Apr 20th, 10:00 AM

Using Data Visualization to Inform Machine Learning Approaches

Eric N. Zielonka

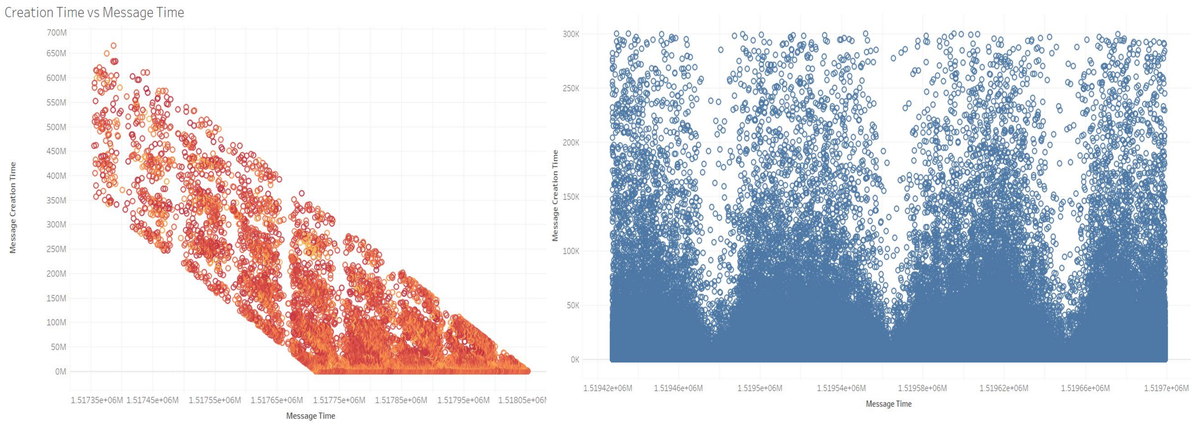
Aleksandr Fritz

Using Data Visualization to Inform Machine Learning Approaches

Eric Zielonka, Aleksandr Fritz

Department of Computer Science, Rowan University
Faculty Advisor – Bo Sun

Error Finding with Tableau

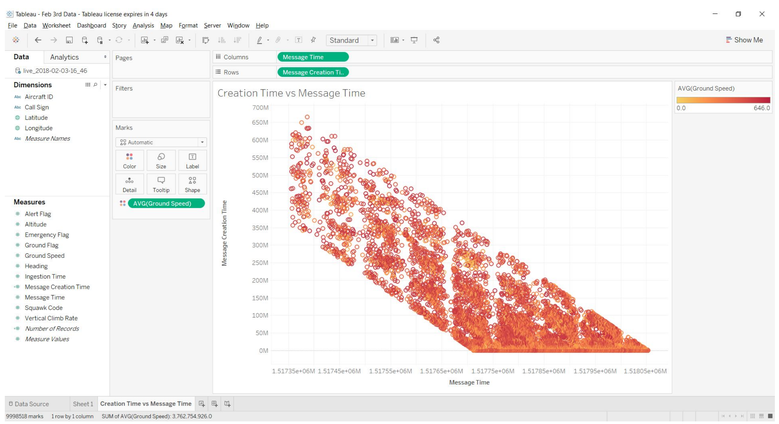


Errors In Data Collection

When the large volumes of data were originally collected to start training our machine learning algorithms, there were large issues. Data Creation Time, the amount of time it took to collect a single data point, was trending downwards, as seen above. The trend should have been flat. After finding this, we adjusted our collection algorithm, and Data Creation Time now trends relatively flat compared to before.

Tableau Approach

Tableau is a quick and easy to use program that can quickly find trends in data. It can take most standard data set types and quickly turn them into varied graphs. It uses Drag and Drop controls to define different axes on a graph. It can also create new derived fields from raw fields. The Data Creation Time field was derived from the Message Time and Creation Time fields in our original data set. By subtracting Message Time from Creation Time, we were able to find how long a message took to process in the collection algorithm.



Acknowledgements

Thanks to ASRC for their funding of this project.

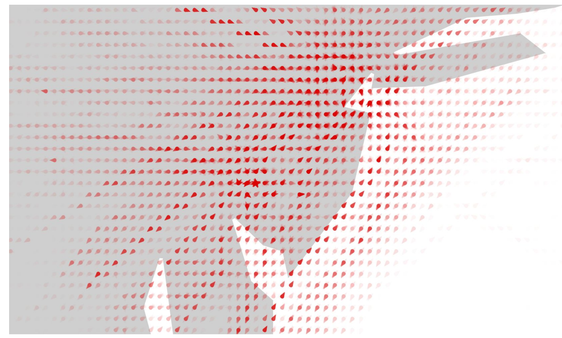
Abstract

Machine learning with big data is a complicated task to tackle. Using data visualizations, one can find trends, anomalies, and patterns to help select the appropriate approach to the problem in machine learning. Using 2D visualizations, we've displayed flight data on interactive maps, visualizing density and property changes in an area. We've also used frequency histograms to view the quantitative properties of each point to look for trends. Using scatterplots, anomalies in data collection were found. Other plots confirmed previously found trends and initial thoughts about the data. These visualizations helped inform a machine learning approach to our problem and avoided major pitfalls further down the road.

Displaying Data in Python

Beginning Attempts

This was the first prototype for showing the variations and density of the data over the whole range of locations. It was made using Python to process the data and D3.js to visualize it. The dataset used in this example was histogram data, which collected all of the data points into "buckets" rounded to .1 degree of latitude and longitude, with similar altitudes, speeds and headings. This map shows darker spots for more dense locations, and each point has arrows pointing in the heading direction. These arrows are also darker for higher frequencies of points with that heading.



Changing Data Points

Each pixel is color-coded to show the average value of all data points whose latitude and longitude coordinates fall within that area. Similar pictures were created for all the properties with numerical values, such as altitude, heading, speed, and vertical climb rate.

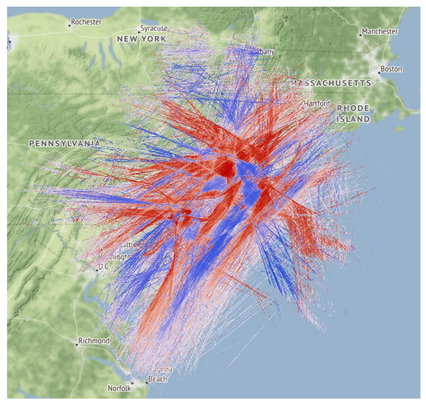
For altitude:

In this case, the gradient goes from blue through purple to yellow. Blue is the lowest altitude, and yellow is the highest.

For climb rate:

In the case, the gradient goes from blue to white to red. Blue is descending, white is around zero or level, and red is ascending.

Climb Rate



Altitude