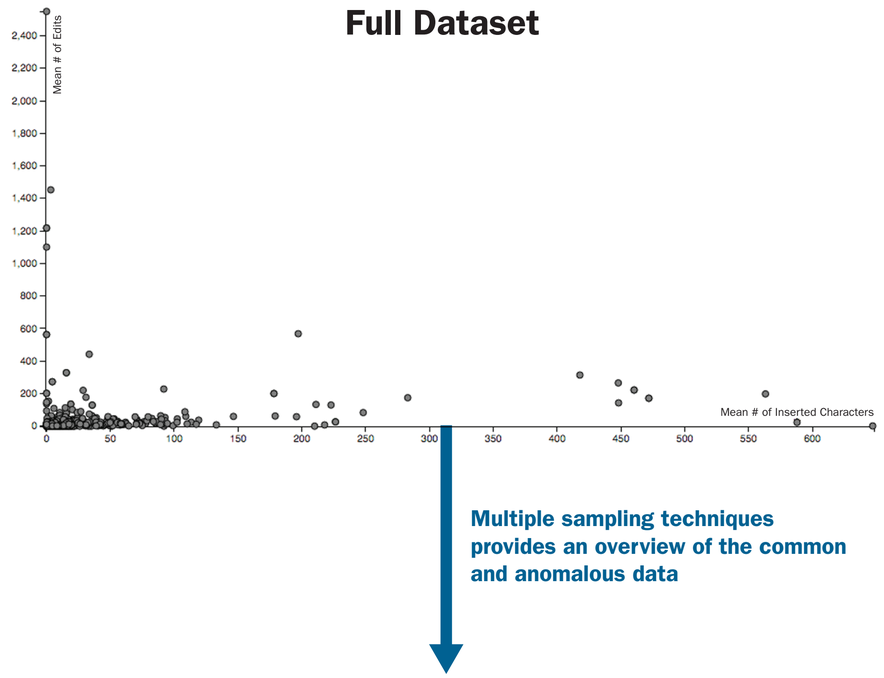
Data Validation for Large-Scale Analytics

Building Tools to Support Data Sampling and Visualization

Large-scale analytics hold great promise for government and industry, and data validation is essential to ensure that those analytics make accurate predictions. We studied practitioners in the field, built data validation tools to support data sampling and visualization, and found that our tools help practitioners generate a diverse set of insights about their data.

Why is Data Validation Important?

Data analysts agree that their biggest challenges are data quality, validating assumptions, and understanding anomalies and errors throughout the process. These challenges are not about the correctness of their code but rather the validation of data analysts' assumptions about their data and subsequent analytics. Without valid data, data science practitioners cannot be sure that their resulting machine learning algorithms are making accurate predictions using relevant features and correct labels.



Today's Data Validation Practices

The state of the art solution to data validation today is human experts who manually sort through predictions and confirm assumptions. However, the process of understanding even a subset of data points is extremely tedious and error prone, especially as the number of data points and features grows.

Data scientists today only have a few data sampling techniques available to them to give them insight into the distribution, common values, and anomalies of their data and they do not sample large datasets efficiently.

Implementing and Visualizing Multiple Sampling Techniques

We hypothesized that using many data sampling techniques would allow practitioners to learn more about their data compared to their current practice. We implemented four data sampling techniques not available on current platforms today to allow data scientists to process and sample large scale datasets:

- Random
- Uncertainty
- Query By Committee
- Density

Each sampling technique selects different subsets of the data, allowing data scientists to capture multiple views of their data more effectively than they can could with current tools.

Contact: Stephanie Rosenthal, PhD srosenthal@sei.cmu.edu