued, strategic positioning requires performing activities either differently or more efficiently than rivals (“Five Competitive Forces” 11). Our partnership with Conviva affords us a large quantity of high quality data for our algorithms to utilize, giving us a slight advantage compared to other services. In order to maintain and build upon this advantage, however, we must focus on developing our products to utilize this data and yield results in a superior manner. Thus, it is clear that our ability to differentiate from competing products and outperform them is key to our business strategy and the following sections describe how we can do so.

New Entrants

“Know yourself and know your enemy, and you will never be defeated” (Sun Tzu 18). This proverb can be applied to almost any competitive situation, from warfare to marketing. Interpreting this teaching in the context of business strategy, we identify that understanding the rivalry among existing and potential competitors is essential to a lasting competitive advantage. This interpretation fits well within the framework of Michael Porter’s five competitive forces. We now examine new entrants through the incumbent advantages and barriers to entry that work to keep this force as a low threat to both of our products. Porter recognized seven incumbent advantages (“Five Competitive Forces” 4-6). The first is supply side economies of scale in which established incumbents have tremendous strength. The code behind a given analysis program is a fixed cost which scales well with an increased number of users, thus reducing the marginal cost of the code with each customer. The servers that receive and process the various users' data are linear, but scale with the number of customers acquired. The real advantage comes from the exponential power of the data supplied by these same customers, a theme we have come back to repeatedly in this paper. As the breadth and quantity of data increase with the combined user base of our customers, our algorithms become increasingly powerful and allow the incumbent product to outperform new entrants. This leads into our second advantage, demand side benefits of scale. As the authority in the field of providing content providers with analytics, incumbents can encourage customer demand by using their data on content quality improvements to provide hard evidence of the bottom line improvement new users can expect. "Increasingly powerful predictive analytics tools will unlock business insights [and drive revenue]" (Kahn 5). Demonstrating that our tools provide access to increases in revenue is key to nurturing demand.

Switching from an incumbent's service provides another barrier to entry, customer switching costs. While switching from one online service to another is not prohibitively expensive considering the benefits offered, the most impacting loss is in the past data the incumbent analysis provider's algorithms had of user's performance. "As we increase the training set size L we train on more and more patterns so the test error declines" (Cortes et al. 241). Via additional training examples, the incumbent's algorithm would consistently outperform the new entrant as the new entrant slowly acquires a pool of data comparable to that of the incumbent.

Just as it does not appear expensive for a customer to switch, it appears feasible for new entrant to join due to minimal physical capital requirements. With Platform as a Service (PaaS) providers, a new entrant merely needs a codified algorithm and a client or two to get started. Still, it is again the data that proves key to providing value to our customers. Importantly, new entrants cannot attain this data until they acquire clients, a classic catch-22 which serves as an inhibiting capital requirement for new entrants.

The global reach of our data partner, Conviva, provides both a size independent advantage as well as an unequal access to potential distribution channels in that it allows for direct international sales in the form of immediate integration of our tools with the systems of our partner's customers. The last relevant advantage as discussed by Porter, concerns restrictive government policy. Privacy concerns do arise when personal data is used, however there are standards for anonymization to be employed when using such data (Iyengar). While governments do allow the use of such data, it has to be acquired by legal means, which means a new entrant is restricted in its means of gathering new data for its algorithms. Thus, after a thorough analysis of the potential new entrants of our industry, the incumbents' advantages suggest that the threat of new entrants is a relatively weak force in our industry.

Existing Rivals

Another category of threats that a successful business strategy must address is that of existing rivals. As Porter described, the degree to which rivalry drives down an industry's profit potential depends firstly on the intensity with which companies compete and secondly on the basis on which they compete ("Five Competitive Forces" 10). We analyze these two parts for each of our products separately.

As machine learning grows in popularity, research into anomaly detection and other analyses of time series data is receiving greater attention both in academia and in industry. A survey of anomaly detection techniques shows a variety of techniques applied in a diverse range of domains (Chandola). Our strategy must take into account the threat of commercialization of technologies into industry competitors. For example, in 1994 Dipankar Dasgupta used a negative selection mechanism of the immune system to develop a "novelty" detection algorithm (Dasgupta). In addition to these potential competitors, there already exist several important industrial competitors working on anomaly detection. In January, 2015, Twitter open-sourced *AnomalyDetection*, a software package that automatically detects anomalies in big data in a practical and robust way (Kejariwal). Our Smart Anomaly Detection product is comparable to products from industry competitors such as Twitter; it is able to integrate with various sources of data, perform real-time processing, and incorporate smart thresholding with alerts. Although our competitors may try to research and develop a superior anomaly detection algorithm, we believe that our superior quantity and quality of data provided by Conviva gives us an edge over our competitors. Thus, we characterize competitive risk for Smart Anomaly Detection as weak. To a large extent, the competitors of Subscriber Analysis include the content providers themselves. Netflix spends $150 million on improving content recommendation each year, with the justification that improving recommendations and subscriber retention by even a small amount can lead to significant increases in revenue (Roettgers). These content providers have the advantage that they have complete access and control over the data they collect. If most companies were able to build an effective churn predictor in-house, the industry would be in trouble. However, we are confident that the quality of our Subscriber Analysis product will overwhelmingly convince content providers facing the classic “buy versus build” question, that building a product of similar quality would demand significantly more resources than simply purchasing from us (Cohn). This confidence is further supported by Porter in the context of the tradeoffs of strategic positioning (“What Is Strategy?” 4-11). In addition to content providers, there also exist competitors such as Akamai and Ooyala, who offer standalone analysis products to content and service providers. These competitors tend to focus on the monitoring and visualization of the data. In contrast, Subscriber Analysis focuses on performing the actual analysis to identify the characteristics and causes of subscriber churn.

Still, our most important advantage over these competitors remains our ability to perform in-depth churn analyses based on the abstraction of session summaries, which consist of a unique combination of metrics exclusively related to service quality. To the best of our knowledge, this is unique to previous and existing works in subscriber churn analyses. Our research has shown that the most prominent existing analysis approaches all incorporate a significant amount of information, often involving direct customer surveys or other self-reported data. Because service quality data is abundant and easy to obtain compared with demographic data, our Subscriber Analysis product can appear extremely appealing to potential customers. This easy to collect and consistent subset of video consumption data means our product has the potential to scale much better than existing approaches which require highly detailed, case-specific, and hard to obtain datasets. However, we cannot guarantee that this algorithmic advantage be sustained as our competitors continue their own research and development. Thus, we conclude that threat of competition to Subscriber Analysis is moderate.

Substitutes

The final element of our marketing strategy concerns the threat of new substitutes. Porter defined substitutes as products that serve the same purpose as the product in question but through different means (“Five Competitive Forces” 11). We first discuss potential substitutes for our Smart Anomaly Detection product.

The gold standard for most alert systems is human monitoring. Analogous to firms hiring security monitors to watch over buildings, video content providers can hire administrators to keep watch over network health. A more automated substitute is achieved through simple thresholding, in which hardcoded thresholds for metrics such as the rate of video failures trigger an alarm when exceeded. Content providers can also utilize third party network performance management software from leaders like CA, Inc. This type of software alerts IT departments of potential performance degradation within the companies' internal networks (CA Inc. 4). Similarly, content providers can pursue avenues besides Subscriber Analysis to reduce churn rates. Examples include utilizing feedback surveys and consulting expert market analysts. Feedback from unsubscribers is an extremely popular source of insight into why customers choose to leave and can go a long way in improving the product and reducing churn rate. These often take the form of questionnaires conducted on the company’s website or through email. In addition, content providers commonly devote many resources towards consulting individuals or even entire departments with the goal of identifying marketing approaches or market segments that generate lower churn rates.

Porter classified a substitute as a high threat when the substitute offers superior price/performance (“Five Competitive Forces” 12). With this in mind, we found that the overall threat of substitutes for Smart Anomaly Detection product is low. In contrast to human monitoring, our product offers a superior value proposition to our buyer. According to Ganjam et. al, many factors, including "multiple encoder formats and profiles, CDNs, ISPs, devices, and a plethora of streaming protocols and video players," affect the end user’s viewing experience (Ganjam 8). The complexity of this delivery ecosystem requires equally complex monitoring with filters to isolate a specific ISP, for example, and to determine if its behavior is anomalous. Such large scale monitoring does not scale efficiently when using just human monitoring. Similarly, simple thresholding poses little threat as a substitute because fine tuning proper thresholds over multiple data streams is difficult and time consuming. Many false positives and negatives still occur, despite such fine tuning (Numenta 11). Network performance management software, on the other hand, poses a considerable threat to us. However, while they are excellent at detecting problems within a content provider’s internal network, they alone cannot increase the quality of service. Xi Liu et al. argue that an optimal viewing experience requires a coordinated video control plane with a “global view of client and network conditions” (Liu 1). Fortunately, thanks to our partnership with Conviva, we have the data necessary to obtain this global view.

Just as with Smart Anomaly Detection, the threat of substitutes for Subscriber Analysis is also low. Although feedback surveys are direct and easy to implement, there are several inherent issues associated with them. Perhaps most prominently, any analysis that uses this data format must make a large number of assumptions in order to deal with uncontrollable factors such as non-response bias and self-report bias (Keaveny). Expert opinion, whether gathered from a department with the company or through external consult, is the traditional and most common approach towards combating subscriber churn. This method, while very effective, tends to be extremely expensive. Still, as demonstrated by Mcgovern’s Virgin Mobile case study, expert opinion can lead to identifying the right market segment, lower churn rates, and ultimately a successful business (McGovern 9).

To mitigate the threat of substitutes, Porter suggests offering “better value through new features or wider product accessibility” (“Five Competitive Forces” 16). For Smart Anomaly Detection, there are several avenues to pursue to provide a better value proposition to our buyers. For example, we can develop more accurate predictors with additional data from Conviva and explore new machine learning algorithms. For Subscriber Analysis, the threat of substitutes continues to be low because, unlike the examples given above, our product can perform effective analyses and generate valuable insights in an automated, efficient fashion. Data obtained through direct customer surveys, while potentially cheap, come bundled numerous disclaimers and can lead to a certain stigma from the subscriber’s perspective. Furthermore, although data obtained through surveys, such as demographic information, might be more helpful in characterizing churners, by focusing on providing churn analysis based only on service quality data, our Subscriber Analysis product has at least one significant advantage. Service quality data from content consumers can be more easily gathered compared to data such as demographic information. Consequently, our product can be more appealing and accessible to content providers, especially those who do not have access to, or would like to avoid the cost of obtaining, personal data about their users. We also point out that both Subscriber Analysis and the substitutes such as those described above can be used in combination with each other. In such a case, our Subscriber Analysis product becomes even more appealing. This is because it can use the data from customer feedback to yield further improved performance. Our product would also make tasks such as identifying appropriate market segments much easier and cheaper to accomplish for content providers.

Strategy Summary

In summary, there are several social and technological trends which make now the right time for commercializing our Subscriber Analysis and Smart Anomaly

Detection products. The most prominent among these are the rapid growth in internet connectivity and the spread of online services. In order to evaluate how well positioned we are to capitalize on the opportunity created by these trends, we developed a business strategy through competitive industry and market analysis from several different perspectives. From the perspective of buyers and suppliers, though we find that buyer power is significant, over time we expect to differentiate ourselves from our competitors by leveraging both the superior size of our dataset and our more efficient overall use of the data. We find that supplier power is low for our industry because the only significant resource we require is available through cloud services, an industry in which we have high buyer power and which is quickly becoming commoditized. From the perspective of rivals, the threat of new entrants is low due in large part to the superior quantity and quality of our data as well as the benefits of scale we would stand to benefit from as incumbents. Similarly, while existing competitors do present a threat, we find that our use of superior data and unique approach gives us a significant competitive advantage over them. Finally, we see a weak threat from the perspective of substitutes because we offer superior value at a cheaper price to our customers that only improves in combination with other techniques. Taken together, our evaluations lead us to believe that there is significant potential for a sustained competitive advantage over competitors, and that now is an opportune time to pursue it.

VI. Intellectual Property

Equally important to a team's ability to build a valuable product and bring it to market is its ability to protect that value. In this section, we explain how we, as a business pursuing the strategy above to bring Subscriber Analysis and Smart Anomaly Detection to market, intend to sustain and protect the value of our work.

The traditional method for protecting the value of a new technology or innovation is obtaining a legal statement regarding ownership of intellectual property, IP, in the form of a patent. Indeed, patents have performed well enough to remain a primary mechanism for IP protection in the US for more than 200 years (Fisher). Unfortunately, when it comes to software, the rules and regulations regarding patents become dangerously ambiguous. The recent influx of lawsuits involving software patents has been attributed to the issuance of patents that are unclear, overly broad, or both (Bessen). Despite software patent laws being an active and controversial topic, these discussions have simply left more questions unanswered. The *Alice Corporation v. CLS Bank* Supreme Court case in 2013 is oft cited as the first source of information about software patentability, and even this case has been criticized for the court's vagueness (*Alice Corporation v. CLS Bank*). As noted by patent attorney and founder of IPWatchDog.com Gene Quinn, a definitive line should be drawn by the courts: a patent describing only an abstract idea, without specific implementation details, is invalid and cannot be acted upon (Quinn).

Thus, faced with the question of patentability, our team must examine the novelty of our Subscriber Analysis and Smart Anomaly Detection products. The goals of Subscriber Analysis and Smart Anomaly Detection are to diagnose the causes of subscriber churn and intelligently detect important changes in measured data respectively. Because these goals are rather broad, there exist a number of existing implementations, both old and new, with similar objectives. As a team considering patentability, we look towards the novelty of our specific approach and implementation. In the course of this introspection, we note that our implementation amalgamates open source machine learning libraries such as SciKit-Learn, published research from both industry and academia, programming tools such as those offered by Databricks, and finally the unique data afforded to us through our partnership with Conviva. With this in mind, we conclude that current patenting processes are flexible enough such that by defining our implementations at an extremely fine granularity, we would likely be able to obtain a patent on our software. However, we strongly believe that there exist several significant and compelling reasons against attempting to obtain a patent for our work. In this section, we elaborate on these reasons and describe an alternative method for protecting our IP which better suits our situation and business goals.

There is an abundance of existing anomaly detection patents of which we must be wary. Several of these patents are held by some of the largest companies in the technology sector, including Amazon and IBM. For example, *Detecting anomalies in Time Series Data*, owned by Amazon, states that it covers “The detected one anomaly, the assigned magnitude, and the correlated at least one external event are reported to a client device” (U.S. Patent 8,949,677). One patent owned by IBM, *Detecting anomalies in real-time in multiple time series data with automated thresholding*, states that in the submitted algorithm, a “comparison score” is calculated by comparing “the first series of [observed] normalized values” with “the second series of [predicted] normalized values” (U.S. Patent 8,924,333). In observance of these patents, we must be wary of litigation, especially when it concerns large technology companies. Recently, many companies in the tech industry, both small and large, have come under fire with a disproportionate number of patent infringement lawsuits (Byrd and Howard 8). Some optimists argue that most companies need not worry, because large technology companies are likely filing patents defensively. However, these companies are often the ones who play prosecutor in these patent infringement cases as well. For example, IBM, a holder of one of these anomaly detection patents, has a history of suing startups prior to their initial public offerings (Etherington). More recently, Twitter settled a patent infringement lawsuit with IBM by purchasing 900 of IBM’s patents (Etherington). In a calculated move by IBM, Twitter felt pressured to settle to protect their stock price in preparation for their IPO. Thus, we must be extremely careful in how we choose to protect our intellectual property. If this means filing a patent, then we must be prepared to use it defensively. This is likely to require a very large amount of financial resources. As we do not currently have these resources to spare and cannot guarantee that the protection offered would be long lasting or enforceable, we seek an alternative to patenting.

The goal of our Subscriber Analysis product is to predict the future subscription status of users based on past viewing behavior. Despite our research on existing patents, our team has been unable to find many patents which pose a legal threat to

Subscriber Analysis. Most active patents on video analytics focus on video performance and forecast, such as Blue Kai Inc's *Real time audience forecasting* (US Patent App. 20120047005). In contrast, the patent field of quantization and prediction of subscriber behavior remains largely unexplored. Despite several commercial solutions on the market, there has not been a corresponding number of patents. Thus, Subscriber Analysis does not face the same level of risk of litigation compared to Smart Anomaly Detection. However, there are a handful of patents in other domains that we need to be wary of. *System and method for measuring television audience engagement*, owned by Rentrak corporation, describes a system that measures audience engagement based on the time he or she spends on the program (US Patent 8,904,419). In short, it constructs a viewership regression curve for different video content and measures the average viewing length. For a new video, the algorithm infers the level of viewer engagement based on the video content and the duration the viewer watched. While viewer engagement is a critical component for predicting behavior in Subscriber Analysis, we also incorporate additional data. These include viewing frequency, content type, and video quality. Under such circumstances, we do not see it as necessary to license patents such as the one above for two reasons. First, and perhaps most importantly, we apply churn analysis in the domain of online video, whereas most relevant patents apply to other older domains. Second, our algorithm incorporates a unique set of features corresponding to the data provided by Conviva.

The decision to pursue and rely on a patent in the software is an expensive one in both time and financial resources as well as a risky one due to the tumultuous software patent environment. As such, while we may pursue a patent, it will not be relied upon for our business model. As such, we have two additional IP strategies to investigate, open sourcing and copyrighting.

Open source software is software that can be freely used, changed, and shared (in modified and unmodified form) by anyone, subject to some moderation (Open Source Initiative). Open sourcing has become increasingly popular; both the total amount of open source code and the number of open source projects are growing at an exponential rate (Deshpande, Amit et al). For the purposes of our endeavor, it is not the novelty of our approach but our dataset and partner provided distribution network that distinguishes us. As the algorithms used are already publicly available, open sourcing our code does not cost us anything but provides us the shield of using open source software for our business and the badge having our code publically exposed and subject to peer review. Our business model would entail providing a value-added service company, dedicated to helping customers integrate their existing systems with our anomaly detection library. Through our partnership with Conviva, we have an established distribution network to our potential customers who we can offer immediate integration with Conviva's existing platform. This is a significant advantage as while open source is openly available to all users, they are primarily for experienced users. Users have to perform a significant amount of configuration before they begin using the code, which can pose quite a deterrent. While we will use the open source codebase as a foundation for our service, we will additionally provide full technical support in designing a customized solution that meets the customer's needs. By pivoting towards this direction, we add additional monetary value to the product that we can sell and bridge the technical gap for unexperienced users, relying on a SAAS implementation style for our business model instead of on a patent.

Copyright for software provides another IP Strategy option. While debate continues to surround software patents, copyrights are heavily applied in software. As expressed by Forbes's Tim Worstall, “there's no doubt that code is copyright anyway. It's a specific expression of an idea and so is copyright.” There are several differences in the protection offered by copyrights compared to that of patents. While a patent may expose a very specific invention or process to the public and protect for 20 years, a copyright offers much broader protection while still providing the threat of lawsuit for enforcement. The copyright lasts 90 years past the death of the author and offers statutory damages (Copyright.gov). In addition, the scope of what it encompasses proves more relevant to our endeavor. “Multiple aspects of software can qualify for copyright protection: the source code, the compiled code, the visual layout, the documentation, possibly even the aggregation of menu commands” (Goldman). By protecting the numerous aspects of our project, copyright provides us adequate security. Besides the advantages of the protection offered, the process is affordable and efficient. Copyright is automatic as soon as a work is completed, though to file for statutory damages, one must formally register for a fee of less than $100 and an application turnaround time of under a year (Copyright.gov). In addition, even prior to completion of the work, we can preregister with a detailed explanation of the work in progress.

All IP strategies come with risks and copyright is no different. While pursuing a strategy of trade secrets would make our code more private, we would risk losing our protection should the secret be compromised. Also, as a general security principle in the computer science field, only the bare minimum should be relied upon to be kept secret to minimize risk of loss. However, completely publicizing our code for our copyright can be equally dangerous as the competition could copy our code with only slight rewrites. To remedy this, we can limit access to the raw code and only publish the required first and last 25 pages of code needed to attain a copyright on the entire work. In addition to this measure, it is our unique dataset that is the source of our code’s advantage over our competitors, and this is already protected by our partner, Conviva, in its aggregated form as a trade secret,

We believe that the novelty of our code and the application of our techniques to our unique dataset would allow us to obtain a software patent. However, while a patent may be most effective at reducing our risk of litigation, we look to alternatives due to the current complexity of filing a software patent and the immense amount of financial resources required to do so. Our research has led us to two very appealing alternatives: open sourcing and copyrighting. For the reasons stated above, we believe that while each of these alternatives have their own risks, their respective merits make them more appropriate for our use than patenting. Moving forward, we plan to employ open sourcing, as we expect that building a large, open community of support will encourage adoption and most benefit our products.

VII. Technical Contributions

Introduction

With the expanding globalization of online services and as “Big Data” becomes an increasingly popular buzzword among companies, there exists a prevalent need for efficient methods for harnessing and utilizing this influx of information. Manually processing data points, such as parsing through the feedback surveys from every client or monitoring a constant stream of logs from all users, has become an overwhelming task for any analytics team. In response, many companies rely on manually monitoring aggregations of this data. Beyond the human limitations of requiring domain knowledge and constant monitoring, these aggregations can lose more intricate, yet significant, relationships within the data. To address these concerns, companies such as Netflix and Google have implemented machine learning techniques into data analysis to provide automatic analysis, including features such as general anomaly detection and content recommendation. Our team examined a specialized domain of online data analysis, online videos. In this paper, we will examine the problem of anomaly detection as well as current endeavors into this field, followed by a discussion of our own primary research, employed techniques, and results towards this end. For this dominant form of online content, we aimed to create a toolset providing both Subscriber Analysis and Smart Anomaly Detection features to process the overwhelming amounts of data produced by a quarter of all internet users watching an online video every day (Statista).

Project Overview

Our endeavor was dual faceted: to make a set of tools to facilitate subscriber analysis and to create tools for automatic anomaly detection. Such tools can be critical to a website’s or service’s success by helping the developers with “Attracting prospects or site visitors; Engaging with the visitors in a bid to drive conversions; Tracking the engagement and conversion performance; Retaining customers for repeat sales" (Campbell). Impacting such critical stages in evaluating and improving the user experience online, data analysis can prove an indispensable tool to online services. However, this goes beyond simply collecting data but processing it in a meaningful way. "'Why do we need another measurement tool in our business?' the most common fear is data overload— collecting more information just because you can inevitably leads to more confusion, not clarity" (Clifton 4). To avoid overburdening those responsible for analysis while also to avoid letting the benefits of such data go to waste, our team divided into two separate subteams to build automatic analytical tools. One team focused on user engagement and designing techniques for minimizing churn rate in the form of Subscriber Analysis. My team focused on Smart Anomaly Detection for both seasonal and nonseasonal online video metrics.

Smart Anomaly Detection

Smart Anomaly Detection entails a system which can accurately inform an administrator when an error occurs or what the error entails or both. To remedy an anomaly, an administrator must spend time in the following steps: detecting the issue, diagnosing the issue, and solving the issue. Our endeavors aimed to minimize the duration of the first two. To this end, we first needed to define what constitutes an anomalous data point versus a non-anomalous data point. Then we had to devise an effective model to provide fast, high accuracy classification between the two. Based on our contextual inquiry, an anomaly is an extreme point, or set of points, outside of some tunable threshold of an expected value for that period. Metrics, such as views on a specific device, which are most indicative of a failure, such as in the form of a CDN failing or a version error, can be monitored in intervals small enough to compromise between promptness of diagnosis and accuracy. To begin our study of anomaly detection, we first begin with the results of our contextual inquiry as well as with an examination of one of the most widely used online analytics tools, Google Analytics.

Knowledge Domain

Google Analytics

Google Analytics provides a free platform for web services to monitor various performance and user metrics. In his text, *Advanced Web Metrics with Google Analytics*, Brian Clifton explains the service, as well as details usage of its various components. He describes the purposes of analysis through the AMAT acronym: “acquisition of visitors, measurement of performance, analysis of trends, testing to improve” (Clifton 11). By recording and monitoring various metrics, such as where a given user visits and for what duration, Google Analytics can provide developers with information to model regular user behavior on their site. With these metrics, various user analytics, including user archetypes, can be produced such as a bounced visitor, “A visitor who views only a single page on your website and has no further actions. This is generally considered a bad experience,” or a user session, “Also referred to as a ‘visit’ or ‘visitor session,’ this is the period of interaction a visitor has with your website” (Clifton 6). These archetypes are useful concepts in developing a website are are directly translatable and prevalent into the domain of online videos in which a bounced visitor may start a video or a service and then leave or in which a user has a viewing session of different videos. Google Analytics monitors a number of analytics to produce these archetypes; the twelve dimensions most prevalent to anomaly detection are: “All traffic, Visitor type (new or returning visitor), City, Region, Country/territory, Campaign, Keyword, Source, Medium, Referral path, Landing page, Exit page” (Clifton 101). These metrics are then provided to the user in the form of an hourly report usually in aggregated form (Clifton 35). To avoid overwhelming the developer with a barrage of incoming metrics, Google Analytics instead prunes and aggregates what it believes are the most important “first-level” metrics. “Examples of first-level metrics include: number of daily visitors, average visit time, geographic distribution, top-visited pages” (Clifton 8). These can be presented in the form of numbers or graphs, but require a human observer not only to be reading the reports, but to have general domain knowledge to recognize relevant patterns.

Whereas some patterns are expected to be gleaned from a human's interpretation of presented metrics, such as recognizing specific site content that delays a user's progress towards conversion, anomalies are automatically detected and delivered in the form of Google Analytics' Intelligence Reports.

> Intelligence provides automatic alerts for significant changes in data patterns from your website. Instead of you having to monitor reports and comb through data, Analytics intelligence alerts you to the most significant information to pay attention to. In addition, you can also create custom alerts and have an email sent to you when this is triggered. For example, intelligence can automatically highlight a 200 percent surge in visits from twitter last monday or let you know bounce rates of visitors from the U.S. dropped by 70 percent yesterday. (Clifton 53)

This seems to be the ideal case for addressing the previous mantra of information overload with big data. Rather than requiring a human to constantly monitor aggregate values and then recognize, via domain knowledge, when an anomaly occurs, a computer can provide automated monitoring and inform the user of an anomaly. However, this only suffices if the automated detection is sufficiently accurate to warrant replacing a human observer. For this we must examine the intricacies of these Intelligence Reports.

> Intelligence Overview: intelligence works by performing statistical analysis on your previous data patterns. Assuming you have reasonable levels of visits to your site each day (more than 100 visits per day) and have enough historical data for the algorithms to work with (at least a month), Google analytics can predict with reasonable accuracy what traffic levels are expected for the current day, week, and month. Comparing predicted values with the level of traffic you

> actually receive allows Google analytics to highlight significant changes and optionally send you email alerts about them. (Clifton 100)

These reports first model a metric's performance based on past data and then uses this to predict values for the following days. These prediction are then compared with observed values and an anomaly is detected if the observed metric is outside of some statistical bounds of the predicted value. While the window of a day for time to detect may be excessively long in the case of online content and hundreds of millions of potential viewers, for now we will focus more on this statistical bound which determines whether a value is anomalous.

> What Constitutes a Significant Change? The Google Analytics definition of a significant change, or what triggers an alert, is when a metric varies by a magnitude of X-sigma or greater from its expected value— where X-sigma is a multiple of the metric's standard deviation... The universal properties of a normal distribution are such that differing from the mean by +/- one standard deviation will account for 68 percent of all measured values. (Clifton 102)

While standard deviation is an effective metric for determining how far from a mean an observed value is, it is not the most robust feature for non-normal distributions. Standard deviation can be easily swayed by extreme values, such as previous anomalies in the data. However, it does offer advantages in algebraic usage as well as in its correlation to pdf in a normal distribution. Unfortunately, as the observed metrics are not necessarily gaussian and as more complicated algebra is not being conducted, this reduction to use standard deviation over other more robust measurements such as median absolute deviation, is possibly unjustified. Should the built in anomaly detection prove insufficient to a developer's needs, custom alerts can also be created. However, custom alerts are limited to thresholds on values or on rates of change (Clifton 103). Not only do these thresholds have to be realized by a human and manually set, but they can also prove ineffectual with seasonal metrics or metrics with a trend over time.

While Google Analytics, provides a good basis for the foundations of an analytics system, our review has discovered a number of concerns to be addressed in our research. First, the level of detail that is abstracted away for humans by creating large aggregate metrics can be utilized by an automated monitoring system focusing on smaller sets such as specific client versions or smaller time windows than a day. Second, assumption of normality of metrics is too extreme and, as such, the use of standard deviation for creating thresholds for anomalies can be improved upon with more robust techniques. Finally, moving beyond “first-level” metrics such as daily visits may reveal seasonality patterns which will require more advanced modeling techniques than fitting a gaussian distribution.

Google Analytics’ Limitations

Various case studies have focused on Google Analytics’ implementation and efficacy. As Plaza described in her evaluation of Google Analytics’ effectiveness:

> “Why use Google Analytics? Firstly, and most importantly for the purpose of this study, it is used because Google Analytics provides time series data. Moreover, it is also employed because Google Analytics is a free service offered by Google that generates detailed statistics about the visits to a website, and which is a user friendly application with the guarantee of Google technology” (477).

Ease of implementation and a friendly UI prove highly significant in the choice for this analytical tool. For automatic, anomalies however, she found other, offline applications more successful: “Google Analytics allows users to export report data in MS Excel format, which when transformed can be analyzed with time series statistical programs” (Plaza 479-480). However, other studies found GA’s anomaly detection useful but limited in its application. Besides an on average two-hour delay in GA reports, Wei Fang found in his study that their own metrics or data, such as log files, could not be imported into GA and that aggregated analytics information was shared with third parties (4). These are aspects we sought to improve upon in our experiments to develop automatic anomaly detection tools, namely in the areas of real time updates, customizable metrics based on the service, and advanced anomaly detection algorithms.

Contextual Inquiry

Having analyzed one of the most popular anomaly detection tools available, we pursued our investigation further by talking to both the engineers and users of such anomaly detection tools, engineers at the video quality analytics company Conviva. As these engineers provide a similar UI to Google's for monitoring video data, we discussed with them what metrics they use and how they classify an anomaly. We first inquired as to what metrics they track to categorize their service's performance. They track video start failures, time spent buffering, viewing attempts, exits prior to video starting, and duration watched, to name a few. Examining the patterns of these metrics, we noted two main categories of metrics: seasonal, such as viewing attempts which fluctuate regularly on a daily basis or perhaps see a regular spike each Wednesday at 7PM when a new episode is released, and nonseasonal, such as video start failures which hover at a steady rate with some noise. We also discussed what the various failure models consisted of and their representation in the data. As with Google Analytics, they consider an anomaly is one point, or a series of points, which is significantly uncharacterized by past behaviour in that metric.

Also as in Google Analytics, the primary feature in determining thresholds was not to make a fixed threshold but rather a human adjustable measure of sensitivity, critical to avoiding oversaturation of false positives on the analyst's part, or, as Clifton described it above, "data overload." At this point, visualization of these metrics for human monitoring, as well as setting fixed threshold based alerts, were the only working system they currently had in place for anomaly detection. The granularity of their reports was on much finer intervals than GA, seconds to minutes, and the quantity of metrics as well as capability of the interface to filter metrics by certain subgroups such as geographic area were very capable of facilitating quick diagnosis of anomalies by a human analyst. As for a more automated approach, they were eager to see what progress could could be made in this area. Throughout our research we met several times with these engineers to discuss our strategies and intermediate results towards creating our Smart Anomaly Detection toolset.

Methods and Materials

Materials

Throughout our research, we employed three critical tools: Conviva's online video data, Databricks' computing platform, and iPython on Apache Spark. Conviva provided months of global online video data from multiple content providers for analysis. Without access to this data set reflective of real online content performance and user experience, we would not have been able to validate our techniques and have been relegated to hypothetical justifications for our methods. As for Databricks and Spark, these tools provided the environment for processing the large data sets available to us efficiently through their memory based cluster computing capabilities, not only in the exploratory data analysis phase but also in production of our final classifier for Smart Anomaly Detection. Spark's memory based RDDs allowed us to rent significant memory resources on EC2 for fast processing and examination of our data set.

Methods Introduction

During the development of effective automatic anomaly detection tools for online video data, my research was primarily focused on exploring various algorithms, validating assumptions for the various statistical methods we intended to use, implementing anomaly detection for nonseasonal metrics, and evaluating our classifiers. During my initial exploratory data analysis, I graphed and helped select, from among our various metrics, attempts and video start failures as the most significant, and failure-wise most telling, representatives of seasonal and nonseasonal metrics respectively. Total attempts are very indicative of the number of viewers at any given time and can indicate whether a video is trending or unreachable such as if the content was improperly uploaded. Video start failures as a ratio of total attempts helps reveal when there is an above average number of failures such as if a Content Distribution Network (CDN) is failing or if a new version of a player is preventing a video from starting. The attempts metric demonstrates regular peaks and dips throughout the day whereas VSF maintains a relatively static mean and distribution, with slight trends over time. For my portion of the toolset implementation, I worked to address the nonseasonal VSF data set, visualized below.

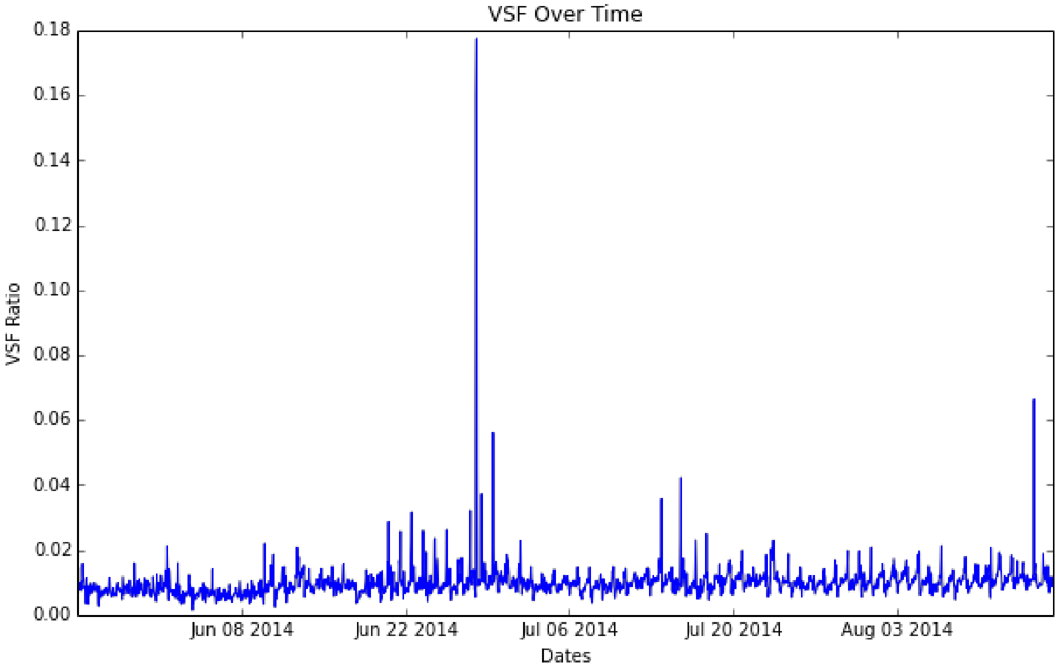

Figure 1

Preliminary Technique Validation

To validate the techniques we would use for identifying anomalies on nonseasonal metrics, first we needed a general way to compare anomaly detection performance across various techniques and various data sets. To begin, we created several graphing functions to produce visualizations which highlight the points that our algorithms label as anomalies and which display the anomalies' location in the overall data set relative to time, as seen in a hourly seasonal data visualization below. The blue line is the observed attempts values and the green line is the predicted value of attempts. The red point is the predicted anomaly. This early stage evaluation tool was effective in providing a quick visual of the efficacy of a particular technique in separating anomalies from regular data.

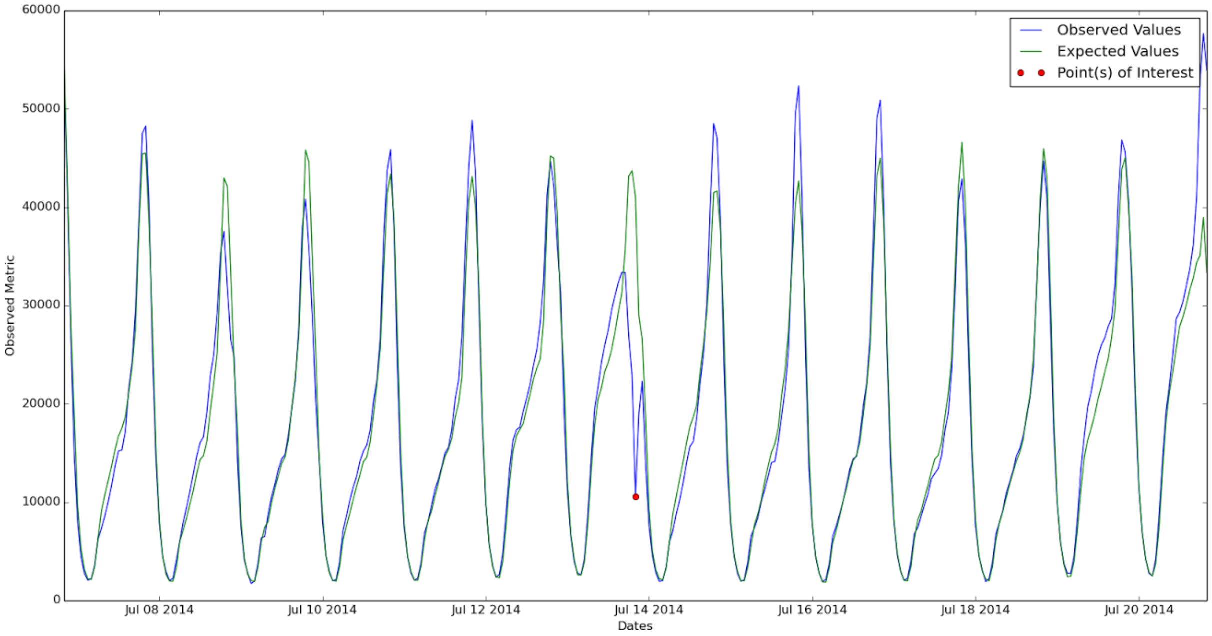

Figure 2

After a technique proved effective under this early visual evaluation tool, we pursued further to examine the technique's performance on smaller time windows, such as minute sized time windows. With these increased number of total points, in order to accommodate gradual trends over time, smaller quantities of previous points were trained on, implementing a moving window model. In addition, we also filtered data sets based on additional metrics such as geography or device version to view this method's effectiveness in not only detecting an anomaly but also diagnosing a more precise location and cause of the anomaly.

Formal Technique Validation

To evaluate an individual technique's performance in a more formal method, we created a test set of several weeks of data aggregated on a metric and on a minute by minute basis in which each point is labeled on a scale of (0-3) in which a 0 would be a non-anomalous point, a 1 would be a an extreme but non-anomalous point, a 2 would be a slightly anomalous point, and a 3 would be a significant and distinguishable anomaly. This labeling was done in coordination with Conviva's engineers for confirmation of the labeling method used. My work included implementing the labeling program in which a labeler would be presented with a point to be labeled as well as several weeks of preceding points. A sample view seen by the labelers can be seen below. The line in the graph represents the amount of attempts relative to time. The rightmost red dot represents the point they are being asked to label. The four preceding red dots and the horizontal lines emanating from them, are guidance points. They are located at exactly 1, 2, 3, and 4 weeks prior. They provide a point of reference for the user in labeling incoming points. The user then uses this information to make a decision on the label of the point as to whether the current point is an anomaly and, if so, how anomalous, via the (0-3) scale.

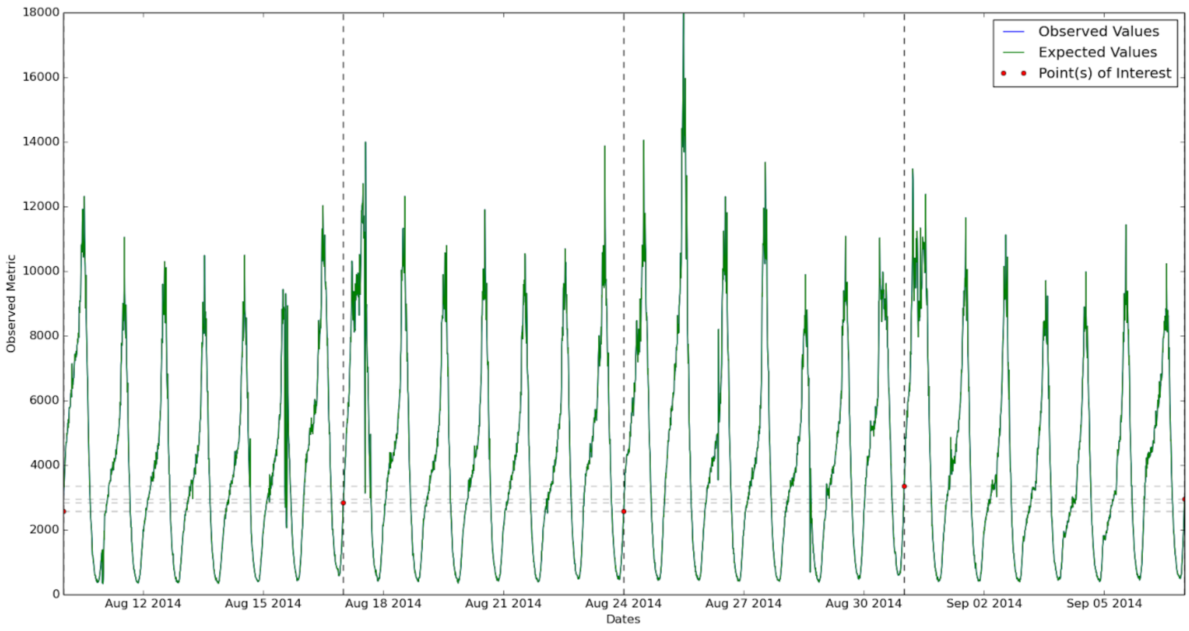

Figure 3

For the VSF data set, we labeled two weeks of 10 minute VSF aggregations, totaling 2,016 points. Of these points, there were 12 labeled anomalies in the form of 3s. For the attempts dataset, we labeled six weeks of data of 10 minute aggregations, expanding the number of weeks labeled due to only one recorded anomaly in the original two week span. The final six week data set for attempts contained 16 labeled anomalies. Once both labeled sets were created, they were separated into two halves. In the case of VSF data, points were separated between the first week and the second week. The first half was then used as a validation set to tune the technique's sensitivity parameter to detect anomalies. The second half of the data was was then used as a test set to evaluate that parameter's efficacy on new data.

After evaluating the accuracy of the different implementations, we then communicated those techniques' final performance to Conviva's engineers as evidence of this Smart Anomaly Detection system's ability to replace human based monitoring techniques.

Potential Seasonal Algorithms

As the majority of my work on the seasonal anomaly detection toolset was in research and final evaluation, we begin by examining the various algorithms considered for its implementation. The most considered techniques include, anomaly detection in ECG by automated external defibrillators, Holt-Winters, and Box-Jenkins based ARMA.

Automatic External Defibrillator (AED) Techniques

As we intended to produce fast, constant, high-accuracy monitoring on cyclical data with possibly limited resources as a monitoring system may be running on thousands of metrics, such as attempts or view duration, over thousands of subgroups, such as client version or geography, AED's seemed a natural point of comparison. AED's monitor the heart rate of a patient and deliver a shock at the proper time when anomalous heart rhythms are detected. Due to the severe consequences of failure, these techniques have to demonstrate high accuracy. As one study in techniques for anomaly detection with AEDs noted, “Thus, we get an accuracy of 90% on this testing dataset using our adaptive window based discord discovery scheme. The false positive rate is 10%” (Choo 132). However, as the accuracy rate one of the highest performing algorithms in the study, a 10% false positive rate seems high for anomaly detection and could cause “data overload.” In the field efficacy may be assisted by humans intervening to place and activate the device when an anomaly appears already in progress. Thus, for AEDs, erring on the side of a false positive may be satisfactory. Due to the urgency of the situation, accuracy must be compromised with speed, though for AED’s there are a number of available techniques to allow for this tradeoff. “For an electrocardiogram signal from body-surface, after a pre filtering process, there are many software methods for fibrillation detection, including the time-field detection method, frequency-field detection method, time-frequency analysis detection method and dynamics analysis method” (Fan 22). Combined with the restrictions of speed and accuracy, these processes must operate with restricted computing resources, relying on the microprocessors offered by the AED. With common AED systems having only 64KB of on chip memory and a speed of 800MHz, these algorithms must be streamlined and lightweight to respond to the urgent response time required (Texas Instruments).

After careful evaluation, we decided AED techniques were inappropriate for our problem. Simply by having a human apply the AED during an incident, the algorithm is already skewed to expect an anomaly rather than normal heart behavior. In addition, most AED techniques require an entire cycle to complete before being compared with the expected cycle and determined to be an anomaly. The amount of time viewed necessary past an event to diagnosis it, which may be milliseconds for an AED, may be minutes or hours for online viewing cycle. Subsequently, we investigated more general anomaly detection techniques based on seasonal modeling, namely Holt-Winters.

Holt-Winters (HW)

Holt-Winters is an algorithm used to model series, via exponential smoothing, commonly compared with Box-Jenkins autoregressive moving average (ARMA) techniques. Though it is commonly accepted to be less generally applicable than the Box-Jenkins model, before dismissing it, we researched whether to use this technique by referencing Chatfield's "The Holt-Winters Forecasting Procedure." In this work, he compared the effectiveness for both sets of techniques when used in a more automated fashion, as well as when tailored to a specific time series. For the automated case, Chatfield described "the univariate Box-Jenkins method gave more accurate forecasts than the HW procedure for about two-thirds of the series analysed" (265). When tuned on Series C, a data set with similar seasonality as in our attempts metric, HW showed no significant advantage over BJ. "Although HW was expected to have inferior accuracy to BJ for these series, the coefficients of determination were all found to be reasonably high, ranging from 77% for Series C to over 99% for Series G. In the latter case, the HW method would almost certainly be judged sufficiently accurate even though the BJ forecasts may be even better" (Chatfield 270). Despite efforts to fine tune HW for performance based on the data's patterns, overall this technique saw minimal improvement and there was no apparent distinction in the data sets for which HW performed superiorly to BJ. "Series for which BJ forecasts are much better than forecasts from the automatic HW procedure do not appear to have any common properties" (Chatfield 276). Despite this lack of separation, Chatfield does note the potential of ARMA to outperform HW in certain cases. "This combination of two sets of 'automatic' forecasts gives results comparable in accuracy to those of Box-Jenkins. Intuitively, the stepwise autoregression forecasts may be able to take account of autocorrelation in the time series which cannot be described by a HW model" (Chatfield 273). Without any definitive support that the HW model was equally viable, we followed Chatfield's unintentional recommendation to pursue an autocorrelated model.

ARMA

The autoregressive moving average uses a modeling technique in which, given a time series, every point is correlated with some subset of points preceding it. Parameters for this subset are calculated to provide a regression for predicting future values based on past values and some noise that is represented by a constant standard deviation value for the series (Box and Pierce 1509). After preliminary tests to confirm the validity of using this technique, we proceeded with this algorithm, discussed in further detail in Benjamin Le's paper, who took the lead on implementing this technique.

Potential Nonseasonal Algorithms

In investigating techniques for nonseasonal metrics, finding accurate representations of the distribution and finding approximate expected deviation serve as essential components in the domain of distinguishing an anomalous point from non-anomalous points. We used VSF ratio as our model for these nonseasonal metrics. A box plot of its distribution can be seen below. In this one dimensional box plot, the y axis represents the different VSF ratio values that were seen in the data. The box represent the location of the data points from the 25th percentile to the 75th percentile, a region termed the interquartile range (IQR). The blue plus signs are the points which fall beyond 1.5*IQR above or below the box, a boundary separated by the black T-shaped whiskers .

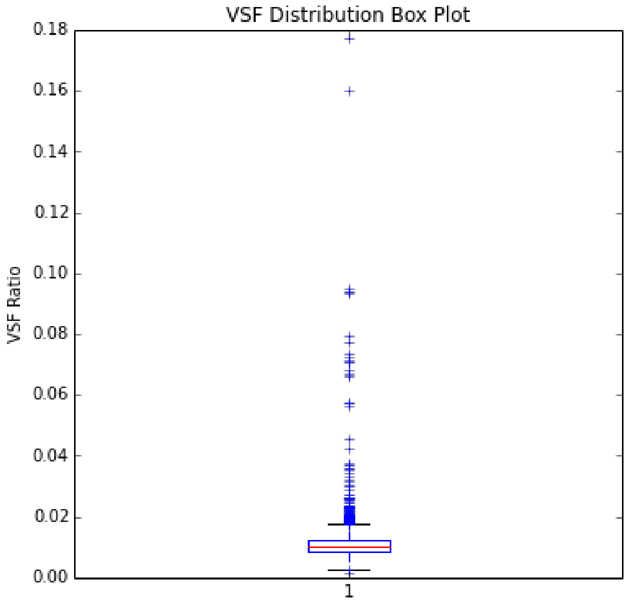

Figure 4

To classify observed points in this distribution, principally we investigated three techniques: standard deviation, MADe, and Weibull distribution.

Standard Deviation

As described in the examination of Google Analytics, the technique of using standard deviation to estimate the likelihood that an observed point is from a given distribution is a well established and used technique. However, to compute this likelihood, Google Analytics, made the assumption that the distribution was normal. An Anderson-Darling normality test, in which, assuming normality, the deviation of both sides from the expected cumulative density function is used to predict normality, on the distribution for VSF produced a highly statistically significant result that the distribution is non-normal, as visible in the preceding box plot. As Songwon Seo describes in his evaluation of several algorithms for outlier detection: “3 SD [standard deviation] Method: x ± 3 SD, where the mean is the sample mean and SD is the sample standard deviation. The observations outside these intervals may be considered as outliers” (9).

However, for non-normal distributions, skew or outliers can greatly sway the standard deviation. In such cases, median absolute deviation can prove a more robust measurement (Walker 24-25).

Weibull Distribution

While the VSF data is not from a normal distribution, it does appear that the regularly occurring points may represent a skewed distribution. In the histogram below, we can view the distribution of VSF data post pruning all points above a threshold of 3 standard deviations away. This skewed shape might be indicative of a Weibull distribution. While still assuming that the data is unimodal, this distribution offers additional fitting parameters, particularly on skew, to match our non-normal data sets (Belyaev). After validating whether a Weibull distribution is an appropriate fit for given metric, the fitted Weibull's pdf can be used to approximate the likelihood of newly observed points and subsequently provide a threshold for anomalous points based on their probability of occurring given the previous distribution. The efficacy of this technique is described in greater detail in Wenxuan Cai's paper who took the lead on this initiative.

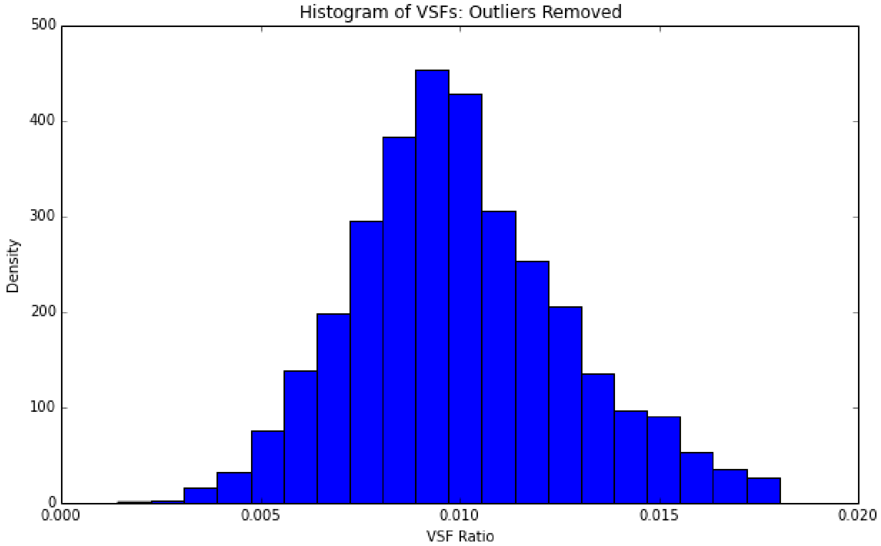

Figure 5

Median Absolute Deviation

In place of standard deviation, a similar metric for estimating deviation is median absolute deviation. “The MADe method, using the median and the Median Absolute Deviation (MAD), is one of the basic robust methods which are largely unaffected by the presence of extreme values of the data set” (Seo 17). Using the median helps insulate against extreme values or potential anomalies in the data. “Since this approach uses two robust estimators having a high breakdown point, i.e., it is not unduly affected by extreme values even though a few observations make the distribution of the data skewed, the interval is seldom inflated, unlike the SD method” (Seo 17). After calculating the median of a data set, we take the absolute value of the difference of each point from the median and then select the median point from these distances to locate the MAD. Multiplying this value by 1.483 produces the MADe which can be used in place of SD in similar thresholding techniques, such as with Z-Scores, to determine whether points are outliers.

$$MADe \; = \; 1.483 * median(|VSF_i - median(VSF_i)|)$$

Preliminary results demonstrated high success in separating anomalies from normally occurring points.

Results and Discussion

With the two weeks of 10 minute interval labeled VSF data for evaluating the MAEe technique, I used the first week to tune the sensitivity parameter, ρ, which appears in the form of the number of MADe’s away from the median to serve as the threshold between anomalous and normal points. For an incoming point to be predicted on, the algorithm generates median and MADe values based on the preceding four weeks of data. If the incoming point is more or less then ρ*MADe away from the median, then the point is labeled as an anomaly. Following tuning the ρ, the algorithm was then run on the second week with the same ρ and evaluated for performance. As our primary metric of performance, we used F-Beta, specifically F1 which is the harmonic mean of both recall and precision. We selected this metric under the reasoning from our discussion with Conviva engineers that while a number of false calls are expected, too many could result in alert fatigue, an indication of the previously mentioned “data overload.” However, on the opposite end of the spectrum, recall prove similarly significant as missing a critical anomaly can result in deteriorations of user experience. The MADe classifier performed above our expectations on the VSF test set and the results are presented in five formats:

- A ROC plot of MADe employed on both weeks to demonstrate the effectiveness of ρ as a classifier. A Receiver Operating Characteristic plot shows, with a binary classification problem, how the threshold can be varied to increase or decrease the true positive rate on the y-axis and the false positive rate on the x-axis. Each blue dot is the MADe method employed with a specific ρ value, with values from 0 to 10 by increments of 0.1. The straight line represents a naive classifier which, in the bottom left corner, would classify all samples as false, and in the top right would classify all samples as true. An ideal classifier would be able to adjust the threshold to attain the top left corner, with a 100% true positive rate and a 0% false positive rate. The farther a classifier is able to pull away from the naive classifier and is able to reach the top left corner, signifies its discriminative ability, and is represented by the area under the curve.
- An F1 graph by ρ. The dotted line represents the F1 performance of MADe on week one with various values of ρ. The green dot represents the performance on week two using the highest ρ tuned from week 1, 5.5 MADe’s.
- A confusion matrix for the final test on week two with our optimal sensitivity value of 5.5 MADe’s.
- A listing of additional results from the highest performing experiment from week 1 and the final test on week 2 in the forms of accuracy, F1, precision, and recall.

- A chart depicting the anomalies labeled by the test set, the red dots, and the anomalies predicted by the MADe classifier with $\rho = 5.5$, the yellow dots, on both weeks of data. Points showing with both sets of dots were correctly predicted.

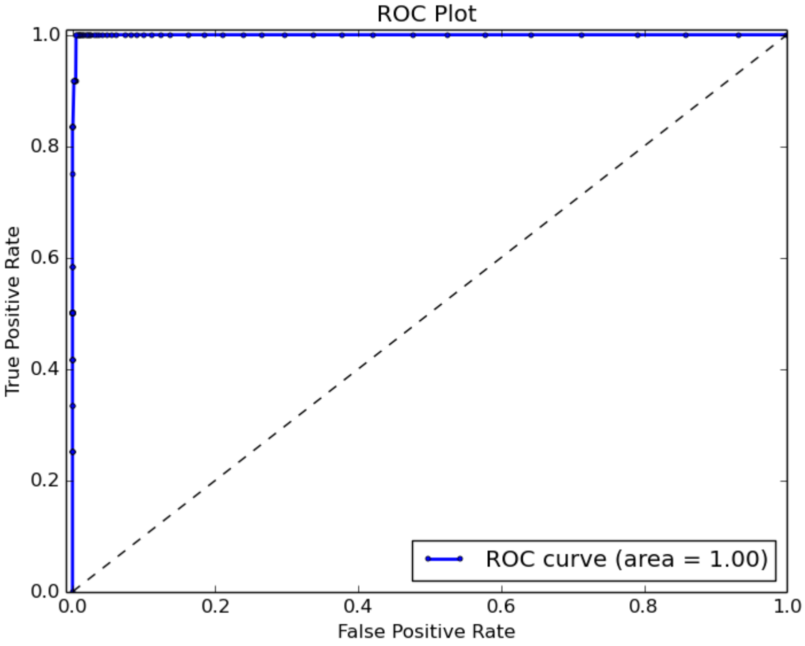

Figure 6

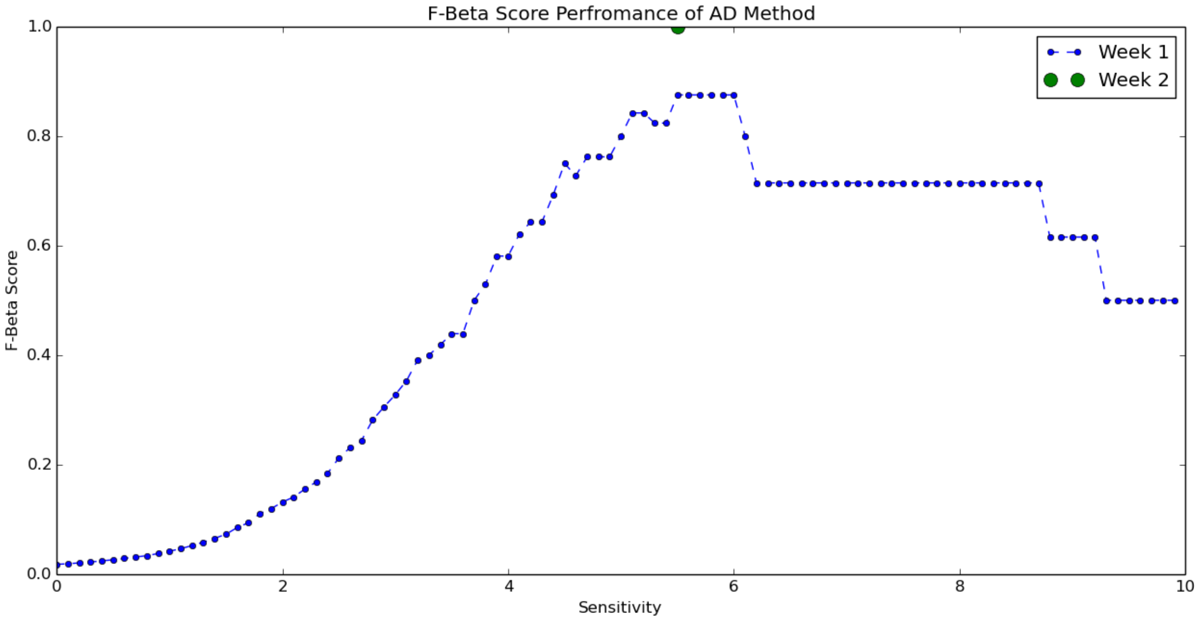

Figure 7

Confusion Matrix for Week 2:

	Predicted Positive	Predicted Negative
Actual Positive	3	0
Actual Negative	0	1005

Figure 8

Additional Metrics for Weeks 1 and 2:

Week	1	2
Accuracy	0.9980	1.0
F1	0.9891	1.0
Precision	1.0	1.0
Recall	0.7778	1.0

Figure 9

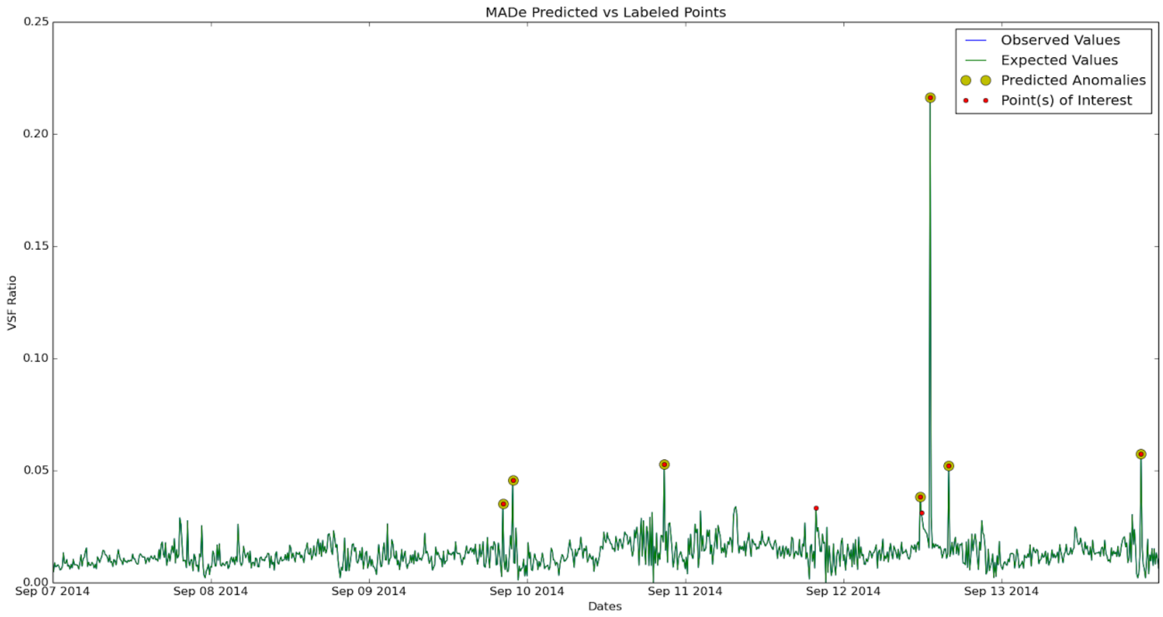

Figure 10: Week 1 Predicted vs Labeled Points

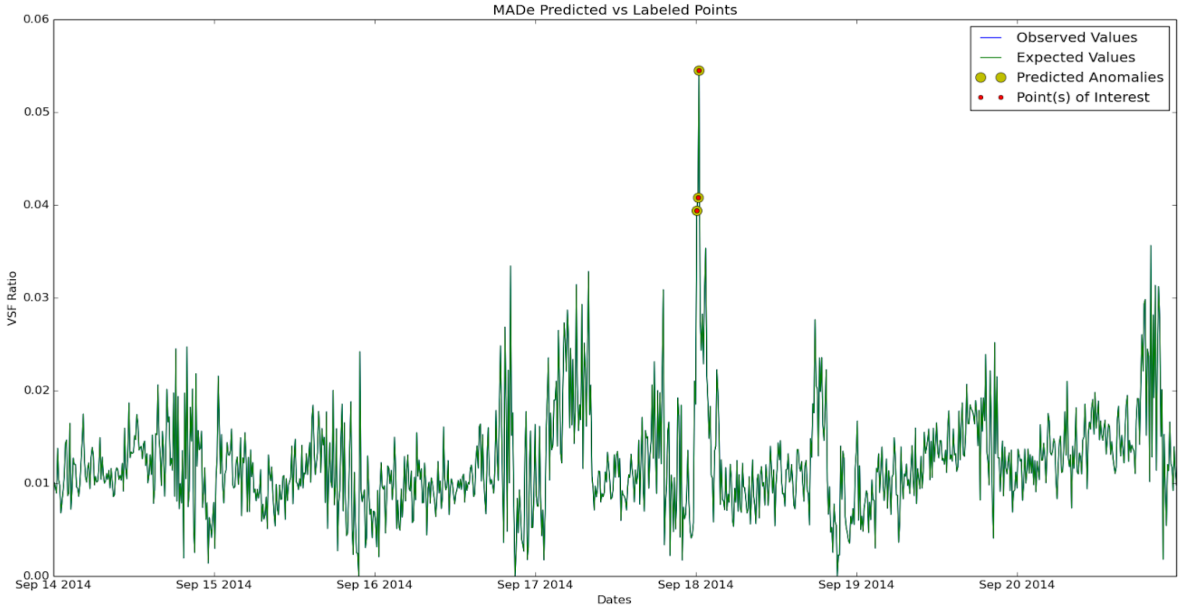

Figure 11: Week 2 Predicted vs Labeled Points

Once our classifier was trained for sufficient accuracy and tested, we then proceeded to analyze smaller aggregate sets for the purpose of anomaly diagnosis. By realizing an anomaly in a smaller aggregation, such as across a specific internet AS, Autonomous System, our tools can help detect an anomaly that would otherwise go undiscovered. In addition, when the larger aggregate set detects an anomaly, these smaller sets can help identify and zero in on a specific cause. For this experiment, we used smaller labeled sets of 100 points each on three specific ASNs, each approximately 20% of the aggregate data set's size. This chart displays the separation between the aggregate set's values, i.e. the observed values, versus the smaller aggregate set's values, i.e. the expected values based on the aggregate, and the anomalies detected in the smaller set by the same MADe classifier with ρ = 5.5, in red.

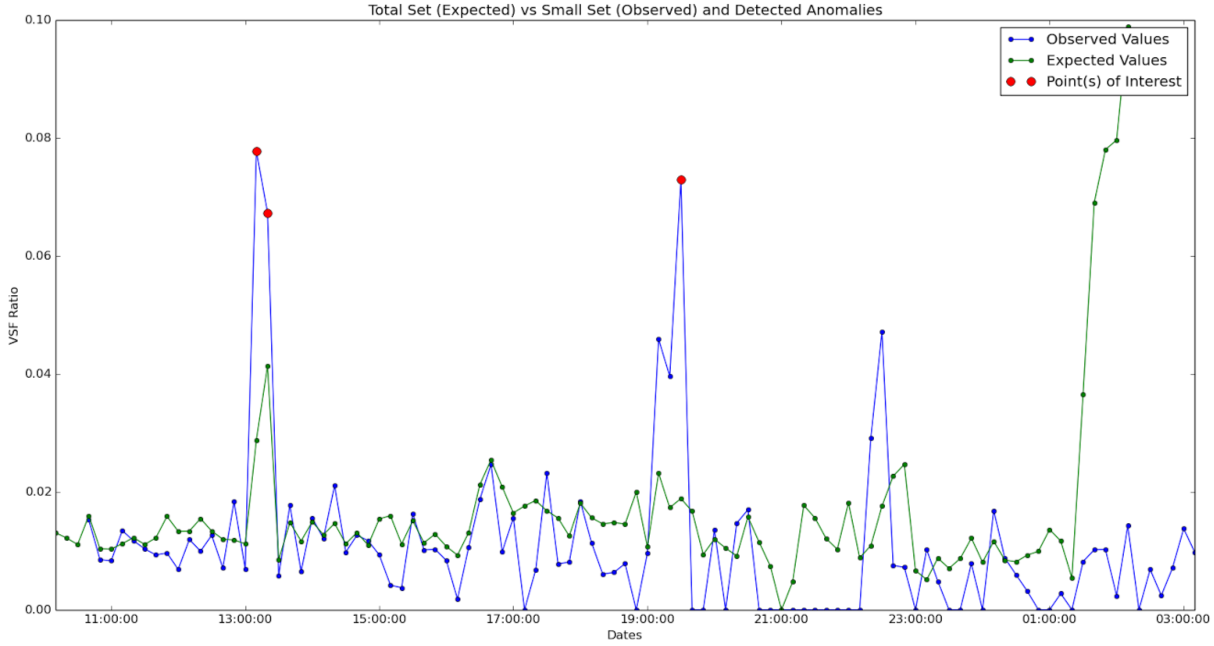

Figure 12

The predicted anomalies were then compared with the human labeled set for the following results, suggesting a highly effective replacement for human monitoring.

Confusion Matrices:

Smaller Set 1	Predicted Positive	Predicted Negative
Actual Positive	2	0
Actual Negative	1	97

Figure 13

Smaller Set 2	Predicted Positive	Predicted Negative
Actual Positive	2	0
Actual Negative	0	98

Figure 14

Smaller Set 3	Predicted Positive	Predicted Negative
Actual Positive	3	3
Actual Negative	0	94

Figure 15

Additional Metrics:

Category	Smaller Set 1	Smaller Set 2	Smaller Set 3
Accuracy	0.99	1.0	0.97
F1	0.8	1.0	0.6667
Precision	0.667	1.0	1.0
Recall	1.0	1.0	0.5

Figure 16

Conviva engineers have been suitably impressed with the effectiveness of our techniques and are eager to see implementations of such techniques to assist content developers in detecting and diagnosing anomalies in their service. While Smart Anomaly Detection is not only a useful tool to avoid “data overload,” our results suggests that our techniques promise to be an effective path for such automation in online services to come.

VIII. Conclusion

Project Status

Based on our original evaluation of the problem of automatic anomaly detection, our endeavors made significant headway toward this end. We aimed to produce tools demonstrated to automatically replicate the human identification of spike based anomalies in online video data. We began by focusing on defining an anomaly as a point statistically unlikely to appear based on normal past behavior, with the likelihood threshold to be be determined by a sensitivity parameter. In addition, we needed to characterize different forms of normal behavior for a metric, such as whether the metric it was seasonal on a daily cycle, or whether it was non-seasonal and hovering at a value with some noise of distribution. Also, we needed a technique which could be retrained and run quickly to allow for real-time monitoring in addition to parallel monitoring on possibly vast numbers of metrics, or restricted subsets of metrics such as from a particular CDN, to diagnose and zero-in on anomalies. We kept these considerations in mind as we selected techniques and successfully implemented scalable, tunable, and rapid algorithms which performed as reasonable automated replacements of the current gold standard of human based monitoring. In addition, our aggregate subset test investigated the possibility of employing these techniques beyond a human labeler to not only direct an admin’s attention to a specific area of concern, but to also detect an anomaly that could be missed on the aggregate scale.

Overall, we successfully completed the majority of our original plans for smart anomaly detection as well as investigated the area of focusing on smaller subsets to facilitate diagnosis of an anomaly's cause. As for future areas of research, initially considered but postponed to dedicate our focus to the more critical, essential components, these will be discussed in the Future Research section below.

Management Insights

As project manager for my team, I was responsible for coordinating team meetings as well as checking in on team members, ensuring tasks were progressing, and assisting parts of the team that seemed to be encountering difficulties. The primary insight I gained from this experience is that team management should not be taken lightly. With varied schedules and only a few valuable hours available from each team member per week, meetings with our advisor, with teams, with subteams, and with our industry partner can quickly overwhelm available resources and meetings must be streamlined, effective, and most importantly timeboxed. In addition, being on point to help resolve issues of organization, task allocation, and direction can be overwhelming for one person who is responsible for an equal share of the research requirement. Halfway through the project, following a team discussion, we more evenly delegated some management tasks, such as resolving a critical issue with our programming environment, to other team members, though my logistical responsibilities continued to prove a formidable burden alongside my own research duties, particularly as deadlines loomed and write ups and reports appeared. Team management, done correctly and efficiently, is an essential tool to keeping a team well oiled and functioning smoothly and in unison. As such, it should be considered a full time position and having a team member dedicated to solely to management may be a worthy investment to maximize a team's potential in each of its members.

Team management aside, there is a significant distinction between a research project and developing a project with business prospects in mind. Namely, in addition to R&D, the researchers must repeatedly halt their progress to provide reports or presentations, or to participate in a groupthink session on various business topics such as marketing or IP. This halting and redirection of efforts requires them to not only persist in their research but also to continually re-predict the final product and update their various market considerations as such. Should the resources have permitted, a more adequate allocation may have existed in maintaining members who separately handled project management as well as business interests, with the PM acting as the liaison and expert counsel between the team and the business manager. In this respect, members could specialize and reduce retooling and team-wide meeting overhead costs. However, the scope of our research as well as the requirements of our business based responsibilities resulted in all team members having to share multiple roles, with those capable or with more acumen in specific areas providing additional time as deadlines required.

Future Research

Our team made significant headway in developing more sophisticated tools for anomaly detection than currently available, though additional areas for research are possible in the forms of multipoint anomalies and adaptation to other domains.

Multipoint Anomaly Detection

While for real-time detection, admins may desire to detect anomalies when the occur, some anomalies may not become apparent until taken in consideration with several points. Proceeding beyond spike detection, detecting multipoint anomalies consists of finding a series of points which, when considered together, prove probabilistically separate from previous points. For those considering this area of further research, we advise two approaches. First, a researcher can utilize our techniques on a group of points with a heightened sensitivity which requires anomaly detected by all points, or some percentage of the points, to return an anomaly. Second, using distribution comparison techniques, a researcher could compare distributions of previous points with the distribution observed in a new set of points and set a sensitivity parameter tuning for when the separation between the two distributions warrants an anomaly. Via these techniques, delayed response time would be mitigated by reduced false positive rates and higher accuracy by incorporating a wider selection of failure models than spike detection.

Adaptation to other domains

The techniques employed in our endeavor are not restricted to solely this domain. Rather, we foresee further adaptation and implementation of these techniques for anomaly detection from various domains, such as general website analytics or hardware monitoring. Future researches can replicate our tools on a variety of data sets demonstrating seasonal and nonseasonal metrics. We intend for our work to serve as the foundation to providing automated anomaly detection not only to the field of online media but to any field with data beyond the scope of human parsing. This data is waiting to be harnessed by machine learning for its value beyond simply a human monitoring some high-level aggregation on a screen; this harnessing will serve as a critical advantage for any business in the digital age, an environment increasingly punctuated by big data.

IX. Acknowledgements

We would like to thank George Necula for advising us throughout the entirety of this Capstone project. We would also like to thank Jibin Zhan from Conviva for introducing the problem space to us and Pat McDonough from Databricks for providing extensive technical support throughout our use of Databricks.

References

Alice Corporation v. CLS Bank. 573 U.S. Supreme Court. 2014. Print.

"AM3703 (ACTIVE)." AM3703. Texas Instruments, n.d. Web. 16 Mar. 2015. <http://www.ti.com/product/am3703>.

Associated Press. “Netflix reeling from customer losses, site outage.” *MSNBC*. MSNBC. 24 July 2007. Web. 15 Feb. 2015.

Belyaev, Yu K., and E. V. Chepurin. "Weibull Distribution." *Encyclopedia of Mathematics*. N.p., n.d. Web. 16 Mar. 2015. <http://www.encyclopediaofmath.org/index.php?title=Weibull_distribution&oldid=18906>.

Bessen, James. “The patent troll crisis is really a software patent crisis.” *Washington Post*. The Washington Post. 3 Sept. 2013. Web. 27 Feb. 2015.

Biem, Alain E. “Detecting Anomalies in Real-time in Multiple Time Series Data with Automated Thresholding.” International Business Machines Corporation. US Patent 8,924,333. 30 Dec. 2014.

Box, George EP, and David A. Pierce. “Distribution of residual autocorrelations in autoregressive-integrated moving average time series models.” *Journal of the American statistical Association* 65.332 (1970): 1509-1526.

"Bringing Big Data to the Enterprise." IBM. N.p., n.d. Web. 13 Apr. 2015.

Brundage, Michael L., and Brent Robert Mills. “Detecting Anomalies in Time Series Data”. Amazon Technologies, Inc., assignee. U.S. Patent 8,949,677. 3 Feb. 2015.

Byrd, Owen, and Brian Howard. 2013 Patent Litigation Year in Review. Rep. Menlo Park: Lex Machina, 2014. Print.

CA Inc. “Manage Your Network Infrastructure for Optimal Application Performance.” *CA Technologies*. n.p. n.d. 13 Feb. 2015.

Campbell, Kunle. "Google Analytics: Tracking 4 Stages of Customer Interaction." *Practical Ecommerce*. N.p., 11 Mar. 2015. Web. 12 Mar. 2015. <http://www.practicalecommerce.com/articles/83108-Google-Analytics-Tracking-4-Stages-of-Customer-Interaction>.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM Computing Surveys (CSUR)* 41.3 (2009): 15.

Chatfield, Chris. "The Holt-Winters Forecasting Procedure." *Applied Statistics* (1978): 264-279.

Clifton, Brian. *Advanced Web Metrics with Google Analytics* (2nd Edition). Hoboken, NJ, USA: Sybex, 2010. ProQuest ebrary. Web. 13 March 2015.

Cohn, Chuck. “Build vs. Buy: How to Know When You Should Build Custom Software Over Canned Solutions.” *Forbes*. Forbes Magazine, 15 Sep. 2014. Web. 7 Apr. 2015.

Connelly, J.P., L.V. Lita, M. Bigby, and C. Yang. “Real time audience forecasting.” US Patent App. 20120047005. 23 Feb. 2012.

Conviva. "About Us." *Conviva*. n.p., n.d. Web. 28 Feb. 2015.

Cortes, Corinna, Lawrence D. Jackel, and Wan-Ping Chiang. "Limits on learning machine accuracy imposed by data quality." *KDD*. Vol. 95. 1995.

Dasgupta, Dipankar, and Stephanie Forrest. "Novelty detection in time series data using ideas from immunology." *Proceedings of the international conference on intelligent systems*. 1996.

Deshpande, Amit and Riehle, Dirk. "The total growth of open source." *Open Source Development, Communities and Quality*. Springer US, 2008. 197-209.

Etherington, Darrell. "Twitter Acquires Over 900 IBM Patents Following Infringement Claim, Enters Cross-Licensing Agreement." TechCrunch. N.p., 31 Jan. 2014. Web. 25 Feb. 2015.

Fan, Aijun, Peng Han, and Bin Liu. "Shockable Rhythm Detection Algorithms for Electrocardiograph Rhythm in Automated Defibrillators." *AASRI Procedia* 1 (2012): 21-26. Web.

Fang, Wei. "Using Google Analytics for Improving Library Website Content and Design: A Case Study." *Library Philosophy and Practice* 9.2 (2007): 22.

Fisher, William W. “Patent.” *Encyclopaedia Britannica Online*. Encyclopaedia Britannica Inc.

Ganjam, Aditya, et al. "Impact of delivery eco-system variability and diversity on internet video quality." IET Journals 4 (2012): 36-42.

Goldman, Eric. "The Problems With Software Patents (Part 1 of 3)." *Forbes*. Forbes Magazine, 28 Nov. 2012. Web. 01 Mar. 2015.

Gottfriend, Miriam. “Bullish Investors See New Hope for Netflix Profit Stream.” *The Wall Street Journal*. The Wall Street Journal. n.d. Web 14 Feb. 2015.

Hanley Frank, Blair. "Amazon Web Services Dominates Cloud Survey, but Microsoft Azure Gains Traction - GeekWire." *GeekWire*. Geekwire, 18 Feb. 2015. Web. 02 Mar. 2015.

Harvey, Cynthia. "100 Open Source Apps To Replace Everyday Software." *Datamation*. N.p., 21 Jan. 2014. Web. 28 Feb. 2015.

Iyengar, Vijay S. 2002. “Transforming data to satisfy privacy constraints.” *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '02). ACM, New York, NY, USA, 279-288. Web. 12 Feb. 2015.

Jasani, Hiral. “Global Online Video Analytics Market.” *Frost & Sullivan*. n.p. 5 Dec. 2014. Web. 12 Feb. 2015.

Kahn, Sarah. "Business Analytics & Enterprise Software Publishing in the US." IBISWorld (2014): 5. Web. 11 Feb. 2015.

Keaveney, Susan M. "Customer switching behavior in service industries: An exploratory study." The Journal of Marketing (1995): 71-82.

Kejariwal, Arun. "Introducing Practical and Robust Anomaly Detection in a Time Series." Twitter Engineering Blog. Web. 15 Feb. 2015.

Lawler, Richard. "Netflix Tops 40 Million Customers Total, More Paid US Subscribers than HBO." *Engadget*. N.p., 21 Oct. 2013. Web. 15 Feb. 2015.

Liu, Xi, et al. "A case for a coordinated internet video control plane." Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication. ACM, 2012.

Mooi Choo Chuah and Fen Fu. 2007. “ECG Anomaly Detection via Time Series Analysis.” In *Proceedings of the 2007 International Conference on Frontiers of High Performance Computing and Networking* (ISPA'07), Parimala Thulasiraman, Xubin He, Tony Li Xu, Mieso K. Denko, and Ruppa K. Thulasiram (Eds.). Springer-Verlag, Berlin, Heidelberg, 123-135.

Mcgovern, Gale. Virgin Mobile USA: Pricing for the Very First Time. Case Study. Boston. Harvard Business Publishing, 2003. Print. 9 Jan. 2010.

“Number of Broadband Connections.” *IBISWorld*. IBISWorld. 3. Web. 12 Feb. 2015.

Numenta. “The Science of Anomaly Detection.” *Numenta*. n.p. n.d. 13 Feb. 2015.

Open Source Initiative. "Welcome to The Open Source Initiative." *Open Source Initiative*. N.p., n.d. Web. 28 Feb. 2015.

Plaza, Beatriz. "Google Analytics for Measuring Website Performance." *Tourism Management* 32.3 (2011): 477-81. Web.

Porter, Michael. "The Five Competitive Forces That Shape Strategy." *Harvard Business Review Case Studies, Articles, Books*. N.p., Jan. 2008. Web. 12 Feb. 2015.

Porter, Michael. "What is Strategy?." *Harvard Business Review Case Studies, Articles, Books*. N.p., Jan. 2008. Web. 12 Feb. 2015.

Quinn, Gene. “A Software Patent Setback: Alice v. CLS Bank.” *IP Watch Dog*. n.p. 9 Jan. 2015. Web. 27 Feb. 2015.

Ramos, Kevin. "Topic: Online Video." www.statista.com. N.p., Oct. 2013. Web. 14 Mar. 2015. <http://www.statista.com/topics/1137/online-video/>.

Roettgers, Janko. "Netflix Spends $150 Million on Content Recommendations Every Year." *Gigaom*. N.p., 09 Oct. 2014. Web. 15 Feb. 2015.

Seo, Songwon. *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. Diss. University of Pittsburgh, 2006.

Shelby County v. Holder. 570 U.S. Supreme Court. 2013. Rpt. in Dimensions of Culture 2: Justice. Ed. Jeff Gagnon, Mark Hendrickson, and Michael Parrish. San Diego: University Readers, 2012. 109-112. Print.

Smith, Sarah. "Analysis of the Global Online Video Platforms Market." -- *LONDON, Jan. 5, 2015 /PRNewswire/* --. Reportbuyer, n.d. Web. 02 Mar. 2015.

Sun Tzu, and James Clavell. *The Art of War*. New York: Delacorte, 1983. Print. 17-18.

Trautman, Erika. “5 Online Video Trends To Look For In 2015.” *Forbes*. Forbes Magazine, 08 Dec. 2014. Web. 14 Feb. 2015.

United States. Cong. Senate. Committee on Commerce, Science, and Transportation. *The Emergence of Online Video : Is It the Future? : Hearing Before the*

Committee on Commerce, Science, and Transportation. 112th Cong., 2nd sess. Washington: GPO, 2014. Web. 15 Feb. 2015

Verbeke, Wouter, et al. "Building comprehensible customer churn prediction models with advanced rule induction techniques." Expert Systems with Applications 38.3 (2011): 2354-2364.

Vinson, Michael, B. Goerlich, M. Loper, M. Martin, and A. Yazdani. “System and method for measuring television audience engagement.” US Patent. 8,904,419. 26 Sep. 2013.

Walker, Helen M. *Studies in the History of Statistical Method, with Special Reference to Certain Educational Problems*. Baltimore: Williams & Wilkins, 1929.

“What Does Copyright Protect? (FAQ) | U.S. Copyright Office.” *What Does Copyright Protect? (FAQ) | U.S. Copyright Office*. N.p., n.d. Web. 01 Mar. 2015.

Worstall, Tom. "The Supreme Court Should Just Abolish Software Patents In Alice v. CLS Bank." *Forbes*. Forbes Magazine, 29 Mar. 2014. Web. 01 Mar. 2015.

Zeithaml, Valarie A. "Service quality, profitability, and the economic worth of customers: what we know and what we need to learn." Journal of the academy of marketing science 28.1 (2000): 67-85.