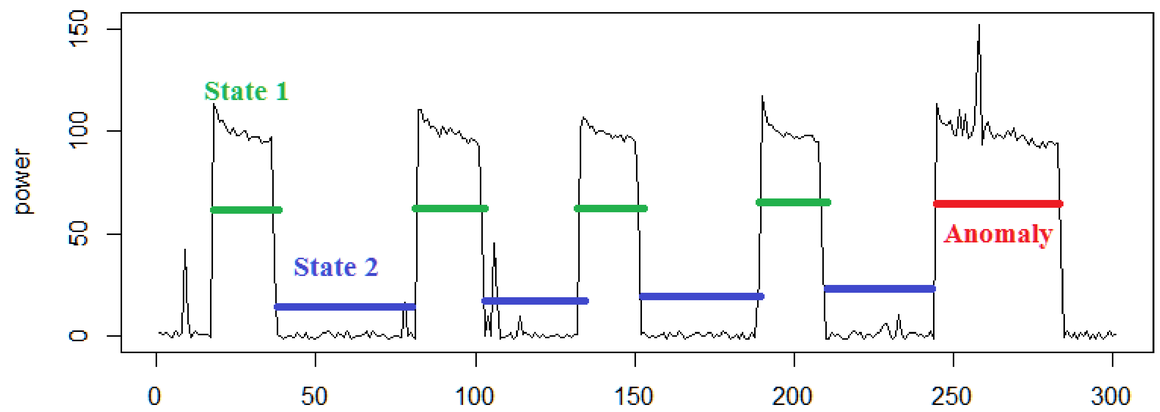
GREENHOUSE: A ZERO-POSITIVE MACHINE LEARNING SYSTEM FOR TIME SERIES ANOMALY DETECTION

Tae Jun Lee (Microsoft), Justin Gottschlich (Intel Labs), Nesime Tatbul (Intel Labs and MIT),
Eric Metcalf (Brown University), Stan Zdonik (Brown University)

Problem

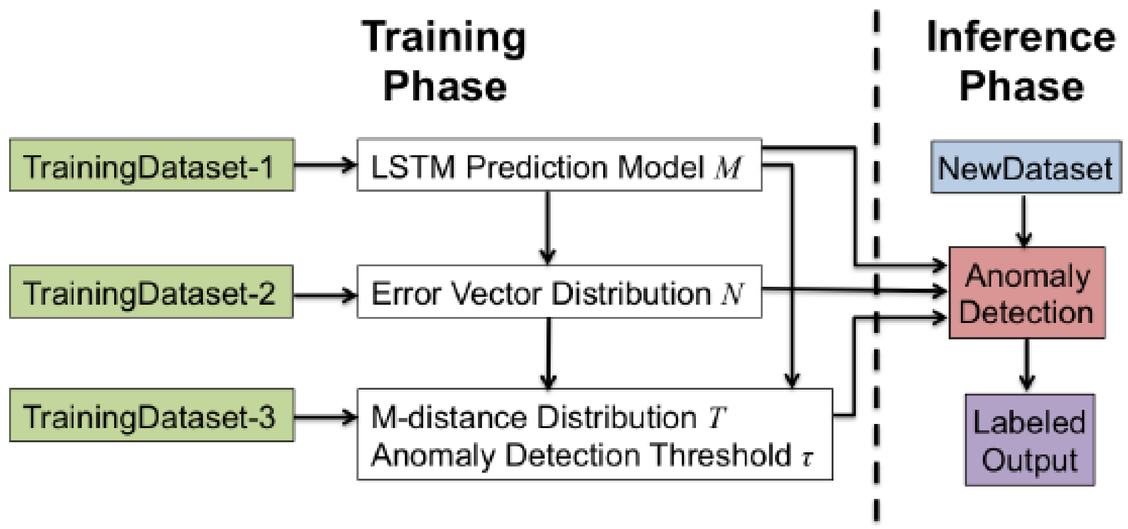
- **Time series anomaly detection**: the process of identifying non-conforming patterns over a period of time.

- **Challenges:**
  - Anomalous data are rare
  - New anomalies occur without advance notice (no data)
  - Time series anomalies can be complex



Greenhouse Solution

- **Zero-Positive Learning (ZPL)**
  - Learn anomalies by training only on non-anomalous data
- **Advantages of ZPL**
  - Reduces storage, training time, power consumption
  - Can detect new, rare, or varying anomalies
- **Additional Components**
  - Long short-term memory (LSTM) for time series
  - Three-pronged training phase combining LSTM, error distributions, and M-distance



Training Phase

1. Split non-anomalous dataset into three partitions.
2. Train an LSTM prediction model M with TrainingDataset-1.
3. Apply M over TrainingDataset-2 to make predictions and compute error vectors. Fit resulting error vectors to a multivariate normal distribution N.
4. Apply M over TrainingDataset-3 to make predictions and compute error vectors. Then compute Mahalanobis distances (M-distances) and fit the resulting M-distances into a truncated normal distribution T.
5. Evaluate the inverse cumulative distribution function of T at a user-specified percentile to be used as the anomaly detection threshold τ.

Inference Phase

1. Apply M over new dataset to make predictions.
2. Compute error vectors.
3. Compute M-distances between error vectors and center of N.
4. Label the time series values whose M-distances exceed τ as anomalies.

Table 1: Predictions and Error Vectors for an example time-series (time, value) = [$(1, v_1)$, $(2, v_2)$, $(3, v_3)$, $(4, v_4)$, $(5, v_5)$, ...]

(a) Making a Prediction

Look-Back (B=3)	Predict-Forward (F=2)
v_1, v_2, v_3	$p_{4.1}, p_{5.1}$
v_2, v_3, v_4	$p_{5.2}, p_{6.1}$
v_3, v_4, v_5	$p_{6.2}, p_{7.1}$
v_4, v_5, v_6	$p_{7.2}, p_{8.1}$
...	...

(b) Computing an Error Vector

Time Point	Error Vector (t)
t = 5	$[p_{5.1} - v_5, p_{5.2} - v_5]$
t = 6	$[p_{6.1} - v_6, p_{6.2} - v_6]$
t = 7	$[p_{7.1} - v_7, p_{7.2} - v_7]$
t = 8	$[p_{8.1} - v_8, p_{8.2} - v_8]$
...	...

LSTM-AD

- State-of-the-art time series anomaly detection algorithm [Malhotra et al, ESANN'15]
- **Greenhouse is LSTM-AD + zero-positive learning + M-distance**
- Greenhouse performance similar to LSTM-AD, despite significantly smaller training data (~25% and ~55% of LSTM-AD)
  - Precision: Performed favorably over LSTM-AD
  - F1 Score: Close to LSTM-AD despite lower recall

Table 2: Greenhouse vs. LSTM-AD [14]

	Precision	Recall	F_1 score
Greenhouse (Twitter_AAPL)	0.49	0.06	0.11
LSTM-AD (Twitter_AAPL)	0.22	0.14	0.17
Greenhouse (nyc_taxi)	0.25	0.58	0.35
LSTM-AD (nyc_taxi)	0.26	0.82	0.40

Ongoing Research

- Managing training datasets to avoid underfitting / overfitting and to preserve the continuity of time series data
- Adapting the algorithm to predict range-based anomalies (see **Range-based Precision/Recall [SysML'18]**)
- Detecting real-time anomalies (Greenhouse in online mode)
- Utilizing human feedback to improve prediction accuracy

intel

BROWN