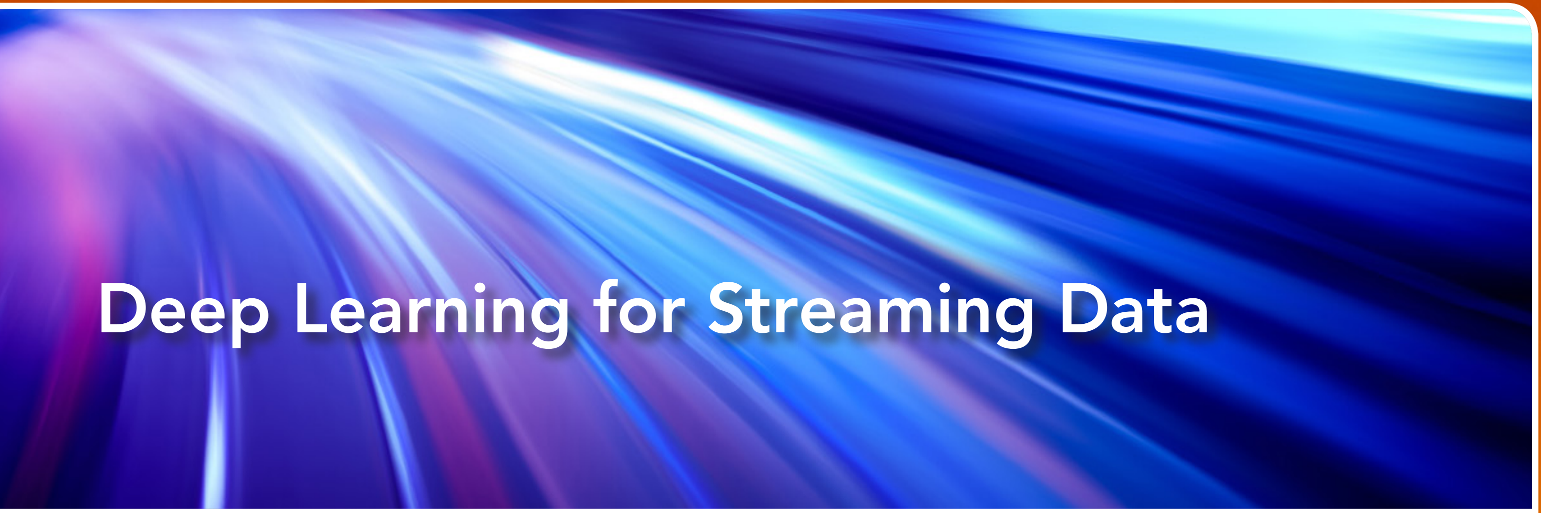
Analysis In Motion Initiative

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



Deep Learning for Streaming Data

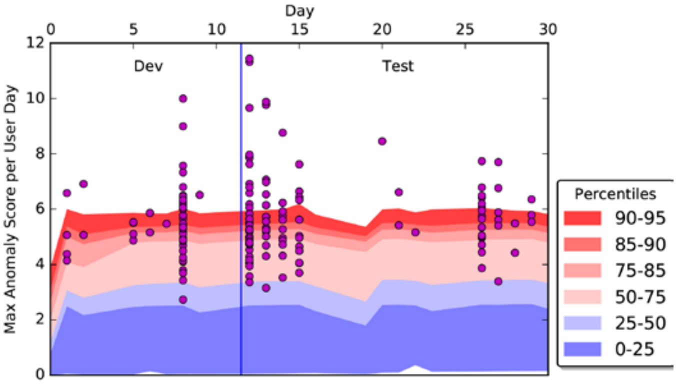
Challenge

Automatically identifying malicious behavior in computer networks poses many challenges. Threats take varied forms, making it impossible to explicitly enumerate all patterns. Raw data sources (e.g., computer and network logs) accumulate so rapidly that storing historical data is prohibitive, necessitating streaming techniques. Labeled data is rarely available, making supervised machine learning unsuitable. Finally, traditional unsupervised learning techniques that have typically been used, require time-consuming feature engineering by domain experts, leaving blind-spots in the system.

Approach

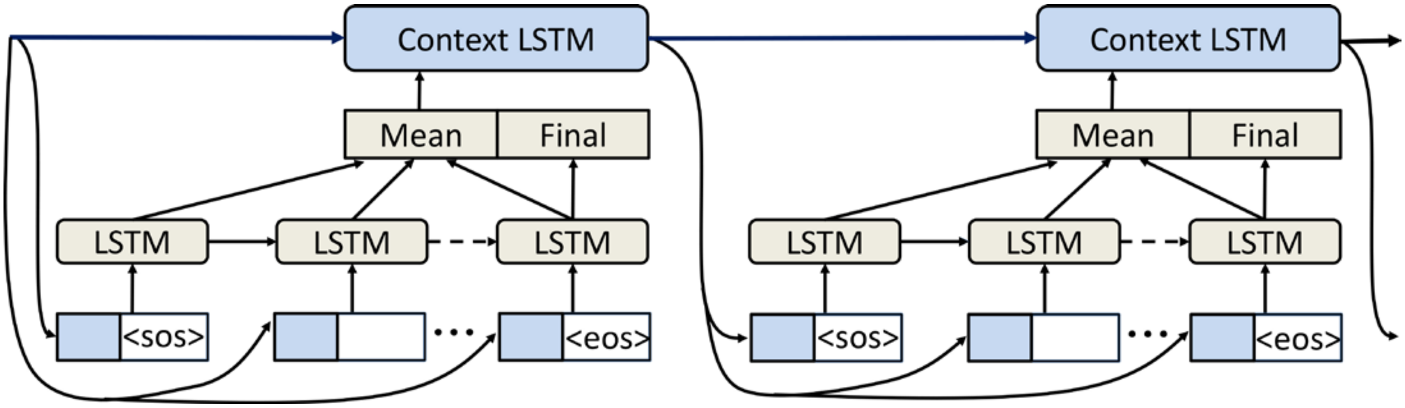
Taking a streaming anomaly detection approach, we leverage large quantities of unlabeled data to learn "normal" sequences of events for users, and we flag unusual event sequences that may indicate malicious behavior. Importantly, we operate directly at the character-level on log-line sequences, eliminating the need for manual feature engineering. From the event-level anomaly scores, we produce a single anomaly score for each user, each day. These flagged

Stream Adaptive Foraging for Evidence (SAFE) automatically identifies and characterizes malicious cyber behavior more effectively than traditional techniques by applying deep learning to streaming data sources.



User anomaly scores over time, with malicious users indicated by purple dots.

U.S. DEPARTMENT OF
ENERGY



The tiered model architecture with upper and lower tiers being implemented as Long Short-Term Memory recurrent neural networks.

users can then be provided to an analyst for further inspection, along with a break-down of the specific event sequences that most contributed to being flagged, aiding interpretability.

Methodology

We model user and computer activity sequences using a two-tiered recurrent neural network (RNN) model. The lower tier is a RNN that receives as input the individual log-line characters, one by one, as well as a "context vector" summarizing all events prior to the current one, and aims to predict the next character in the log-line. The upper tier is a RNN that tracks the dynamics of user behavior across sessions; it takes as input a sequence of log-line summaries (given by lower-tier hidden states). It produces, as output, the aforementioned context vector. The two tiers are jointly trained.

The model alternates between two modes: evaluation and training. When the data for a new day arrives, the model first evaluates the data; event-level anomaly scores are computed as the average log-probability that the model assigns to the true characters in the log-line. After all of the day's events have been scored, the model runs through them a second time, using stochastic gradient descent to update model weights to increase the log-probabilities of the true characters. Note that the model both detects anomalies and continues to train in an online fashion, requiring a single day's worth of data to be buffered.

Not only does this approach eliminate the costly feature engineering stage, but by eliminating blind spots, it also outperforms aggregated feature models on the Los Alamos National Laboratory cybersecurity dataset.

Impact

SAFE increases the effectiveness and reduces the cost of identifying and characterizing malicious behavior in computer networks and other streaming data sources.

CONTACT

Nicole Nichols
Scientist
(206) 528-3441
nicole.nichols@pnnl.gov

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

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

U.S. DEPARTMENT OF
ENERGY