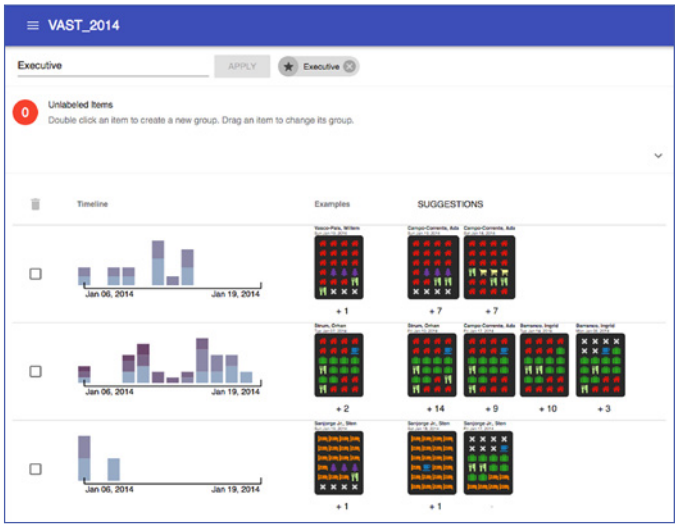
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Interactive Machine Learning

Challenge

Data scientists know that machine learning never works out of the box—it always requires a human touch. Clustering algorithms require tweaking, and trained classifiers are fragile. Real world systems require a human-in-the-loop to correct obvious mistakes and improve performance. Existing interactive machine learning systems, however, are designed for experts (i.e., data scientists) who are not likely the real end-users of the system. Furthermore, existing techniques lack scalability and overwhelm the user, forcing her to interact with all of the data in abstract statistical visualizations.

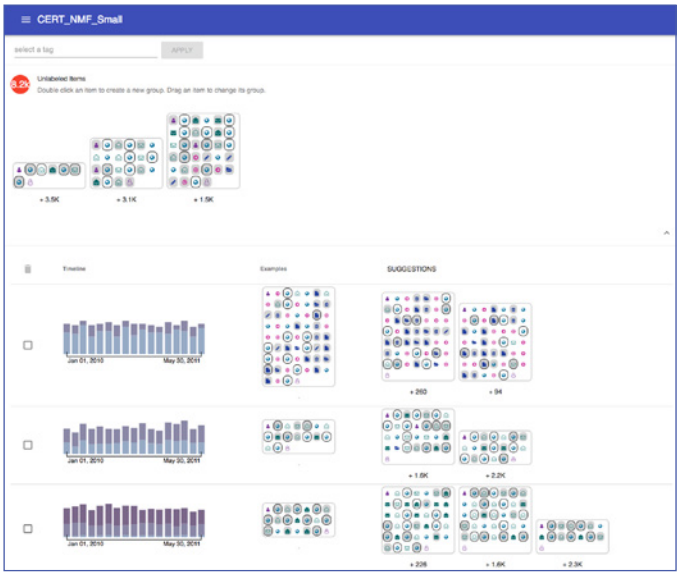


CHISSL allows a user to easily construct and refine groups from examples and then rapidly get recommendations of related content. In this figure, we are showing different patterns of life, but CHISSL has been applied to other domains including cyber, images and social media.

Often analysts are handed huge piles of data—news articles, images, tweets—that need to be summarized, organized, or triaged. CHISSL, which stands for Computer-Human Interaction + Semi-Supervised Learning, is an easy to use tool to accomplish these tasks.

Approach

We designed, CHISSL, an interactive machine learning tool, to address these challenges. CHISSL works for non-expert users, allowing users to drag and drop items to refine the model and better organize their data. While CHISSL scales to large amounts of data, the user interface keeps things simple by limiting what the user sees to the "tip of the iceberg."



CHISSL generalizes to other domains like cyber security and insider threat detection. This figure shows grouping of similar user activity sequences, where activities are actions like visiting web pages, sending emails, and connecting external devices.

Methodology

Our analytics pipeline has three key steps: encoding, representation learning, and clustering. Encoding transforms the raw data into a more structured format (e.g., event sequences or bag of words) and allows the user to incorporate domain knowledge. In the representation step, we use state-of-the-art deep learning techniques to transform encoded data to dense representations. Lastly, we apply tried-and-true agglomerative clustering to organize the data into a hierarchy. Behind the scenes, the user interface leverages this hierarchy to determine an effective "tip of the iceberg" to display, and also to rapidly incorporate user feedback.

Impact

CHISSL is fast, easy to use, and generalizable. To date we have tested CHISSL on a variety of application domains including social media, image analysis, geo-temporal analysis, and cybersecurity. Even for large datasets, CHISSL can incorporate user feedback at interactive speeds.

CONTACT

Dustin Arendt
Co-Principal Investigator
(509) 371-6902
dustin.arendt@pnnl.gov

Mark Greaves
Initiative Lead
(206) 528-3300
mark.greaves@pnnl.gov

Svitlana Volkova
Co-Principal Investigator
(509) 372-6585
svitlana.volkova@pnnl.gov

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

U.S. DEPARTMENT OF ENERGY