Rare events, known as anomalies, can make a large impact on the world despite the fact they are infrequent and improbable. People use a different terms for these phenomena ranging from *outliers*, *oddities*, and *edge conditions*, to *boundary cases, pathological cases* or ***black swan** events*. What they are called doesn't matter as much as the strategy for predicting their arrival, something easier said than done.

Anomaly Detection

A Cloud-Based Predictive Analytics Framework Brings Sophisticated Capabilities to Small & Medium Size Businesses

There are plenty of examples of anomalies in Nature: “hundred year” floods, devastating earthquakes, volcanic eruptions, and pest activities like locust swarms. One theory says that the anomaly of a “near Earth object” striking the Earth drove dinosaurs to extinction. Sadly, the developed world also has its share of examples too; stock market crashes, plagues and pandemics, terrorist attacks, or the sinking of the Titanic. Clearly, it would be great to predict these events to help us better prepare for the in-enviable.

Strangely, some anomalies occur and people even know they are going on but actually finding them is a complicated challenging problem. This occurs when rare events closely resemble common harmless events. A range of crimes like identity theft, fraud, treason, and insider threats occur by people inside an organization who are seemingly otherwise trustworthy. Organizations would like to prevent these, and detecting them is now made easier using our approach of user activity analysis.

Searching for a needle in a stack of needles can drive home a point.

As the saying goes "it is like finding a needle in a haystack". While this is a fanciful description and rarely does anyone actually search for a needle in a haystack this alludes to the challenge of finding a unique object amidst countless similar ones. To make matters even worse, imagine if you have never even seen a needle, and you aren't sure exactly what one looks like even when you are looking right at it.

Extreme Value Statistics

Today, people want to use computers to find rare, yet important events they may never have seen before by ***calculating for the unknown***. Credit card companies and banks hunt for fraud, and insurance companies predict catastrophes using special techniques on a daily basis. They leverage teams of smart people and complex mathematics. Internet companies like Google, Twitter and Yahoo are concerned with making their own predictions of black swan events too. Security specialists profile network traffic looking for break-ins or attacks. All these groups construct calculations and statistics to locate departures from previous activity patterns and then focus closely on the deviations.

In general, there are two basic approaches, one is called *unsupervised* which basically means software works without any previous examples of the event it is looking for. Another approach is called *supervised* means the software is given examples of what it should trying to find. The difficulty with rare events is there aren't enough good examples for any substantial event characterization. With either approach, the results improve when changes are made to direct software toward situations you feel are correct decisions.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

K-means clustering is a popular approach for determining groups in an unsupervised analysis.

On the surface one might guess that the deviations are found anywhere there are peaks of unusual activity. This is only partially true. Actually, measurements focus on clusters of activity and the variations *inside* them as well as the sheer distance or difference from other activity.

For some anomalies time is uniquely important. This might be when administrators are monitoring the number of failed password attempts that happen *within a minute* or the number of downloads from a file server. This is called *time series*. In other cases, time is not as crucial like when someone uses a credit card that doesn't belong to them. That can happen day or night, and maybe the geographic location is more significant.

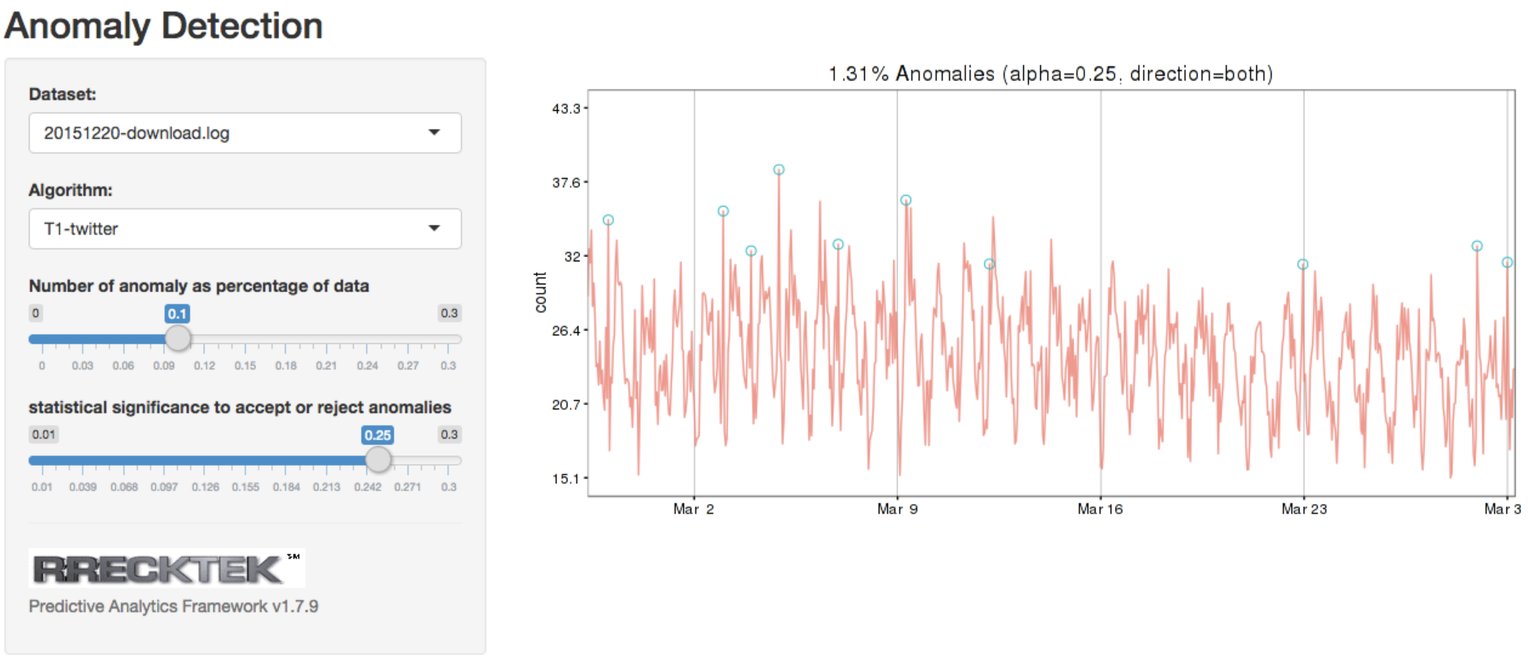
The Predictive Analytics Framework is available in the AWS Marketplace.

The company RRecktek offers a component inside its Predictive Analytics Framework that allows both novices and experts to work with anomaly detection firsthand. It doesn't take a lot of effort to get it working. It has been tested on more than half a dozen different public cloud providers, and is also available for use in your private enclave as well.

After users add their data to the system using a secure communications channel like *scp*, they can use a simple web browser to manage the controls. The capability can analyze log files from your applications, or spreadsheets produced by business processes. Your analysis process can use a highly optimized version of the R programming language and a web interface provided by the free version of Shiny server by RStudio.

We include the free software created by RStudio. Commercial versions are available from them with additional capabilities.

Any additional R development, modifications or data manipulation is supported through the RStudio Server IDE that is already installed and optimized, but the capability is operational without any changes.



Anomaly Detection

Dataset: 20151220-download.log

Algorithm: T1-twitter

Number of anomaly as percentage of data: 0.1

statistical significance to accept or reject anomalies: 0.25

Predictive Analytics Framework v1.7.9

Using only a web browser users can select which dataset to analyze and which algorithm to use. There are also two slider configuration controls. One control slider allow users to select the number of anomalies as a percentage of data, this means you might only focus on one percent of occurrences or widen the search all the way to thirty percent. A second control provides a means of tuning accuracy or t*he statistic significance necessary to identify an activity as an anomaly.*

If you are interested in finding out more, or have any questions or comments you are encouraged to contact RRecktek's principal, Ronald P. Reck, to schedule a demonstration. He can be reached at <rreck@rrecktek.com>. If you fail to receive a timely response please try again.